Improving Speech Intelligibility through Spectral Style Conversion

Tuan Anh Dinh
M. S., Japan Advanced Institute of Science and Technology, 2016

Presented to the
Center for Spoken Language Understanding
within the Oregon Health & Science University
School of Medicine
in partial fulfillment of
the requirements for the degree
Doctor of Philosophy
in
Computer Science & Engineering

July 2021

Center for Spoken Language Understanding
School of Medicine
Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the Ph. D. dissertation of

Tuan Anh Dinh

has been approved.

Alexander Kain, Thesis Advisor
Associate Professor

Meysam Asgari
Assistant Professor

Xubo Song
Associate Professor

Peter Heeman
Associate Professor

Kris Tjaden
Professor

Acknowledgements

I would like to thank my advisor, Alexander Kain, for his guidance and endless support during my Ph. D. degree. He was patient to nurture and believe in me. He never blamed me for the mistakes I made. He encouraged me when the results turned out to be bad.

I would like to thank my thesis committee members: Meysam Asgari, Xubo Song, Peter Heeman, and Kris Tjaden for their insightful comments and discussion.

I would like to thank my office-mates, Liu Chen and Robert Gale, for their support and friendship. Our time at Austria was so memorable. I want to thank Ognyan Moore and Philip Robinson for their support to my internship and job hunting.

I would like to thank current and former faculty of Center for Spoken Language Understanding for their important input and motivation. Thanks go to Jan van Santen, and Steven Bedrick for their comments and discussion, as well as Patricia Dickerson for her great administrative support.

Last, but not least, I want to thank my parents for their sacrifices to support me. I also want to thank my brother, Duc, for his supports.

This work was partially supported by NIH grants R01DC004689, R01DC016621 and R03DC013990. Any opinions, findings, conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the NIH.

Contents

List of Tables

List of Figures

Abstract

Improving Speech Intelligibility through Spectral Style Conversion

Tuan Anh Dinh

Doctor of Philosophy
Center for Spoken Language Understanding
within the Oregon Health & Science University
School of Medicine

July 2021
Thesis Advisor: Alexander Kain

Oral communication is the most important way for delivering information in our daily life. Unfortunately, the quality of such communication can be degraded by 1) speech disorders (e.g. dysarthria) and 2) surrounding environments (e.g. noise or reverberation). Style conversion is a technology that modifies the source speaking style of a speaker to sound like a more intelligible target speaking style of either the same or different speaker. For speech enhancement, style conversion helps either typical or disordered speech become more intelligible prior to its presentation in adverse environments. The technology can make the widely-used speaking-devices in commercial (e.g. mobile phone, GPS), medical (e.g. assisted-speech) and military (e.g. ground troop relays) circumstances, more intelligible. Moreover, the technology can become instrumental for the next generation of speaking-aid and hearing aid devices, which is highly demanded now.

In the dissertation, I consider new machine learning based-approaches for style conversion. Inspired by the intelligibility gain of *clear* (CLR) speaking style over *habitual* (HAB) speaking style, I propose several HAB-to-CLR spectral mappings approaches for intelligibility improvement.

In the first approach, I propose a machine-learnable, compact and interpolable representation for spectral style conversion, which was realized by the means of the *Variational Autoencoder* from the high-dimensional Mel-cepstrum coefficients. In a vocoding experiment, I showed that using a 12-dimensional VAE-based representation (VAE-12) achieved significantly better perceived speech

quality compared to a 12-dimensional Mel-cepstrum feature. In a voice conversion experiment, I showed that mapping VAE-12 resulted in significantly better perceived speech quality compared to a 40-dimensional Mel-cepstrum feature, with similar speaker accuracy, thus demonstrating the efficiency of mapping in a low-dimensional latent feature space. In a HAB-to-CLR conversion experiment, I showed that this VAE-12 together with a custom skip-connection deep neural network significantly improved the speech intelligibility for one speaker with *mild* dysarthria, with the average keyword recall accuracy increasing from 24% to 46%.

In the second approach, I propose the use of *conditional Generative Adversarial Nets* (cGANs) in HAB-to-CLR spectral mappings for typical speakers and speakers with mild dysarthria. Specifically, our cGANs-based spectral style mapping can address the over-smoothing issue of our previous feed-forward networks-based spectral style mapping. I evaluated the performance of the cGANs in three tasks: 1) speaker-dependent one-to-one mappings, 2) speaker-independent many-to-one mappings, and 3) speaker-independent many-to-many mappings in terms of intelligibility. In the first task, cGANs outperformed a traditional deep neural network mapping in terms of average keyword re-call accuracy and the number of speakers with improved intelligibility. In the second task, I significantly improved intelligibility of one of three speakers, without any source speaker training data. In the third and most challenging task, I improved keyword recall accuracy for two of three speakers, but without statistical significance.

In the third approach, I propose two conversion methods to improve naturalness and intelligibility of alaryngeal speech (LAR), which is more distorted than mild dysarthria. Specifically, the first method utilized a feed-forward network for predicting binary voicing/univoicing and the degree of voicing (aperiodicity). The second method adopted cGANs to learn alaryngeal speech spectra to clearly-articulated speech spectra. To address the unusable fundamental frequency (F0) information of alaryngeal speech, I created a synthetic fundamental frequency trajectory with an intonation model consisting of phrase and accent curves. For the two conversion methods, I showed that adaptation always increased the performance of pre-trained models, objectively. In subjective testing involving four LAR speakers, I significantly improved the naturalness of two speakers, and I also significantly improved the intelligibility of one speaker.

In the fourth approach, I report preliminary results of improving speech intelligibility using duration conversion. Although these results were not positive, I show potential directions for further study on duration conversion and speech intelligibility.

Overall, the results show the potential of applying machine learning techniques in mapping speech to improve its intelligibility. These methods can improve speech intelligibility for typical speakers, speakers with mild Parkinson's disease, and more serious case of alaryngeal speech.

Chapter 1

Introduction

1.1 Motivation

1.1.1 Unintelligible Speech

Speech is probably the most important biosignal for human communication. The way that people typically talk is referred to as habitual speech [172]. However, habitual speech becomes less intelligible in noise conditions. Habitual speech is also hard to understand for people with hearing impairments (e.g., due to age) and non-native speakers. Current speaking devices (e.g., Amazon Alexa, and Apple Siri), which create synthetic speech similar to the habitual speech of people, can also be difficult to understand, especially in noisy environments, which exacerbate many speech conditions (Figure 1.1). There are also cases involving atypical speakers whose speech is hard to understand (Figure 1.2). For example, mild dysarthria, which is a speech motor disorder, usually results in a substantive decrease in speech intelligibility, especially in noise conditions. Alaryngeal speech, which is produced by people who have undergone laryngectomy, is even harder to understand.

1.1.2 Listener Side Solution

One approach to increase the intelligibility of speech in noise is to use noise suppression and cancellation. Specifically, Michelsanti explored a conditional Generative Adversarial Nets-based approach for mapping the spectral features of noisy speech to those of intelligible speech [143]. Jean-Marc focused on the main perceptual characteristics of speech – spectral envelope and periodicity – for noise suppression in real time with low complexity [218]. Isik proposed POCoNet involving a large U-Net with DenseNet and self-attention blocks with frequency-positional embeddings for high-quality noise cancellation [92].

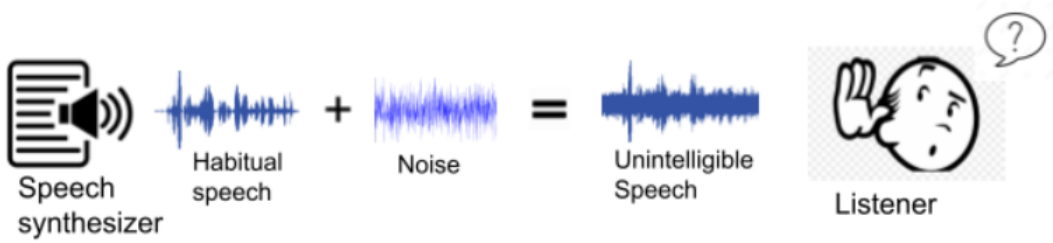


Figure 1.1: Synthetic speech of speaking devices is degraded by noise

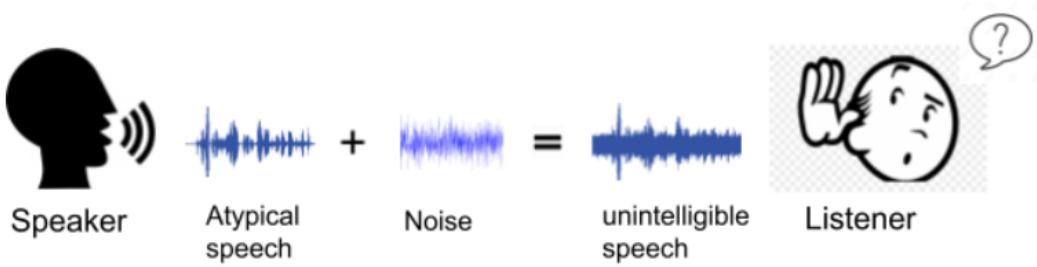


Figure 1.2: Atypical speech is hard to understand, especially in noise

However, the above approaches require listeners to use *noise cancellation devices* (e.g., noise-cancelling headphones), which take as input a noisy speech signal and output an enhanced signal with higher intelligibility and quality. Thus, the processing happens on the *listener side*. Yet there are many cases, such as transit announcements, where listeners do not have noise cancellation devices; thus, a listener-side solution might not be practical in many cases.

1.1.3 Lessons from Real Speakers: Habitual versus Clear Speech

In real life conversation, people adjust their voice to overcome communication difficulties due to speaking disorders, hearing impairment of listeners, and background noise. For example, teachers speak louder and more slowly to help students understand better. To make habitual speech more intelligible in noise, speakers adopt special *clear* speaking styles, which are resilient to changing environments and listeners' specificity. This clear speech is highly articulated. Specifically, researchers reported the extension of phoneme duration in clear speech [173, 60, 61]. In combination with extended phoneme duration, the longer and more frequent pauses lead to a significant decrease of speaking rate, from 160–200 words per minute (wpm) in habitual speech to 90–100 wpm in clear speech [173, 121]. These studies also showed that clear speech has higher intelligibility than habitual speech in adverse environments. The relationship between acoustic changes of prosodic and spectral features and the intelligibility gain of clear speech was investigated in a number of studies [95, 203]. The spectral and prosodic variations from habitual to clear speech are probably

significant contributors to the improved intelligibility of clear speech. In addition, Brenk also found that the combination of spectrum and duration is a contributing factor to the varied intelligibility of slow speech [27].

1.1.4 Speaker Side Solution

In contrast to a listener side solution, a speaker side approach is to convert habitual speech directly from speakers into clear speech prior to its distortion due to background noise. The conversion into clear speech should make the speech more resilient to noise (see Figure 1.3 and Figure 1.4). Specifically, habitual speech (typical or dysarthric speech) is converted into clear speech using signal processing or machine learning-based methods. Thus, the approach is known as as *style conversion*, which aims to modify a sentence uttered in a source speaking style to sound like it is uttered in a target speaking style. In other words, style conversion alters the style-dependent characteristics of speech signals, such as spectral and prosodic features, in order to *improve the perceived intelligibility* of linguistic content. Different from listener side approach, the speaker side approach has the benefit of processing a clean input signal, which is more convenient than processing noisy signals on the listener side.

Another speaker-side approach is to convert the speech of atypical speakers into the intelligible speech of normal speakers. The approach is preferred when the clear speech of atypical speakers is not available or it is impossible for the atypical speakers to create intelligible speech.

1.1.5 Previous Work on the Speaker Side Solution

There has been previous work on the speaker side solution. Koutsogiannaki investigated the characteristics of the short-term energy of clear speech compared to habitual speech [118]; then she applied filters to habitual speech to create these characteristics of clear speech. Her technique resulted in modified speech with higher intelligibility. However, her approach also showed a trade-off between intelligibility and naturalness of modified speech because those speech modifications are impossible for speakers to make. Moreover, her approach did not model the conversion from habitual to clear speech.

To make dysarthric speech become more intelligible, Mohammadi utilized HAB-to-CLR spectral style conversion on habitual vowels using a Gaussian Mixture Model [145]. In a different approach, Kain converted dysarthric speech into typical speech using a Gaussian Mixture Model [96]. Kazuhiro and Othmane converted alaryngeal speech into typical speech using deep neural networks [109, 16]. The machine learning-based methods (e.g., deep neural networks) showed the

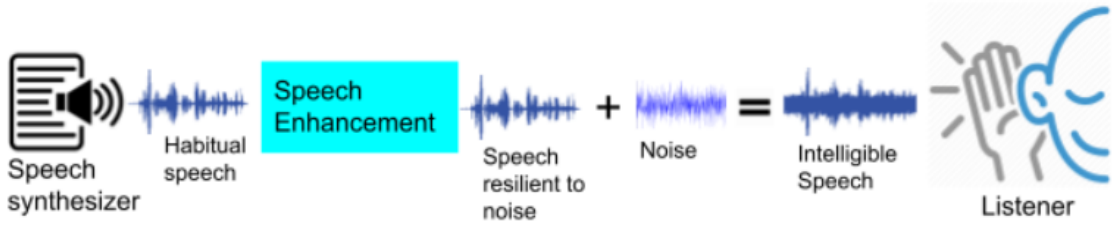


Figure 1.3: Make habitual speech (generated by speech synthesizer) more resilient to noise

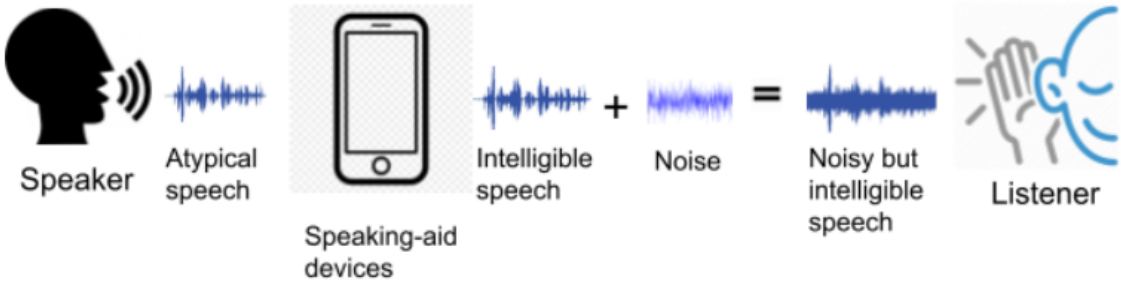


Figure 1.4: Make atypical speech (spoken by people with dysarthria) more resilient to noise

most promising results; but there is still room for improvement.

1.2 Thesis Problem and Statement

Modifying the habitual speech of typical and atypical speakers on the speaker side to increase intelligibility in noise is a challenging problem. My thesis is that the speech intelligibility of typical and atypical speakers can be improved automatically by learning how they map their voice and make it more intelligible.

Specifically, I converted habitual speech of typical speakers into clear speech, which is known as style conversion, using machine learning-based methods. I also evaluated the efficacy of these style conversion methods on a more challenging case involving an atypical speaker with mild dysarthria. I further pursued the challenge of converting alaryngeal speech, which is barely understandable compared to mildly-dysarthric speech, into intelligible speech.

1.3 Specific Aims of the Dissertation

First objective: To determine effective spectral representations for spectral voice and style conversion. Up until now, high-dimensional (e.g., 40-dimensional) Mel-cepstral coefficients (MCEP), a commonly used short-term spectral feature in speech processing, have been

mapped between source and target in voice and style conversion. Although high-dimensional MCEPs have shown an exceptional quality of speech vocoding (analysis and synthesis), they may not exhibit the necessary interpolability for voice and style conversion. An interpolable feature ensures that even when two or more parameter vectors are averaged (e.g., as part of a mapping procedure), the result remains near the manifold of possible speech. I contrasted two new sets of spectral features: 1) *probabilistic peak tracking* (PPT) features, which is a *formant*-like hand-crafted feature, and 2) *manifold* features, which is machine learnable by Variational Autoencoder (VAE) [115]. The hypothesis is that compact and interpolable spectral features are more effective for voice and style conversion mappings between source and target. The two sets of features were integrated into a high quality vocoder, WORLD [152]. I extensively evaluated the two sets of spectral features by comparing them to each other and to baselines, which are two commonly used spectral representations: *line spectral frequency* (LSF) and *Mel-cepstrum coefficients*, in speech reconstruction, voice conversion, and style conversion tasks (Chapter 3).

Second objective: To develop effective HAB-to-CLR spectral mappings using well-established machine learning algorithms. Motivated by the success of conditional Generative Adversarial Nets (cGANs) [93] in machine learning, I utilized cGANs to map the spectral features of habitual speech to those of clear speech. The hypothesis is that the cGANs-based mappings can do detailed spectral modifications, unlike commonly used statistical and rule-based methods [145, 117], which can achieve better performance of spectral style conversion mappings. Specifically, cGANs were investigated in three spectral style conversion mappings: 1) one-to-one mappings, 2) many-to-one mappings, and 3) many-to-many mappings for intelligibility improvement in noisy environments. I compared the performance of cGANs-based mappings to a baseline of feed-forward networks with custom skip-connections in one-to-one style conversion mappings. I extensively evaluated the performance of the three mappings on both typical speakers and speakers with mild dysarthria (Chapter 4).

Third objective: To develop effective conversion methods from alaryngeal speech to intelligible speech, using well-established machine learning algorithms. Alaryngeal speech is unnatural sounding and difficult-to-understand speech for several reasons, including poor voice quality, poor voiced/voiceless differentiation, and poor articulatory precision [112]. It is important to note that intelligible speech is different from clear speech in the previous objectives. The hypothesis is that utilizing machine learning-based spectral conversion mappings, voicing,

and voicing degree prediction can compensate for severe speech disorders (e.g., due to laryngectomy). Therefore, I propose an approach that has two parts for transforming alaryngeal speech (LAR) to intelligible speech (INT). The first part predicts binary voicing/univoicing and the degree of voicing (aperiodicity) using feed-forward networks. The second part is for LAR-to-INT spectral mappings using cGANs. Moreover, to address the unusable fundamental frequency (F0) information of LAR speech, I created a synthetic fundamental frequency trajectory with an intonation model consisting of phrase and accent curves. I evaluated the LAR-to-INT conversion methods on an alaryngeal speech database (Chapter 5).

Fourth objective: To investigate the performance of uniform and non-uniform duration style conversion. I show preliminary results of improving speech intelligibility using duration conversion. The hypothesis is that phoneme (and even sub-phoneme) segments are changed unequally (or non-uniformly) during the process; and a non-uniform duration style conversion is better in improving speech intelligibility in comparison to a uniform conversion. Therefore, I evaluated the performance of uniform and non-uniform duration conversion in terms of intelligibility in an ideal case when (oracle) sentence and phoneme-level scaling factors are given (Chapter 6).

1.4 Contributions

The contribution of the first objective is a compact and interpolable manifold feature, which is effective for speech reconstruction, spectral voice conversion mappings, and HAB-to-CLR spectral style mappings. For speech reconstruction, the manifold feature, which is realized by VAE, was as good as commonly-used spectral features. For voice conversion mappings, using the new representation obtained better speaker similarity than using high-dimensional MCEPs and LSFs. For style conversion mappings, I significantly increased the sentence-level intelligibility of dysarthric speech in noisy environments with a subjective evaluation. This is reported in Tuan Dinh, Alexander Kain, Kris Tjaden, *Using a Manifold Vocoder for Spectral Voice and Style Conversion*, Interspeech, 2019.

The contribution of the second objective is a novel cGAN-based spectral style mapping between habitual and clear speech. I increased the sentence-level intelligibility of dysarthric speech in noisy environments, subjectively and significantly. This finding is published in Tuan Dinh, Alexander Kain, Kris Tjaden, *Improving Speech Intelligibility through Speaker Dependent and Independent Spectral Style Conversion*, Interspeech, 2020.

The contribution of the third objective is two conversion methods and a fundamental frequency (F0) synthesis method for LAR-to-INT speech mappings. I improved the perceived naturalness and

intelligibility of alaryngeal speech, significantly. This finding is published in Tuan Dinh, Alexander Kain, Robin Samlan, Beiming Cao, Jun Wang, *Increasing the Intelligibility and Naturalness of Alaryngeal Speech Using Voice Conversion and Synthetic Fundamental Frequency*, Interspeech, 2020.

The contribution of the fourth objective is an analysis on the effect of changing speaking style on phoneme and sentence duration, which showed that phoneme duration was not changed uniformly. However, my effort of conducting non-uniform style duration conversion did not outperform the uniform conversion. I attribute the failure to the artifacts created by duration conversion algorithms. Further work should be done to reduce the artifacts in order to examine the performance of duration conversion on speech intelligibility.

1.5 Dissertation Outline

In Chapter 2, I present the 1) differences between clear speech and habitual speech, 2) relationship between acoustic features and improved intelligibility of clear speech, 3) a literature review of voice conversion methods, 4) subjective measurements of intelligibility, and 5) a literature review of automatic approaches for intelligibility improvement. In Chapter 3, I contrast manifold features to probabilistic peak tracking features in speech vocoding, voice conversion, and style conversion. In Chapter 4, I present my cGAN-based mapping for spectral style conversion. In Chapter 6, I present my preliminary study on uniform and nonuniform duration style conversion for intelligibility improvement. In Chapter 5, I present two conversion methods and a fundamental frequency synthesis method for improving naturalness and intelligibility of alaryngeal speech. Finally, I summarize my contributions and present possible future directions in Chapter 7.

Chapter 2

Background and Related Work

In this chapter, I review the background literature. In Section 2.1, I review the literature regarding intelligibility and acoustic differences between habitual and clear speech. In Section 2.2, I review the relationship between acoustic features and speech intelligibility. In Section 2.3, I review voice conversion frameworks, because I used voice conversion techniques for improving intelligibility. In Section 2.4, I review automatic methods for improving intelligibility. Finally, I review different assessment methods of speech intelligibility in Section 2.5.

2.1 Habitual and Clear Speech

Picheny used the term *habitual speech* (HAB) to refer to speech produced under the instruction, "speak in the same manner as you would during an ordinary conversation" [172]. In contrast, he used *clear speech* (CLR) to refer to speech produced under the instruction "speak clearly, as you would when talking to hearing-impaired listeners". Although the term "clear" implies higher intelligibility of perceived speech, it is possible that clear speech does not have intelligibility advantages over habitual speech for some listeners [59]. In this dissertation, I refer to habitual and clear speech as the speaking styles in response to the above-mentioned instructions. In Section 2.1.1, I review work on the intelligibility of habitual versus clear speech under a variety of conditions, and in Section 2.1.2, I review work on the acoustic differences between habitual and clear speech.

2.1.1 Intelligibility of Clear Speech

There has been great interest in the intelligibility gain of clear speech over habitual speech, which have both been examined with various listener groups, including a) 18–32 aged *normal-hearing* listeners [120, 133, 59, 137], b) 61–88 aged normal-hearing listeners [84], c) 60–89 aged *hearing-impaired* listeners [172, 181, 213], d) 19–33 aged, simulated hearing-impaired listeners [137], and e) school-aged children with and without *learning disabilities* [25]. The common finding is that

clear speech is more intelligible than habitual speech, across all speakers. For example, Pinechy reported a significant intelligibility difference of 17 percentage points for hearing-impaired listeners compared with habitual and clear *nonsense sentences* [172]. The improved intelligibility of clear speech was found to be independent of 1) listeners, 2) presentation levels, and 3) frequency-gain characteristics. However, Ferguson concluded that the advantage of clear speech was listener-dependent [60]. Ferguson also found that there is no advantage for clear vowels when elderly hearing-impaired listeners identified the front vowels, which they attributed to the raising of F2 values to the hearing-loss region of the hearing-impaired listeners (e.g. 2000–2500 Hz). Similarly, Maniwa reported that simulated hearing-impaired listeners benefited from clear speech in discriminating sibilants in /s/–/ʃ/, and /z/–/ʒ/ pairs, but they did not benefit from clear speech in discriminating voiceless non-sibilants in /f/–/θ/ pairs [137]. Ferguson and Maniwa showed that the advantage of clear speech depends on the hearing ability of listeners as well as on the kinds of hearing loss (if any). Thus, the intelligibility advantage of clear speech depends on the the age of listeners and their hearing ability [60].

In addition to listener conditions, speech materials have been investigated for intelligibility differences between habitual and clear speech. There are two principal kinds of speech materials: 1) word-level materials, with nonsense syllables allowing more control over the phonemes to be evaluated, and 2) sentence-level materials resembling daily communication. Examples of the first type include vowels in /b/–/V/–/d/ context [60], and VCV syllables where the consonant is a fricative [137]; examples of the second type include nonsense sentences [172, 171] and *meaningful sentences* [181, 24]. A drawback of using meaningful sentences is existing semantic cues, which could be used by listeners to compensate for the reduced intelligibility [74]. Several studies found that elderly listeners (aged 65-77) are relatively better at using these semantic cues than younger listeners (aged 22-29) due to well-preserved linguistic knowledge with age, although aging reduced hearing ability in the presence of background noise [191, 225, 174, 73]. Additionally, it is important to note that phoneme-level intelligibility should not be used to obtain sentence-level intelligibility [7].

2.1.2 Acoustic Differences between Habitual and Clear Speech

There has been great interest in investigating the acoustic differences between habitual and clear speech [173, 60, 25, 121]. Interestingly, the findings were not always in agreement due to the variability of speakers, speech materials, and analysis methods. Major findings have focused on the differences in three main aspects of speech: 1) prosodic (a combination of fundamental frequency, energy and phoneme duration), 2) spectral (such as formant and formant-normalized spectrum),

and 3) phonological aspects.

The prosodic studies of clear speech show a slight increase in mean and variability (or range) of fundamental frequency (F0) [173, 25, 121]. Also, an increased consonant-vowel energy ratio (CVR) has been reported in clear speech for stops and fricatives [25]. In contrast, the increased CVR was only reported in affricates of clear speech [121]. Picheny found greater *root-mean squared* (RMS) intensities for unvoiced stop consonants in clear speech [173]. Researchers also reported the extension of phoneme duration, especially in the tense vowels: /iː/, /u/, /ɑ/, and /ɔ/ in clear speech [173, 60, 61]. In combination with the reduced phoneme duration, the longer and more frequent pauses lead to a significant decrease of speaking rate from 160–200 words per minute (wpm) in habitual speech to 90–100 wpm in clear speech [173, 121]. Additionally, Krause showed an increased amplitude modulation for low modulation frequencies (up to 3–4 Hz) of clear speech on a small number of speakers [121]. Krause concluded that the increased depth of envelope helped syllables to be better distinguished from one another. Krause et al. also investigated the duration between the time of burst and the onset of the voicing (VOT). They found increased VOTs for voiceless stop consonants in clear speech for one of the speakers.

Spectral studies of clear speech show an expanded vowel space via formant frequencies [173, 60, 25]. Other work shows higher energies of long-term average spectra at higher frequencies [121], with Krause showing a deceased spectral tilt. The second formant displacement from the target has been shown to be significantly less in clear speech; specifically, Picheny found that formant displacement was dependent on vowel duration more for lax vowels for both habitual and clear speech [173], and Moon found more variation of formant frequencies in lax vowels [149]. Additionally, Godoy compared the averaged short-term spectral envelope of clear and habitual speech [66], which showed that clear speech has higher energy in two frequency bands: [2000, 4800] and [5600, 8000].

Phonological studies showed: 1) vowel reduction (e.g., vowels becoming shwa-like), 2) degemination (e.g., two similar phonemes merged into one sound), and 3) alveolar flaps occurred more often in habitual speech [119]. In contrast, bursts of the stop consonants in word final position tended to be released more often, and the sound insertion of a schwa after a voiced consonant occurred more often in clear speech [173, 121].

In conclusion, existing studies show the main acoustic differences between clear and habitual speech as: 1) increased F0 mean and range of clear speech relative to habitual speech, 2) longer phoneme duration in clear speech, 3) increased amplitude modulation for clear speech, 4) increased vowel spaces in clear speech, 5) higher energies at higher frequency regions in clear speech, and 6) phoneme insertions (e.g., schwa) occur more often in clear speech.

In the next section, I discuss the contribution of the features to speech intelligibility.

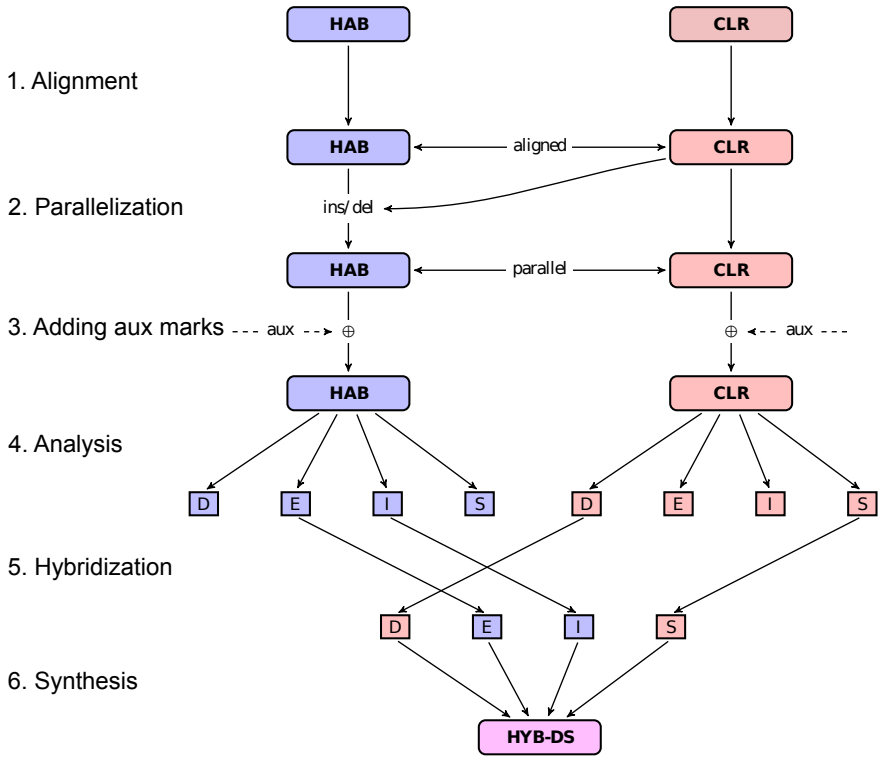


Figure 2.1: Hybridization method to investigate the acoustic cause for the improved intelligibility of clear speech (CLR). HAB denotes habitual speech. The features in consideration are duration (D), energy (E), intonation (I), and spectrum (S).

2.2 Acoustic Features and Speech Intelligibility

The correlation between changes in certain acoustic features (e.g., fundamental frequency) and speech intelligibility have been considered [77]. The relationship between stimulus variability and spoken word recognition has also been investigated [190]. Moreover, a hybridization method (as shown in Figure 2.1) has been used to examine the acoustic causes of intelligibility gain in clear speech [203] and slow speech [27]. I review the relationships between different prosodic and spectral features and speech intelligibility. I do not limit the review to the intelligibility of clear speech, but I summarize general studies of acoustic features and speech intelligibility [190, 26, 77, 27, 203].

2.2.1 Prosodic Features

Fundamental Frequency (F0)

Fundamental Frequency (F0) studies show no correlation between mean F0 and sentence intelligibility when considering gender [26]. Other studies show decreased or increased F0 values with

a global percentage of 10, 20, or 30%, which did not have any impact on word identification rate [189]. A phonetic relevance hypothesis reported that F0 mean is not relevant to intelligibility [189]. In other research, the range of F0 (the difference between maximum F0 and minimum F0) showed significant correlations with sentence intelligibility for one out of 20 speakers [26]. In contrast, another study showed that the correlation between F0 range and word intelligibility is not significant [77]. The differences between the these studies include 1) speech material (sentence versus word intelligibility), 2) measurement of F0 values (logarithmic versus linear scale), and 3) speakers.

The role of F0 as an important hint for English phoneme identification remains arguable. English vowels can be described in terms of the frequencies of the lowest three formants and formants transitions, regardless of F0 values. Similarly, other findings also confirmed that phoneme identity of non-tonal languages such as English is virtually independent of F0 [168]. In contrast, other findings showed that the distance between F0 and F1 affect the perception of vowel height. Although the mean F0 values of clear speech tends to be higher, the relationship between F0 and intelligibility remains debatable.

Energy

There were two types of energy measurements that were considered: 1) the consonant-vowel ratio (CVR), which is the relative energy ratio between consonants and neighboring vowels, and 2) overall energy, which is the average energy of speech signal. One study found that the artificial amplification of the CVR can improve intelligibility (on the order of 10 percentage points) at both the VCV word and nonsense sentence-level [78]; however, this study did not specify which consonants to amplify or the specific level of amplification. Although there is an increased CVR in clear speech due to stop release burst and fricatives [25], the CVR may not be a contributing factor to the increased intelligibility of naturally spoken speech. However, artificial amplification of the CVR was effective for intelligibility improvement.

In contrast, overall energy or intensity was reported to significantly affect intelligibility [63]. Overall energy is not a factor of interest; therefore, most studies normalize the overall amplitude for both habitual and clear speech to the same level. Even after normalizing energy, clear speech is still more intelligible than habitual speech, which indicates that other features cause improved intelligibility. In a recent study, Brenk investigated the effect of root-mean-squared (RMS) energy on intelligibility gains and losses of slowed speech, using a hybridization technique [27] (Figure 2.1). He concluded that the RMS energy trajectory was an important contributing factor to the increased or decreased intelligibility of slowed speech for people with dysarthria secondary

to Multiple Sclerosis.

Duration

A variety of studies looked at the effect of phoneme, word, and sentence-level duration on speech intelligibility. One study reported a positive correlation between monosyllabic word duration and word-level intelligibility [77]. In contrast, another study showed no correlation between speaking rate, which was measured from overall sentence duration, and sentence-level intelligibility [26]. Natural and artificial changes in speaking rate resulted in impaired identification of spoken words; this finding showed the importance of speaking rate for intelligibility [189].

When matching the speaking rate of habitual and clear speech (with differences of no more than 25 wpm), clear speech was still more intelligible than habitual speech (59% intelligibility of clear speech compared to 45% intelligibility of habitual speech), which showed that other features cause the improved intelligibility [121].

Hillenbrand reported phoneme duration as an important cue for vowel identity [85]. Varying the vowel duration of /hVd/ syllables degraded vowel identity and significantly affected the vowel contrasts of (/ɑ/-/ɔ/-/ʌ/), and (/æ/-/ɛ/). Bradlow reported a positive correlation between stop closure duration and rate of /d/ detection as in "walled town" [26]. They also showed that a long duration of /s/ relative to the surrounding vowels as in "play seems" led to syllable affiliation ("place seems"). Thus, the *inter-segmental timing* was concluded to be important for speech intelligibility.

Although Hillenbrand found that the phoneme-level perceptions were less important than word- or sentence-level perception [85], one should pay attention to inter-segmental timing in a controlled experiment at the phoneme-level. The failure of intelligibility improvement by uniformly stretching phoneme duration can be attributed to the disrupted naturalness of inter-segmental timing. Uchanski addressed the errors in inter-segmental timing of modified speech by borrowing the phoneme duration from clear speech [214].

Although a variety of studies have examined the relationship between speech intelligibility and phoneme, word, and sentence duration, the findings are not consistent. Therefore, based on these studies, conclusions that the duration (or speaking rate) alone is the acoustic cause of increased intelligibility of clear speech cannot be drawn.

Along with clear speech, a variety of studies have investigated the intelligibility benefit from slowed speech. Specifically, the intelligibility gain of slow speech was investigated for speakers with dysarthria secondary to Parkinson's disease (PD) and Multiple Sclerosis (MS) [206]. The

comparison of intelligibility between habitual and SLOW speaking rates showed increased intelligibility for both speaker groups when slowing down speaking rates. In a follow-up study, Stipanic further confirmed the variable effects of rate reduction on intelligibility quantified by orthographic transcription [192]. Similarly, both intelligibility gains and losses following rate reduction were reported for a group of speakers with hypokinetic dysarthria of matching severity [140]. In another work, Van investigated the effects of a number of rate control techniques [165], which did not show an improved intelligibility with slower speaking rate at group level. However, five out of 19 showed a meaningful increase of more than 8% intelligibility with slower speaking rates. In a follow-up study, he examined the effect of rate control methods on intelligibility for a larger group of 27 speakers with different types of dysarthria [164]. Compared to his previous study, half of the participants showed a significant increase in intelligibility secondary to at least one rate control method. However, pooling the group results over all rate control methods showed an overall reduction of intelligibility when slowing down speaking rate. Another study showed the same trend with scaled intelligibility evaluation for speakers with dysarthria secondary to PD and MS [205]. A recent study also showed that sentence duration alone was not a contributing factor to intelligibility associated with slowed rate [27] for speakers with dysarthria secondary to MS. In conclusion, these studies focused on varying effects of rate reduction methods on intelligibility in speakers with various types of dysarthria and neurological diagnoses, and these studies did not have an agreement on the role of rate reduction on improving speech intelligibility.

Pauses

The pause is a part of speaking rate along with phoneme duration. Krause reported that when controlling the speaking rate in habitual and clear speech, the pause frequency and duration were nearly equivalent in both habitual and clear speech, which showed that pause frequency and duration are not contributing factors to increased intelligibility of clear speech [121]. However, the study only involved 18–29 aged listeners with normal hearing ability. It might be long pause duration benefits people with hearing loss or elderly (over 60) listeners. Further investigation on the topic is necessary but outside of the scope of this dissertation.

Amplitude Modulation

A variety of studies showed that temporal envelope is an important factor for speech intelligibility [51, 52]. Temporal envelope is a change in the amplitude and frequency of sound perceived by humans over time. The researchers reported that the amplitude modulation in the range between 4 Hz and 16 Hz is the most important for sentence intelligibility, and the amplitude modulation

as low as 2 Hz is important for phoneme identification [51, 52].

In another study, Liu examined the importance of temporal envelope and fine temporal structure on speech perception in *auditory chimera* speech [134], which showed that the temporal envelope contributed more to clear speech at a high signal-to-noise ratio, while the fine structure contributed more at a low signal-to-noise ratio. However, the auditory chimera was less intelligible than the original habitual speech, which showed a negative influence of processing artifacts on their findings. Similarly, modifying the temporal intensity envelope had detrimental effects on intelligibility because of processing artifacts [122]. As a result, minimizing the processing artifacts is necessary for a further investigation of amplitude modulation.

2.2.2 Spectral Features

Formant Frequencies

Hillenbrand investigated the important role of formant movement for speech intelligibility using naturally produced speech, as well as synthesized speech with either original formants or flat formants [86]. One of the findings is that synthesized speech with original formants had higher rates of vowel identification than signals with flat formants, which showed the importance of formant movement in vowel identification. Moreover, Smits reported that the formant transitions related to prevocalic voiced stops were more effective than the bursts of the same stops for stop identification [188]. They concluded that the important role of formant transition was highly dependent on the vowel context.

In addition, Turner showed the effect of lengthening formant transitions of the stop consonants on the synthesized syllables with hearing-impaired listeners [212], which showed that the stop identification rates increased rapidly when stretching the formant transitions of the stop consonants from 5 ms to 160 ms. The performance became close to perfect at a transition of 20 ms and longer for normal hearing listeners; however, not all hearing-impaired listeners benefited from longer formant transitions. Therefore, the advantage of lengthened formant transitions was limited by a listener's hearing loss [212].

In another study, Moon reported formant undershoot of the second formant frequency in vowels /i:/, /ɪ/, /ɛ/ , and /ei/ to be less dramatic in the clear speech style than in the habitual speaking style [149]. In CVC materials, the first and second formant frequencies of tense vowels reached their target frequencies and had less variance in clear than in habitual speech. In contrast, Krause found that the formant values extracted from the vowel midpoints were not closer to the formant target frequencies nor less variant in clear speech spoken at habitual speaking rates (*clear/normal*)

than in habitual speech [121]. They argued that the formant contour of clear/normal speech might have reached the formant target frequencies closer than the formant contour of habitual speech, and measurement at one time point might not be sufficient to capture the differences.

Many researchers have been interested in the relationship between vowel space and speech intelligibility. They showed that speakers with larger vowel spaces are more intelligible than speakers with reduced spaces [26, 77]. Specifically, the speakers who had wide F1 ranges (defined as the difference between F1 for /i:/ as in "easy" and F1 for /ɑ/ in "pot") tended to be more intelligible than speakers with a smaller F1 range. The F2 range (defined as the difference between F2 for /i:/ and F2 for /ɔ/) is significantly correlated with sentence intelligibility [26].

In another study, Ferguson concluded that steady-state formant values for back vowels, dynamic formant movement, and duration for front vowels were important cues for the vowel identities with young normal-hearing listeners [60].

Spectral Balance

In a variety of studies, speakers tend to raise vocal effort and overall energy to make speech more intelligible. The process of raising vocal effort correlated with increased values of F0 and formant amplitudes of F1, F2 and F3; moreover, the formant amplitudes in the higher range grew more than those in the lower range [130]. Krause found narrower formant bandwidths in clear speech than those in habitual speech at normal speaking rate of 200 wpm, which showed higher formant amplitudes in the short-term spectra of clear speech compared with habitual speech [121]. A raised energy in the 1–3 kHz frequency range of long-term average spectrum (LTAS) is significantly correlated with intelligibility [121, 77]. However, Hazan found that the slope of the LTAS did not correlate with intelligibility [77]. A recent study compared the short-term average spectrum (STAS) of clear and habitual speech [66], which showed that clear speech has higher energy in two frequency bands: [2000, 4800] and [5600, 8000]. In conclusion, the increased energy in 1) the frequency range of 1–3 kHz for LTAS, and 2) the two frequency bands: [2000, 4800] and [5600, 8000] for STAS were responsible for the improved intelligibility of clear speech.

Speaker Characteristics

A number of studies showed that speech quality significantly affects speech intelligibility. The variation in speaking style (normal, nasalized, child-directed, whispered, excited, and elongated) presented in a single block reduced word intelligibility relative to a single speaking style [189], which shows that the speech quality is relevant to word intelligibility. In other research, Hazan showed that less-intelligible speakers were perceived as sounding "mumble, unpleasant, muffled,

or weak", relative to the more intelligible speakers [77]. However, the study also showed that the quality dimensions of voice excitation (harsh/smooth, creaky/non-creaky, husky/not-husky) were not correlated with intelligibility [77].

Another important contributing factor of speech intelligibility is gender difference. Bradlow and Hazan found that female speakers were more intelligible than male speakers [26, 77]. Bradlow also observed that female speakers featured: 1) wider F0 range, 2) larger vowel space, 3) more precised inter-segmental timing, and 4) less frequent alveolar flapping relative to male speakers [26, 25]. However, it is unclear whether an intelligibility of 93.4% for female speakers (compared to an intelligibility of 81.1% for male speakers) could be attributed to one factor or a combination of these factors.

2.2.3 Combination of Features

The above studies showed high correlations between various acoustic features and speech intelligibility. Note that the high correlations do not necessarily imply causality. The acoustic causes of improved intelligibility of clear speech [95, 203] and slowed speech [27] were investigated using a hybridization technique (Figure 2.1). In this method, researchers prepare parallel data of clear and habitual speech. Then, they insert different components of clear speech into habitual speech to increase its intelligibility. The clear components can be duration (D), energy (E), intonation (I), and spectrum (S) (Figure 2.1). Researchers look for clear components that increase the intelligibility of habitual speech. Moreover, these studies also looked at the combination of features. Kain found that a combination of spectrum and duration was sufficient to improve sentence-level intelligibility of habitual speech for one speaker; while F0, energy, phoneme sequence, and pause information were not [95]. In a follow-up study, Tjaden investigated the acoustic variables explaining intelligibility variation for two speakers with dysarthria secondary to Parkinson's disease [203]. The results showed that a combination of clear spectrum and duration yielded a 13.4% improvement of transcription intelligibility; while only clear energy yielded 8.7% improvement, and only clear spectrum yielded 18% improvement [203]. In a recent study, Brenk showed that a combination of duration and short-term spectrum was an important contributing factor to the intelligibility changes of slow speech for people with dysarthria secondary to Multiple Sclerosis [27].

2.3 Voice Conversion

The voice conversion framework (VC) can be used for style conversion (SC); thus, SC is closely related to VC. VC is a process of transforming a source speaker's speech so it sounds like a target

speaker's speech. In other words, the speaker-related factors of speech is mapped between source and target speakers, while the linguistic content is preserved. Different from VC, SC techniques are used to map style-related factors of speech between source (e.g habitual) and target (e.g clear) speaking styles in order to improve the perceived intelligibility. Although a previous attempt in my research group in applying a VC technique for SC achieved only a modest result [145], the VC techniques are worthwhile to investigate for the task of SC.

In this section, I review the fundamentals of VC. Figure 2.2 presents an overview of a VC framework, which contains two phases: 1) *training phase*, and 2) *conversion phase*. During the training phase, parallel utterances, which contain pairs of utterances from source and target speakers with the same linguistic content, are prepared. In the *speech analysis* step, the speech waveform of the source and target utterances is converted into *speech features* (e.g fundamental frequency (F0), and spectrum) (Section 2.3.1); then, the speech features are further analyzed into *mapping features* (Section 2.3.2). The *time alignment* step (Section 2.3.3) aligns the sequences of mapping features between source and target speaker. Lastly, the *train mapping function* step produces a *mapping function* from the aligned mapping features (Section 2.3.4 and Section 2.3.5).

In the conversion phase, mapping features are obtained from a new utterance of the source speaker. During *map the features* step, converted features are calculated by applying the mapping function on the source mapping features. The final *speech synthesis* step produces an *converted utterance* from the converted features.

2.3.1 Speech Features

In general, a vocoder is responsible for 1) extracting speech features from a waveform in the initial speech analysis step, and 2) reconstructing a waveform from speech features in the final speech synthesis step. The performance of the vocoding (analysis and synthesis) process determines the best quality that VC or SC systems can achieve. Most analysis and synthesis techniques are frame-level (frame-by-frame), which splits speech signals into small, overlapped frames to ensure the statistical stationary of the speech features in the frames. The length of the frame can be either constant or relative to the pitch period of the signal (known as *pitch-synchronous analysis*).

Different vocoders have different assumptions on speech models, which can be classified into two main categories: 1) *source-filter* models and 2) signal-based models. The source-filter model assumes that an *excitation* signal (related to vocal cord movement and frication noise) passes through the vocal tract (represented by a *filter*) to produce speech signals. The excitation signal (or *source* signal) and the filter are assumed to be independent from each other. Two commonly used filter models are: 1) *all-pole* (e.g., linear predictive coding (LPC)) [13] and 2) *log-spectrum*

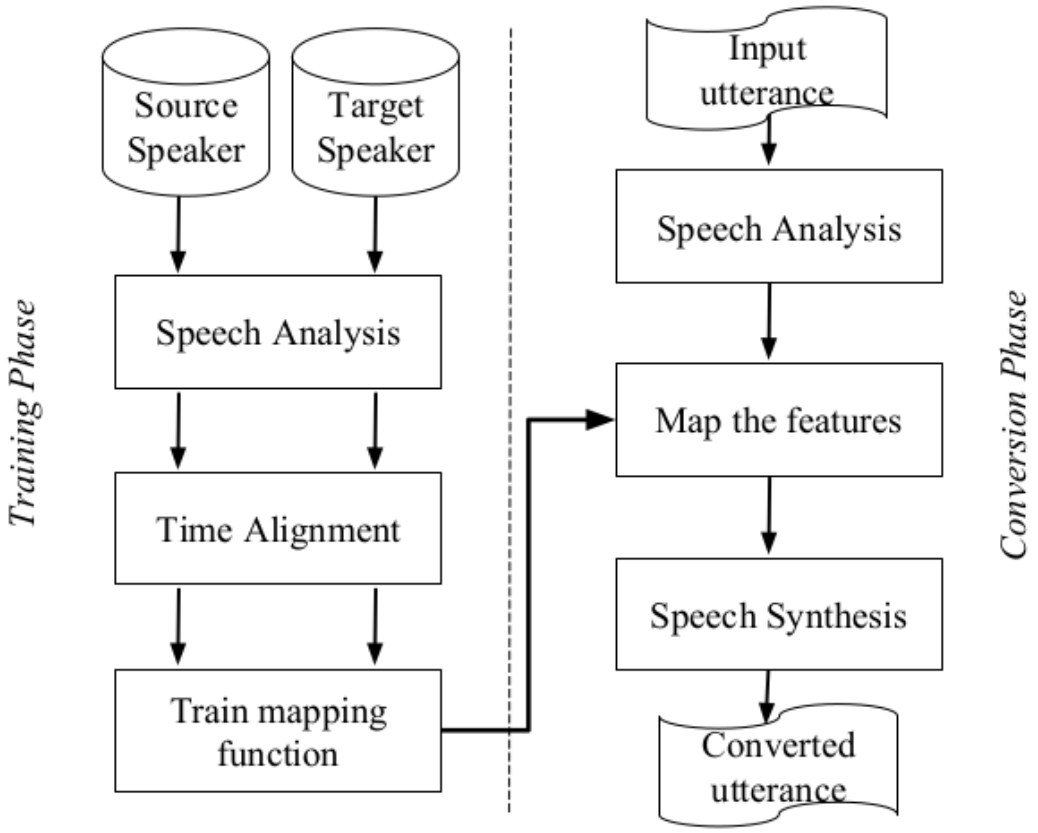


Figure 2.2: Voice conversion framework [56]

filters (e.g., mel-log spectrum approximation (MLSA) [91]). In the two models, a spectral envelop filter represents the vocal tract. Although the independence between source signal and filter is assumed, the pitch periods show up as sharp peaks and deep valleys (known as *harmonics*) in the spectral envelope. Clearly, the occurrence of pitch information in a spectrum violates the independence assumption between source signal and filter. In an attempt to alleviate the interference between signal periodicity and the spectrum, Kawahara proposed the STRAIGHT vocoder, which has a pitch-adaptive time-frequency spectral smoothing [106]. He updated it with the TANDEM-STRAIGHT vocoder in order to provide a unified computation of spectrum, fundamental frequency and aperiodicity [107]. However, Morise also proposed an improvement on STRAIGHT to address its inefficient spectral smoothing method, which involves calculating the short-term Fourier transform twice. He proposed the CheapTrick algorithm in WORLD vocoder [150, 152], which proved to be a more efficient spectral analysis and smoothing method than that of STRAIGHT. Note that the smooth spectrum is easier to model and manipulate.

Researchers calculated the excitation signal in a variety of rule-based methods including 1) a pulse/noise model using periodic pulse/noise for voiced/unvoiced speech segment, 2) glottal excitation models [36, 222], 3) residual signals [97, 195], 4) mixed excitation [167, 166], and 5) band aperiodicity [82, 34]. In another approach, the excitation signal was predicted using deep learning-based methods, including 1) LPCNet [219], 2) GlottDNN [4], and 3) GlottGAN [21].

In contrast to source-filter models, the signal-based models do not make any restrictive assumptions (e.g., independence of source signal and filter); therefore, they usually have higher quality. The disadvantage is that they are less flexible for modification. Some examples of the model are: 1) pitch-synchronous overlap-add (PSOLA) [157] using varying frame sizes related to pitch to create short frames of speech signal, 2) Linear Predictive PSOLA allowing simple vocal tract modifications [216], and 3) Harmonic plus noise models (HNM) decomposing speech signal into harmonics (sinusoids with frequencies relevant to pitch) [193].

2.3.2 Mapping Features

Typically, the speech features (e.g. spectrum) are not effective to manipulate with mapping functions in VC or SC; therefore, the speech features are further processed to calculate the mapping features that are more suitable for manipulation. I review the commonly used mapping features in the following.

Spectral envelope

Researchers used the logarithm of the magnitude spectrum as mapping features, which required more constraints for the VC mapping functions to work with the high-dimensional mapping features [217, 194]. To emphasize the perceptual information, the frequency scales were warped to Mel- or Bark-scale. Note that the features are highly correlated.

Cepstrum

A small set of cepstral coefficients can be used to represent a spectral envelope. To obtain the cepstral coefficients, a *discrete cosine transform* was applied on the logarithm of the magnitude spectrum. To emphasize the perceptual information, researchers warped the frequency scales of the magnitude spectrum on Mel-scale, which resulted in *mel-cepstrum coefficients* (MCEP) [90]. The cepstral coefficients are uncorrelated.

Line spectral frequencies

Line spectral frequencies (LSFs) are associated with frequency and formant structure. The features have better quantification and interpolation proprieties [169], which are preferred by statistical methods [94]. The frame-based LSF parameters monotonically increase; therefore, they are highly correlated.

Formants

Formant frequencies and corresponding bandwidths can be used to approximate magnitude spectrum [144, 229]. Due to the compactness of the features, the quality of synthesised speech is limited when modifying the formants.

Among the mapping features, mel-cepstrum coefficients and line spectral frequencies are the most commonly used features [148]. However, appropriate mapping features need to be both compact and interpolable, and thus ideally suited for regression approaches that involve averaging. Interpolability ensures that even when two or more parameter vectors are averaged, the result remains near the manifold of possible speech; this property does not hold for MCEPs, and line spectral frequencies. In Chapter 3, I examine novel spectral mapping features that satisfy both compactness and interpolability; and I use the MCEPs and line spectral frequencies as baselines to evaluate the novel mapping features.

2.3.3 Time-alignment

Typically, a parallel corpus of utterances are used in the training phase of VC systems, which consists of utterance pairs from source and target speakers with the same linguistic content. After extracting mapping features from the utterances, the sequences of source and target mapping features are aligned, which helps the source and target mapping features be equal in lengths. Commonly, a *dynamic time warping* (DTW) algorithm is used to obtain the best time alignment between each utterance pair [3, 98]. The alignment results in equally-long source and target sequences of mapping features. The DTW alignment assumes that the same phonemes of the speakers have similar features; however, the assumption is not always true and might result in sub-optimal alignments, since the speech features are not speaker-independent.

The impact of the frame-level alignment on the performance of VC mapping is investigated in a number of studies, especially when one frame aligns with multiple other frames[156, 81, 67, 146]. The findings showed that a combination of DTW and a simple voice activity detection (VAD) technique achieved a successful alignment [81]. Other studies reported to filter out the source-target training pairs that are unreliable, based on a confidence measure [211, 178]. The time-alignment process is typically independent from training mapping function between source and target features, which may lead to sub-optimal results of the two processes. Therefore, a recent study examined a sequence-to-sequence model with attention in learning the time-alignment and a mapping function, simultaneously [199].

2.3.4 Spectral Mapping

After time-alignment, the VC mappings must be learned to represent the relationship between the source and target spectral (mapping) features. Given the source mapping features $\mathbf{X}^{train} = \left[\mathbf{x}_1^{train}, ..., \mathbf{x}_N^{train}\right]$ and target mapping features $\mathbf{Y}^{train} = \left[\mathbf{y}_1^{train}, ..., \mathbf{y}_N^{train}\right]$, $\mathbf{x}^\top = (x_1, ..., x_D)$ and $\mathbf{y}^\top = (y_1, ..., y_D)$ are D-dimensional vectors, the goal of the training stage is to build a mapping function $\mathscr{F} : \mathbf{Y}^{train} = \mathscr{F}\left(\mathbf{X}^{train}\right)$. At conversion time, an unseen source features $\mathbf{X} = [\mathbf{x_1}, ..., \mathbf{x}_{N^{test}}]$ of length N^{test} is transformed by the mapping function $\mathscr{F}$ into estimated target features $\hat{\mathbf{Y}}$ (as in Equation 2.1).

$$\mathscr{F}(\mathbf{X}) = \hat{\mathbf{Y}} = [\hat{\mathbf{y}}_1, ..., \hat{\mathbf{y}}_{N^{test}}] \tag{2.1}$$

Generally, the mappings are performed frame-by-frame, which means that each frame is mapped independently of other frames $\hat{\mathbf{y}} = \mathscr{F}(\mathbf{x})$. To model the time-dependence of speech frames, recent models often consider more context to go beyond the frame-by-frame mapping.

Codebook Mapping

Abe used vector quantization (VQ) to reduce the number of source-target pairs in an optimized way [3]. During the training phase, he formed a *codebook* with M code vectors using *hard clustering* on source and target features separately. The code vectors were denoted as $\mathbf{c}_m^x$ and $\mathbf{c}_m^y$ for source and target features, for m=$[1, ..., M]$ respectively. At conversion time, the closet centroid vector of the source codebook $\mathbf{c}_m^x$ was retrieved and the corresponding target codebook $\mathbf{c}_m^y$ was selected.

$$\mathscr{F}_{\mathrm{VQ}}\left(\mathbf{x}\right) = \mathbf{c}_m^y \tag{2.2}$$

where $m = \arg_{\eta=[1,...,M]} \min d\left(\mathbf{c}_\eta^x, \mathbf{x}\right)$.

The advantage of the VQ approach is its compactness due to the use of clustering approach to determine the codebook. The disadvantage is the discontinuity of generated feature sequences. The disadvantage was alleviated by using a large codebook, which requires more parallel data. The quantification error was reduced using fuzzy VQ, which utilizes soft clustering [184, 12, 211]. Given an unseen source feature x, a continuous weight $w_m^x$is computed for each codebook using a weight function. The mapped feature is a weighted sum of the centroid vectors

$$\mathscr{F}_{\text{fuzzy VQ}}(\mathbf{x}) = \sum_{m=1}^{M} w_m^x \mathbf{c}_m^y \tag{2.3}$$

where $w_m^x = \mathrm{weight}(\mathbf{x}_m^x, \mathbf{c}_m^y)$.

The weight function was calculated using a variety of methods: 1) Euclidean distance [184], 2) phonetic information [186], 3) exponential decay [11], 4) vector field smoothing [76], 5) statistical approaches [129]. The traditional VQ is a special case of fuzzy-VQ, where only one centroid has a weight of one, and the rest have zero contribution.

Mixture of Linear Mappings

Instead of using centroid vectors, Valbret proposed the *linear multivariate regression* (LMR) which linearly transforms source mapping features $\mathbf{x}$ into target mapping features [217].

$$\mathscr{F}_{LMR}\left(\mathbf{x}\right) = \mathbf{A}_m\mathbf{x} + \mathbf{b}_m \tag{2.4}$$

where $m = \arg_{\eta=[1,...,M]} \min d\left(\mathbf{c}_\eta^x, \mathbf{x}\right)$, and $\mathbf{A}_m$ and $\mathbf{b}_m$ are regression parameters. Similar to VQ, the disadvantage of the approach is the discontinuity in the predicted features when the clusters change between neighboring frames. Inspired from fuzzy-VQ, the linear regression was updated to solve the discontinuities

$$\mathscr{F}_{weighted\,LMR}\left(\mathbf{x}\right) = \sum_{m=1}^{M} w_m^x \left(\mathbf{A}_m\mathbf{x} + \mathbf{b}_m\right) \tag{2.5}$$

where $w_m^x = \text{weight}(\mathbf{c}_m^y, \mathbf{x})$.

To estimate the parameter of the mapping function, Kain proposed a joint-density of the source-target mapping feature vectors, called *joint-density Gaussian mixture model* (JDGMM) [97]. A joint feature vector $\mathbf{z}_t = \left[\mathbf{x}_t^\top, \mathbf{y}_t^\top\right]$ is created from a source mapping feature $\mathbf{x}_t$ and a target mapping feature $\mathbf{y}_t$. He then fits a Gaussian mixture model (GMM) to the joint data. A known issue of GMM-based mappings is to generate speech with muffled quality. The reason is that generated features are averaged; resulting in wide formant bandwidths in the converted spectra. This problem is also known as over-smoothing, because the converted spectral envelops are smoothed. Post-processing techniques were used to compensate for the over-smoothing issue [207, 208, 196].

Neural Network Mapping

Typically, the association between source and target mapping features is complex and not linear. To represent the non-linear relationship, researchers used *artificial neural networks* (ANNs). ANNs have a number of neurons which are grouped into multiple layers. An ANN performs a non-linear mapping function with the form of $\mathbf{y} = f\left(\mathbf{W}\mathbf{x} + \mathbf{b}\right)$ where f is called the activation function (e.g., sigmoid, tangent hyperbolic, rectified linear units, or linear function). ANNs have two (or more) layers, which is defined as

$$F_{ANN}\left(\mathbf{x}\right) = f_2\left(\mathbf{W}_2 f_1\left(\mathbf{W}_1\mathbf{x} + b_1\right) + b_2\right) \tag{2.6}$$

where $\mathbf{W}_i$, b_i, f_i represents the weight, bias and activation function for the i^{th} layer, respectively.

ANNs with more than two layers are called *deep neural networks* (DNNs). The sizes of initial input and final output layers are constant, which depends on the size of source and target mapping features. However, the sizes of the intermediate layers are empirically decided.

Narendranath attempted to use ANNs to map formant frequencies [161]. Later, Makki used *principal component analysis* to calculate a compact representation of speech features as mapping features [136]. Desai investigated the performance of a three-layer ANN in mapping mel-cepstral features [42]. Moreover, a variety of ANN architectures was examined for VC mapping including: 1) Feedforward architecture [42, 14], 2) restricted Boltzmann machine [33], 3) joint architectures [33, 147], and 4) recurrent architectures [160]. Generative models were investigated in VC such as 1) Generative adversarial network (GAN) models [102, 103] and variational autoencoders [19, 87]. To capture long-term dependency information, Xie used sequence error minimization instead of frame error minimization [227]. Another way to model time dependency is to use RNNs, which implicitly model temporal behavior by looking at information from previous input frames in addition to current input frames [159, 160]. Kameoka investigated convolution *sequence-to-sequence* (seq2seq)

models in VC mapping, which convert not only spectral aspects but also prosodic aspects of source speech, simultaneously [100]. In a follow-up study, Tanaka improved the performance of the seq2seq models using attention models [199].

Frequency Warping Mappings

Motivated by the differences of formant frequencies, formant bandwidths and energies in each frequency band between different speakers, researchers focused on the manipulation of formant location, formant bandwidths and energy in certain frequency bands. The advantage of this approach is to accept high-dimensional mapping features (e.g., harmonic vocoders), which provide higher speech quality compared to more compact vocoders (e.g., LSF vocoders). Specifically, the conversion of source spectral features into target spectral features is achieved by warping the frequency axis to adjust formant location and bandwidth, and then adjusting the energies in each frequency band [57, 58].

There have been a variety of attempts to use frequency warping mappings in VC. Valbret conducted the mapping directly on log-spectral features, which subtracted the source spectral tilt before warping, and added the target spectral tilt after warping to the log-spectrum [217]. In another study, source formant frequencies and bandwidths were converted to match target values [144, 210]. Other researchers investigated a number of *vocal tract length normalization* (VTLN) techniques. Sundermann examined piece-wise linear, power, quadratic, and bi-linear VTLN functions [194]; while Morley estimated VTLN parameters using an iterative algorithm [154].

In addition to formant information, the average energy of spectral bands is an important factor of speaker individuality. Typically, researchers subtracted the source spectral tilt before the frequency warping, then they added the target spectral tilt to take care of the average energy. In another study, Tamura applied a simple amplitude scaling to shift the average energy of source speech [198]. In a different approach, Godoy combined frequency warping and amplitude scaling to add more degrees of freedom to the mapping [68, 69].

There have been numerous extensions of the frequency warping mappings such as in combination with GMMs [55, 230], dictionary-based methods [185, 215], maximizing spectral correlation [202], equalizing formant frequencies as preprocessing step [146], and exemplar-based approach [201].

2.3.5 Prosodic Modeling

Although prosodic features (pitch, duration, spectral balance, energy) are important for speaker identity [83, 154], the majority of VC literature focuses on spectral mappings. For duration modelling, *decision tree* [39] and duration-embedded hidden Markov models [226] were investigated.

The commonly used pitch transformation is a linear transformation, which globally shifts the source pitch contour using average and standard deviation of the source speaker' s F0 and target speaker's F0.

$$\hat{F0}^y = \frac{\sigma^y}{\sigma^x}(F0^x - \mu^x) + \mu^y \tag{2.7}$$

where μ and σ represent mean and standard deviation of the log-scaled F0, $F0^x$ denotes F0 of the source speaker, and $\hat{F0}^y$ denotes predicted F0 of the target speaker [31].

In another approach, the spectral features and F0 were modelled jointly, which improved both spectral and F0 conversion [54, 75]. In yet another study, the new F0 values were calculated from a converted spectrum using GMM [53].

2.4 Automatic Approaches for Improving Intelligibility

An approach to increase the intelligibility of speech in adverse environments is to use noise suppression and cancellation [138, 32, 113, 224, 143, 218, 92]. In this approach, noisy speech signal is processed to emphasize the speech components and alleviate the noise components (see Section 1.1.2). The approach requires noise cancellation devices (e.g., noise cancelling headphones) for users, which takes as input a noisy speech signal and outputs an enhanced signal with higher intelligibility and quality.

However, noise cancellation devices are not always available for users. Another approach is to alter the speech signal prior to presentation in a noisy environment; these techniques can be classified into several categories, including utilizing audio and signal properties such as amplitude compression [163], dynamic range compression [20, 22], peak-to-rms reduction [177], and formant-enhancement [28]. Other techniques exploit the knowledge of a noise masker such as optimizations based on a speech intelligibility index [180] or glimpse proportion measure [197, 200].

There are also techniques that consider the intelligibility gains due to a clear speaking style [172, 60, 59], inspired by the acoustic characteristics of clear speech such as spectral flattening and vowel space expansion [8, 66]. In this section, I review the intelligibility improvement techniques that modify speech signals prior to presentation in noisy environments. The techniques can be classified into two main groups: 1) rule-based techniques, and 2) statistical techniques.

For rule-based techniques, Gordon showed that elderly listeners have difficulty processing brief consonant cues such as burst portions of stops [74]. Therefore, they increased the intensity of the consonants in consonant-vowel (CV) syllables to improve the consonant identification rate. Successful results were achieved by intensifying the consonant energy in CV and vowel-consonant-vowel (VCV) sequences for normal hearing listeners [71, 77] and hearing-impaired listeners [72]. Hazan adjusted the degree of amplification of the burst and aspiration to improve the intelligibility of nonsense sentence materials [77]. In another approach, intensifying the CV ratio (known as energy redistribution) using voiced/unvoiced information as well as increasing the spectral energy center of gravity by high-pass filtering increased monosyllabic word intelligibility over unmodified speech [187]. However, Gordon showed that lengthening only consonant duration in CV syllables was unable to affect intelligibility for normal-hearing listeners aged 65–72 [71]. In the same study, consonant identification was improved by a combination of intensifying the consonant energy and lengthening consonant duration.

It was found that extending pause duration for people with hearing loss [214], and inserting additional pauses between words in a sentence for both young and elderly people with or without hearing loss [73] did not benefit intelligibility of meaningful sentences. In contrast, Liu reported an 13% absolute improvement of intelligibility by inserting pauses between words, and normalizing root-mean-squared (RMS) energy of the speech [134]. However, the author did not exclude pauses when calculating RMS energy, which resulted in increased energy as well as an increased signal-to-noise ratio of the speech. Therefore, the absolute improvement could be attributed to the increased signal-to-noise ratio instead of the inserted pauses.

Other studies focused on modifying the modulation of spectral envelope. Narne increased the depth of modulation of the speech envelope by 15 dB, which resulted in an improved consonant identification rate by listeners with auditory neuropathy [162]. Kusumoto showed that modulation enhancement of the temporal envelope from 1 to 16 Hz improved consonant recognition rates by 6% points in a reverberant condition with normal hearing listeners [124]. However, the word-level success of the modulation enhancement techniques does not necessarily transfer to sentence-level success [78].

Inspired by the acoustic changes of clear speech, Godoy combined spectral shaping and dynamic range compression to expand the vowel space, which resulted in improved intelligibility in several signal-to-noise ratios [66]. The disadvantage of the technique is the trade-off between naturalness and intelligibility. In fact, the quality of their modified speech degraded perceptually because their spectral modification could not be achieved by humans. Moreover the simple modifications may overlook important features that impact speech intelligibility.

In addition to rule-based techniques, researchers attempted to use statistical techniques for intelligibility improvement. Kain attempted to use Gaussian mixture model (GMM)-based voice conversion to transform the vowels of speaker with dysarthria to closely match the vowel space of a non-dysarthric (target) speaker, which resulted in improved intelligibility of dysarthric vowels of one speaker [96]. Previously, hybridization was utilized to investigate the acoustic causes of improved intelligibility in clear speech [95, 203] (see Section 2.2), which reported that it should be possible to automatically increase the intelligibility of speech by learning a mapping between habitual and clear features, or SC. However, a previous mapping experiment, which utilized a GMM, only showed very modest improvements, and was conducted only on vowels [145]. The mappings can be limited by: 1) inappropriate mapping features, and 2) over-smoothing problem of the mapping techniques [207], which introduced artifacts to the modified speech as well as degrading the naturalness of modified speech. I address the limitation of mapping features in Chapter 3; and I address the over-smoothing issue of mapping techniques in Chapter 4.

2.5 Speech Intelligibility Assessment

Speech intelligibility assessment focused on the performance of listeners when speech is presented in noisy environments [116]. Specifically, researchers evaluated speech intelligibility in different levels, including 1) phoneme-level intelligibility such as phoneme identification accuracy [51, 52], vowel identification accuracy [212, 86], stop consonant identification accuracy [188], consonant identification accuracy [71, 74, 162], 2) word-level intelligibility such as word identification accuracy [189], and 3) sentence-level intelligibility such as sentence transcription accuracy [205], *keyword recall accuracy* [27]. It's important to note that phoneme-intelligibility cannot be used to predict sentence-level intelligibility [7]. Additionally, Horvitz evaluated speech intelligibility using a *intelligibility preference test* [9]; specifically, they evaluated speech intelligibility of a system relative to another system. In Chapter 3, 4, and 6, I utilize the keyword recall accuracy test to evaluate the performance of spectral style conversion. In Chapter 5, I utilize the intelligibility preference test.

2.6 Conclusion

In this chapter, I reviewed the studies on 1) the acoustic differences between habitual and clear speech, 2) the contributions of acoustic features to speech intelligibility, 3) fundamentals of voice conversion techniques, 4) speech modification techniques for intelligibility improvement, and 5) evaluation methods for speech intelligibility. The spectrum and duration proved to be important

contributing factors for increased intelligibility of clear speech. Researchers attempted to convert the spectral envelope of habitual vowels to closely match the spectrum of clear vowels, which resulted in modest results. The modest results can be attributed to 1) inappropriate acoustic features for mappings, and 2) over-smoothing problem of the mapping techniques, which introduced artifacts to the modified speech as well as degraded its naturalness. I will address these two limitations in the upcoming chapters in order to improve the performance of SC mappings.

Chapter 3

Spectral Features for Voice and Style Conversion

In this chapter, I contrast two new sets of spectral mapping features: 1) probabilistic peak tracking (PPT) features, which are formant-like hand-crafted features (Section 3.1), and 2) manifold features, which are machine learnable by a Variational Autoencoder (Section 3.2).[1] The two sets of features are integrated into an existing high quality vocoder, WORLD. I extensively evaluate the two sets of features by comparing them to each other and to two baselines in three different tasks: speech reconstruction (Section 3.3), voice conversion (Section 3.4), and style conversion (Section 3.5). The baselines are two commonly used spectral representations: line spectral frequency and mel-cepstrum coefficients.

3.1 Probabilistic Peak Tracking Features

Motivated by the importance of formant frequencies and formant bandwidths on speech intelligibility (see Section 2.2), I propose formant-like hand-crafted features that are the frequencies of the peaks in the magnitude (energy) spectrum. Moreover, I assume that the peak frequencies change slowly and continuously over time, which ensures the smoothness of the peak frequency contours. This assumption, however, sometimes causes the peak frequency contours not to pass through spectral peaks. Therefore, peak bandwidths are used to represent the presence or absence of magnitude peaks: a wide bandwidth represents the absence of a peak at that frequency, while a narrower bandwidth represents the presence of a peak. As a result, a spectrum is represented by a small number of peaks frequencies and corresponding peak bandwidths.

[1] In this chapter, Section 3.1 was part of my qualifying exam and is not published. The remaining sections are based on a paper published in Interspeech [44], *Using a Manifold Vocoder for Spectral Voice and Style Conversion*, Tuan Dinh, Alexander Kain, Kris Tjaden.

Specifically, I use nine peak frequencies to represent the speech spectrogram. These nine peak frequencies consist of one *glottal formant frequency* [23], four peak frequencies in the lower frequencies and the other four in the higher frequencies. The glottal formant frequency is linear proportional to fundamental frequency [47]. In vowels, such as /a/, the four low-band peak frequencies are important because they are intended to capture the first four formants. In some consonants, the low-band peak frequencies are not as informative as high-band peak frequencies. For example, in frication such as /s/, the low band frequencies are not informative; while, the high band frequencies are informative. When a peak bandwidth is very wide, the peak frequency is probably not informative.

In the rest of this section, I calculate an initial estimation of peak frequencies using the statistic of formant frequencies in each phoneme category (Section 3.1.1). I examine how the formant frequencies change with respect to the change in the spectrogram; I want my peak frequencies to change as rapidly as the formant frequencies (Section 3.1.2). Using the initial estimation of peak frequencies and how rapidly peak frequencies change over time, I compute the nine peak frequencies on a spectrogram (Section 3.1.3). One can improve the results of the primary tracking using a secondary tracking (Section 3.1.4). After obtaining the peak frequencies, I estimate their corresponding peak bandwidths to reconstruct the spectrogram (Section 3.1.5). I present how to integrate the PPT feature in a vocoder (Section 3.1.6).

3.1.1 Initial Peak Frequencies

In this section, I estimate initial values for the peak frequencies. The initial peak frequencies add a global constrain that will be used by the tracker in Section 3.1.3.

I used the WORLD magnitude spectrogram [152]. The spectrogram is smoothed in both time and frequency domain; and the smoothness reduces spurious peaks created by pitch and harmonics. It brings an advantage for peak tracking based on the magnitude spectrogram when the dominant peaks become clearer than those peaks created by harmonics.

I first calculate the histograms for each formant frequency in each phoneme category in a TIMIT formant database [41], which provides phoneme labels, phoneme boundaries, and the first four formant frequencies F1–4. Figure 3.1 shows the histogram of the formant frequencies for the phoneme /a/ and the corresponding modes. I select the mode of the histogram of each formant frequency for each phoneme category as the initial estimate of the four low-band peak frequencies. For the initial estimate of the four peak frequencies in the high band, I do not have formants for them in the database. Instead, I select equally spaced values (5000, 6000, 7000, 8000 Hz). Finally, I add an initial estimate of glottal formant frequency at 200 Hz. As a result, each spectrum in

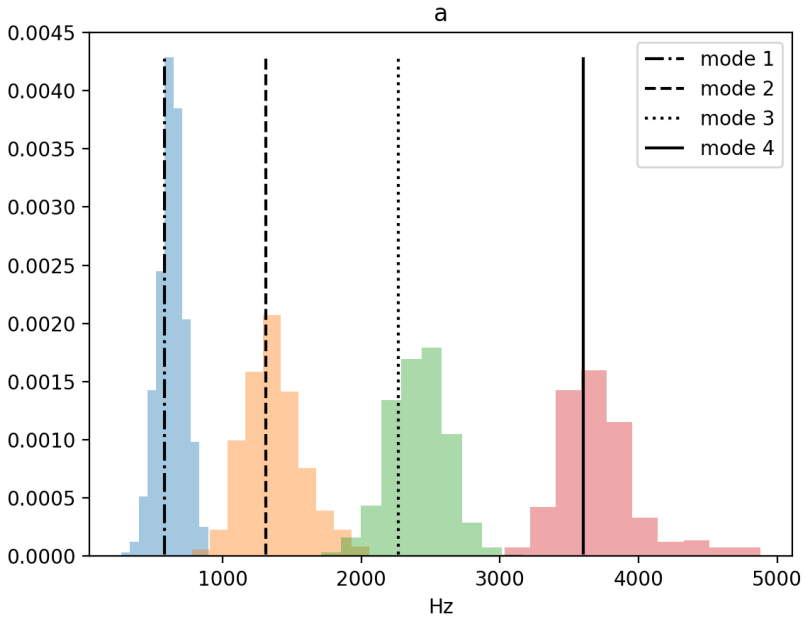


Figure 3.1: Histogram of the first 4 formant frequencies F1–4 of phoneme /a/ with 4 modes

a phoneme category k is represented by a 9-dimensional vector $\mathbf{m}_k$ consisting of four low-band frequencies, four high-band frequencies, and one glottal formant frequency. During testing, given a phoneme label, one can retrieve the nine formant frequencies and use them as the *initial peak frequencies*.

In real-life applications, however, the phoneme labels are not always available. Therefore, I built a phoneme classifier. The classifier receives as input acoustic feature $\mathbf{X}_t$ and outputs the posterior probability $p(k|\mathbf{X}_t)$ of $\mathbf{X}_t$ belonging to phoneme category k.

The classifier is based on the work of Song [223], a convolution neural network (CNN)-based phoneme classifier.[2] The network's structure (see Figure 3.2) consists of four convolution layers with sizes 128, 256, 384, and 384. The first two convolutions layers have max pooling of size (2×2). The filter sizes of the convolution layers are (3×5), (3×5), (3×3), (3×3), respectively. I select padding schemes for the convolution layers as *valid*, *valid*, *valid*, *same*, respectively. The convolution layers are dedicated to extract meaningful features from a time-frequency representation of speech. A speech spectrum is represented as 32 log-Mel filter-banks. I stack 12 frames before and after the current frame to make a 25 $\times$32 input image of the spectrogram. As a result, the input

[2]My implementation differs from his in that I used one less convolution layer and one less dense layer. I also specify the padding scheme and early stopping.

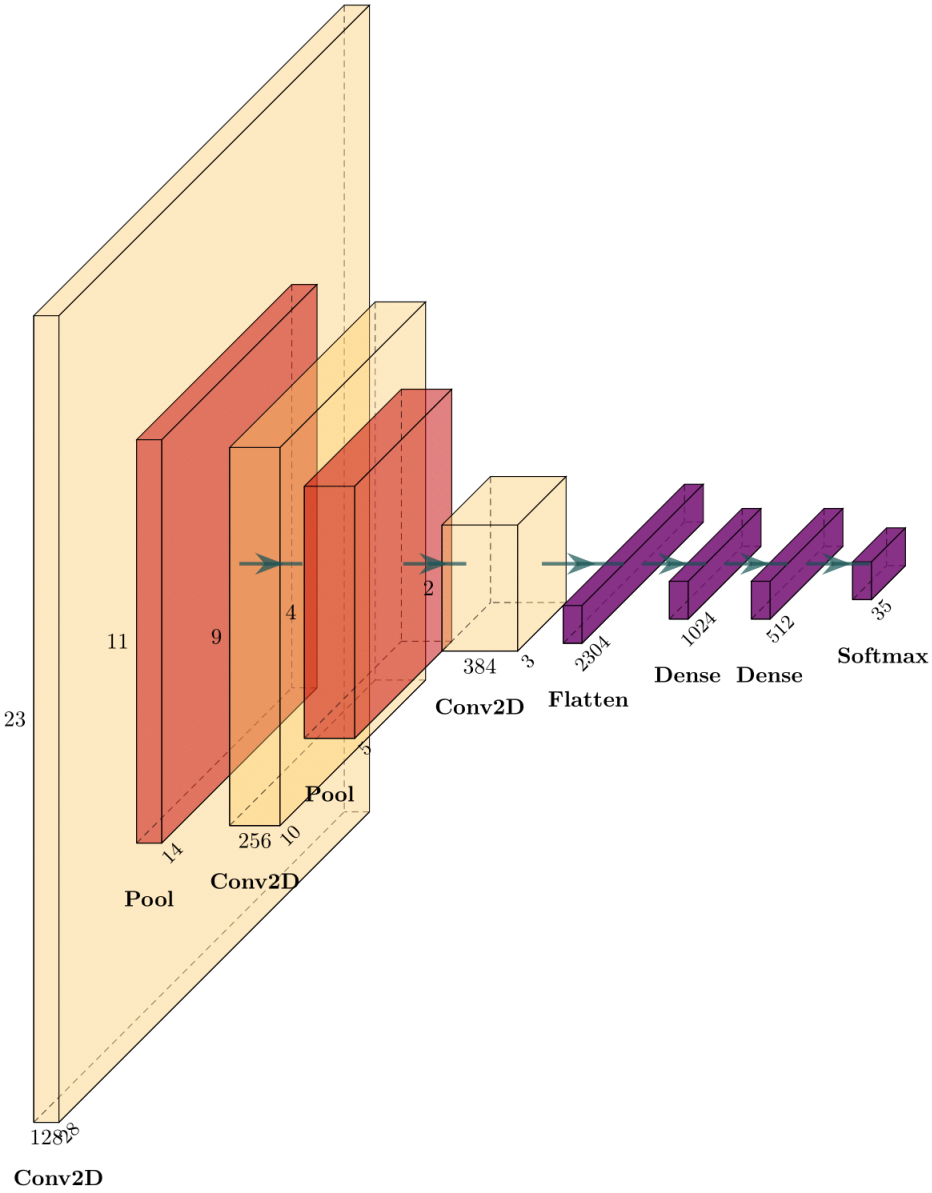


Figure 3.2: CNN architecture for phoneme classification

size for the CNN is (25×32). For the frames near the beginning or the end of sentences, I copy the first or the last frame, respectively. The convolution layers are followed by 2 densely connected layers with the sizes 1024 and 512, and finally a sigmoid layer for phoneme classification, generating a phonetic posterior-gram. Parametric ReLU is used as activation functions for all layers. To help parametric ReLU work effectively, batch normalization layers are applied before parametric ReLU. To avoid over-fitting, dropout layers are applied after the convolution and dense layers with a dropout rate of 20%. The Adam optimizer [114] is used to optimize the categorical cross-entropy function. Early stopping is used to stop the training process when there is no progress in the validation set.

I trained the classifier on the TIMIT formant database, which has 35 phoneme categories [41]. The 630 speakers in TIMIT are divided into train/validation/test as 462/144/24 designated speakers. I eliminate the spoken dialect samples for all speakers. I use a data generator to train the model on one sentence at a time. The frame error rates for validation and test were 21% and 22%, respectively, which is very close to the state-of-the art performance [223].

Using the phoneme classifier, the initial peak frequencies of each spectrum can be obtained as weighted sum of posterior probability $p(k|\mathbf{X}(t))$ and the nine formant frequencies of phoneme category k as follows:

$$\hat{\mathbf{peak}}(t) = \sum_{k=1}^{K} \mathbf{m}_k \cdot p(k|\mathbf{X}(t)) \tag{3.1}$$

where $\mathbf{m}_k$ is the mode (or the basic vector) of formant frequencies of class k, and p is the posterior probability that an acoustic vector $\mathbf{X}(t)$ at time t belongs to class k. The Equation 3.1 says that the values $\hat{\mathbf{peak}}(t)$ does not change when two frames are in the same phonetic classes k. It happens at the middle of a phoneme when the evidence of the phoneme class $p(k|\mathbf{X}(t))$ is high. When those frames are at the transition of two phoneme classes, the likelihood of previous phoneme fades out while that of the next phoneme fades in. It creates a smooth transition of peak frequencies between two phoneme classes. During the peak tracking, I assume peak frequencies on a spectrum should not be far away from the initial peak frequencies. An example of the initial peak frequencies are shown as dashed blue lines in Figure 3.3.

3.1.2 Spectral and Formant Frequency Change

In this section, I use the TIMIT formant database to examine the relationship between how the spectrogram changes from one time point to the next and how the formant frequencies change as well. The results of the analysis will be used in the next section when I track the peak frequencies.

To represent the spectral change, I define the average absolute difference between the features

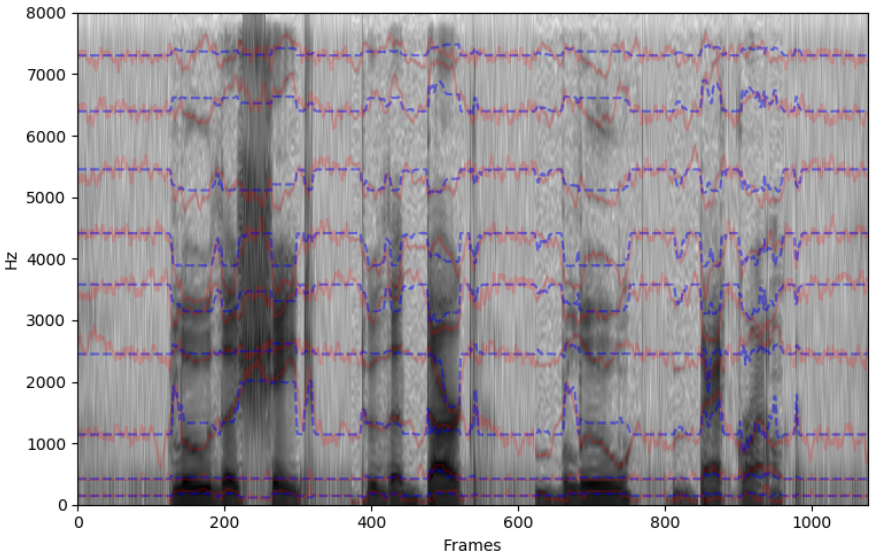


Figure 3.3: Initial peak frequencies (dashed blue lines) and final PPT results (solid red lines)

of two consecutive *non-overlapping* normalized (i. e. means are subtracted from) log-spectra as follows:[3]

$$\mathrm{SC}(t) = \frac{1}{256}\sum|\mathbf{SPEC}(t+2) - \mathbf{SPEC}\,(t)\,| \tag{3.2}$$

in which, the scalar $\mathrm{SC}(t)$ is the spectral change at time t; $\mathbf{SPEC}(t)$ and $\mathbf{SPEC}(t+2)$ are log-spectrum from 0–4 kHz at time t and $t+2$, respectively; $\mathbf{SPEC}(t)$ has a length of 256. The reason for using a length of 256 is that I apply a Fourier transform with a length of 1024 on 16 kHz speech signals on my corpus; therefore, a spectrum has a length of 256 to represent 4 kHz.

To represent the formant change, I also define the average first-order difference between first four formant frequencies F1–4 as the formant change (FC) as follows:

$$\mathrm{FC}(t) = \frac{1}{4}\sum|\mathbf{formant}(t) - \mathbf{formant}(t-1)| \tag{3.3}$$

in which the scalar $\mathrm{FC}(t)$ is the formant frequency change at time t; $\mathbf{formant}(t)$ and $\mathbf{formant}(t-1)$ are the vectors of the first four formant frequencies at time t and $t-1$, respectively; the formant frequencies are the first four formant frequencies available in TIMIT dataset (Section 3.1.1).

For each audio file in the TIMIT dataset, I calculate the SC and FC. I fit a linear model $\mathrm{FC} = \alpha \times \mathrm{SC}$ with intercept set to 0 to obtain a linear coefficient (slope) $\alpha = 0.7$ as in Figure 3.4.

[3] I calculated the difference between frames at time t and $t+2$, since two consecutive frames are 50% overlapped.

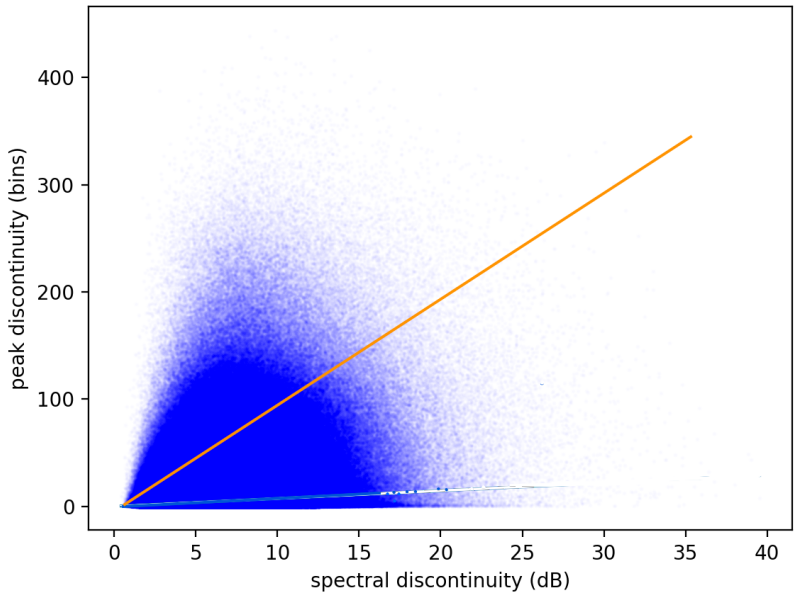


Figure 3.4: distribution contour (orange line) of spectral change (spectral discontinuity) SC and formant change (peak discontinuity) FC

I assume the peak frequency PC change is the same as the formant frequency change FC. It does make sense to apply alpha to the lower four, but I use this for all nine peak frequencies. In testing, with a given spectral change $\mathrm{SC}(t)$ at time t, I calculate the peak frequency change $\hat{\mathrm{PC}}(t)$ at time t, which represents how rapidly the peak frequencies shift, as follows

$$\hat{\mathrm{PC}}(t) = 0.7 \times \mathrm{SC}(t) \tag{3.4}$$

3.1.3 Primary Peak Tracking

In this section, I calculate peak frequencies using a probabilistic peak tracking procedure taking as inputs a spectrogram, initial peak frequencies (see Section 3.1.1) and how rapidly the peak frequencies change (see Section 3.1.2). The peak tracking procedure involves an *observation probability* of a peak frequency having a value of f Hz, a *transition probability* of a peak frequency to change from one point to another point, and the Viterbi algorithm.

First, I calculate the observation probability. I consider a normalized WORLD log-spectrogram **norm-spec** as the probability of a frequency f being a peak frequency:

$$p\left(f \text{ is a peak frequency} | \textbf{norm-spec}(t)\right) = \textbf{norm-spec}\left(t, f\right) \tag{3.5}$$

where $1 \leq k \leq 9$. I assume that the initial peak frequencies $\mathbf{peak}_k(t)$ follows a normal distribution with the initial peak frequencies $\hat{\mathbf{peak}}_k(\mathbf{X}(t))$ as means;

$$p\left(\mathbf{peak}_k\left(t\right)|\mathbf{X}(t)\right) \sim \mathcal{N}\left(\mathbf{peak}_k\left(t\right); \hat{\mathbf{peak}}_k(t), \sigma\right) \tag{3.6}$$

of 500 Hz. I do not use a pre-defined, speaker-independent range to search for a peak frequency. I define an observation probability of a peak frequency having a value of f as follows:

$$p\left(\mathbf{peak}_k\left(t\right) = f\right) = p\left(f \text{ is a peak frequency}|\mathbf{norm} - \mathbf{spec}(t)\right) \times p\left(\mathbf{peak}_k(\,t\,)|\mathbf{X}(t)\right) \tag{3.7}$$

Second, I calculate the transition probability. Recall that I assume that the peak change $\hat{\text{PC}}$ of peak frequencies from one point to another is not high and it should be proportional to the change SC on a spectrogram. I define the *transition probability* of peak frequencies $\mathbf{peak}_k$ from time t to $t-1$ with $1 \leq k \leq 9$ as follows:

$$p\left(\mathbf{peak}_k(t)|\mathbf{peak}_k(t\text{–}1)\right) \sim \begin{cases} 1 & |\mathbf{peak}_k(t) - \mathbf{peak}_k(t\text{–}1)| < \hat{\text{PC}}(t) \\ \mathcal{N}\left(|\mathbf{peak}_k(t) - \mathbf{peak}_k(t\text{–}1)|; \hat{\text{PC}}(t), 500Hz\right) & otherwise \end{cases} \tag{3.8}$$

where $\hat{\text{PC}}(t) = 0.7 \times \text{SC}(t)$ (see Equation 3.4).

Third, with the above observation and transition probability, I use the Viterbi algorithm to track the peak frequency trajectories individually except I track the glottal formant frequency and the first peak frequency jointly because the two frequencies are close to each other. I first track lower frequencies, then higher ones. A result of the probabilistic peak tracking is shown as solid red lines in Figure 3.3.

3.1.4 Secondary Peak Tracking (Optional)

A speech spectrum could have more than nine peak frequencies. By tracking more peaks (which I referred to as secondary peaks), the spectrum can be represented more precisely. The secondary peak tracking is optional because primary peaks are probably sufficient in modeling speech spectrum. The secondary peaks could be tracked in a secondary path and forced to be between primary peaks. I do not explore this in this work.

3.1.5 Bandwidth Computation

In this section, I explain how I compute the peak bandwidths. Peak bandwidth is computed in an iterative process so that the computed peak bandwidth and peak frequencies can best reconstruct the original spectrum using an all-pole model.

To calculate the range of the peak bandwidths, I use the widely-used formula of formant bandwidths:

$$\text{bw} = \frac{-\log r}{\pi} \cdot F_s \tag{3.9}$$

where F_s is the sampling rate, r is the pole location; $r \in [0, 1]$. As a result, I have bw $\in (0, +\infty)$.

I define a *loss function* that can be used in a iterative process for calculating bandwidths bw $\in (0, +\infty)$ as follows:

$$\frac{1}{512} \sum \mathbf{w} \left(\mathbf{ori}_t - g \cdot \mathbf{syn}_t\right)^2 \tag{3.10}$$

in which, $\mathbf{ori}_t$ is a original spectrum, $\mathbf{syn}_t$ is a frequency response of an all-pole filter with the filter coefficients calculated from the peak frequencies $\mathbf{peak}(t)$ (see Section 3.1.3) and the peak bandwidths $\mathbf{bw}_t$ (see Section 3.1.5) at time t. I normalize the 9-dimensional frequency vector $\mathbf{peak}(t)$ and I add 2 extra real poles (zero and one) to $\mathbf{peak}(t)$ as a requirement of my implementation of an all-pole filter.

The scalar g is used to equalize the root mean square energy (RMSE) of $\mathbf{ori}_t$ and $\mathbf{syn}_t$, and it is calculated as follows:

$$g = \frac{\text{RMSE}(\mathbf{ori}_t)}{\text{RMSE}(\mathbf{syn}_t)} \tag{3.11}$$

Moreover, I define an error weight vector $\mathbf{w}$ to emphasize the errors at the peak frequencies as follows:

$$\mathbf{w} = 0.5 + 0.5 \cdot \frac{\mathbf{ori}_t - \min(\mathbf{ori}_t)}{\max(\mathbf{ori}_t) - \min(\mathbf{ori}_t)} \tag{3.12}$$

I use the L-BFGS-B algorithm [29, 228] to search for the best values of bandwidth $\mathbf{bw}_t$ in $(0, +\infty)$ that minimize the loss in Eq. 3.10.

3.1.6 Integrating PPT Features into a Vocoder

I integrate PPT into the WORLD vocoder [152, 150], but it can be integrated into other vocoders (for more details about vocoders, see Chapter 2). First, I extract fundamental frequency, magnitude spectrogram, and aperiodicity using the WORLD vocoder. Second, the peak frequencies and corresponding peak bandwidths are computed from the magnitude spectrogram. Third, the magnitude spectrogram is reconstructed from peak frequencies and peak bandwidths using all-pole model. Finally, I synthesize speech from fundamental frequency, reconstructed spectrogram, and aperiodicity (Figure 3.5).

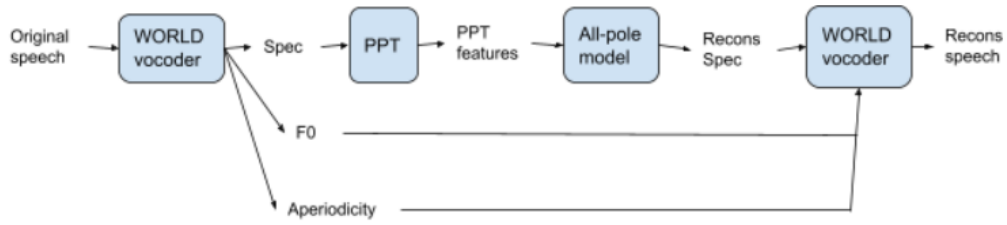


Figure 3.5: Integrate PPT into WORLD vocoder

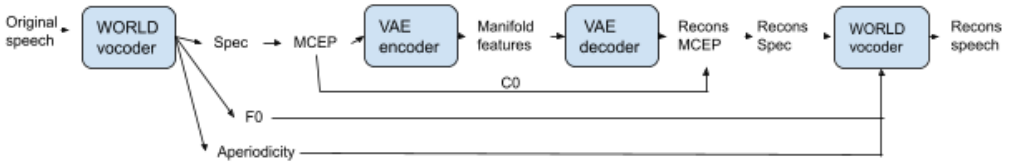


Figure 3.6: Integrate VAE into WORLD vocoder

3.2 Manifold Features

I propose a new type of feature that is both compact and interpolable, and thus ideally suited for regression approaches that involve averaging. Both compactness and interpolability are realized through projection of high-dimensional acoustic features onto a lower-dimensional manifold that is learned from a large multi-speaker database of speech data. Thus, the features are specialized to only model acoustic events related to speech, as opposed to music or other sources. Moreover, interpolability ensures that even when two or more parameter vectors are averaged, the result remains near the manifold of possible speech; this property does not hold for MCEPs, linear predictive coefficients, or the log-magnitude discrete-time Fourier spectrogram. Manifold learning is implemented with the use of a variational autoencoder.

3.2.1 Variational Autoencoder

The variational autoencoder (VAE) is a latent variable generative model, which combines variational inference and deep learning [115]. The latent variable generative model $p_\theta(\mathbf{x}|\mathbf{z})$, also called a decoder, is a deep neural network (DNN) with parameters θ. The inference model $q_\phi(\mathbf{z}|\mathbf{x})$, also called the encoder, is represented by another DNN with parameters ϕ. The latent variable $\mathbf{z}$ is a compact representation of the observation $\mathbf{x}$, generated by an encoder mapping the input space into its corresponding latent space. In this section, the VAE encoder predicts the mean $\mu_\mathbf{z}$ and log-variance $\log \sigma^2_\mathbf{z}$ of the posterior distribution $q_\phi(\mathbf{z}|\mathbf{x})$ from a 39-dimensional input vector (as shown in Figure 3.7). The decoder predicts the observation $\hat{\mathbf{x}}$ from samples of $\mathbf{z}$. I learn the parameters

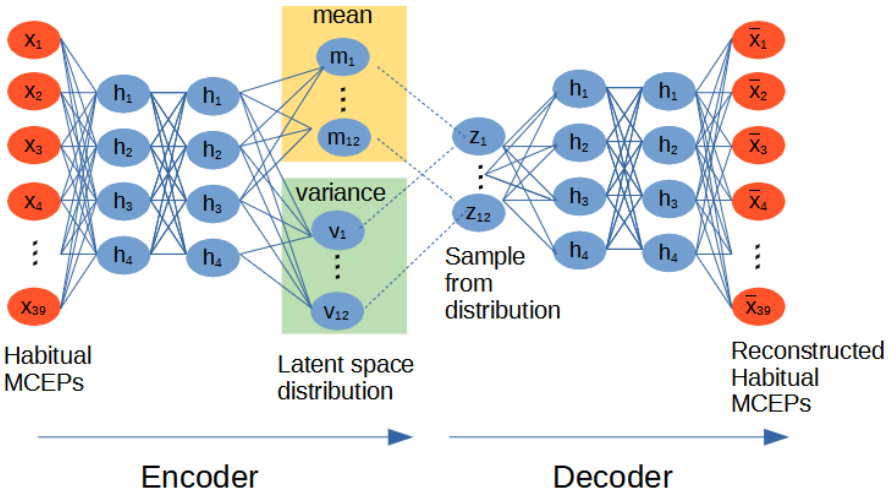


Figure 3.7: Structure of variational autoencoder. This figure is based on a figure in [170]

θ and ϕ by maximizing the variational lower bound $\mathcal{L}(\theta, \phi; \mathbf{x})$ given by

$$\mathcal{L}(\theta, \phi; \mathbf{x}) = \mathop{\mathbb{E}}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log(p_\theta(\mathbf{z}|\mathbf{x}))] - D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x})|p_\theta(\mathbf{z}|\mathbf{x})) \tag{3.13}$$

where D_{KL} denotes the Kullback-Leibler divergence.

Often, the inference model $q_\phi(\mathbf{z}|\mathbf{x})$ is parameterized using a diagonal Gaussian distribution $\mathcal{N}\left(\mathbf{z}; \boldsymbol{\mu}, diag(\boldsymbol{\sigma}_\mathbf{z}^2)\right)$. The prior is modeled as an isotopic parameterless Gaussian distribution $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I})$. To yield a differentiable network after sampling, I use the common technique of re-parameterizing the random variable $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})$ as a deterministic variable $\mathbf{z} = \boldsymbol{\mu}_\mathbf{z} + \boldsymbol{\sigma}_\mathbf{z} \odot \boldsymbol{\epsilon}$, where $\odot$ denotes an element-wise product, and vector $\boldsymbol{\epsilon}$ is sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Recent studies showed the efficacy and interpolability of VAE-based latent representation in modeling and transforming speech [88, 19, 89].

My VAE encoder consists of three fully-connected layers with 256 nodes each and a 12-dimensional Gaussian parametric layer modeling $\mathbf{z}$. No activation function is applied to the Gaussian parametric layer. For other layers, I use rectified linear units. The decoder is identical except for the Gaussian layer. I use $\mu_\mathbf{z}$ as my compact representation of $\mathbf{x}$. To train the VAE, I use the TIMIT [41] dataset. Of the 630 available speakers I select all 462 speakers designated for training and all 144 designated for validation. As is convention, I eliminate the spoken dialect samples (SA sentences) for all speakers. I train with the Adam optimizer [114], a mini-batch size of 256, and early stopping.

3.2.2 Integrate Manifold Features into a Vocoder

For the initial analysis and final synthesis, I use the WORLD vocoder [152, 150]. Our process is inserted as additional steps after analysis and before synthesis. Specifically, I first calculate the 40-dimensional MCEP (MCEP-40) from the inverse Fourier transform of the 512-point mel-warped log spectrogram. (Reconstruction from MCEP-40 resulted in no perceivable degradation as compared to the vocoder using the high-resolution spectrogram.) In a second step, I subtract the mean, calculated on the whole dataset, with the goal of reducing the channel effect of a particular dataset (cepstral mean subtraction); I also exclude the zeroth coefficient C0, representing energy. The resulting 39-dimensional vector is then encoded as a 12-dimensional latent representation by the VAE, and then immediately decoded (as shown in Figure 3.7). I re-add the zeroth-coefficient C0 unmodified. Finally, I calculate a new high-resolution spectrogram from the resulting MCEP, and synthesize a new speech waveform, using the original fundamental frequency and aperiodicity information (Figure 3.6).

3.3 Experiment: Reconstruction Quality

In this section, I compare Probabilistic Peak Tracking feature and manifold feature to each other and to the baselines of line spectral frequency and mel-cepstrum coefficients (for more details about the baselines, see Section 2.3.2) in the task of speech reconstruction. The manifold features are realized by VAE; thus, I denote it as VAE. In addition to my proposed systems with 20-dimensional PPT features (PPT-20), 12-dimensional VAE features (VAE-12), I also implemented two other systems for comparison. The MCEP-12 system used 12^{th}-order MCEP to represent the spectrogram; here I chose the order/dimensionality to be the same as the VAE-12 system. The LSF-20 system used 20^{th}-order linear predictive coefficients converted to line spectral frequencies (LSF) to represent the spectrogram, calculated using the autocorrelation method derived from the inverse Fourier transform of the squared magnitude spectrogram. The LSF order was chosen to produce an expected log-spectral distortion (LSD) approximately similar to that of VAE-12.

To evaluate reconstruction quality, I use the voice conversion challenge (VCC) 2016 database [209], which features 5 male and 5 female speakers. Each speaker has 162 parallel utterances. We arbitrarily selected two female (SF1, TF1) and two male speakers (SM2, TM1). Using these four speakers and all available sentences, the mean (and standard deviation in parentheses) LSD in dB produced by the three vocoding systems were as follows: VAE-12 8.0 (3.0), MCEP-12: 9.18 (3.0), LSF-20: 8.7 (3.4), PPT-20: 9.37 (2.73). Objectively, it appears that the three systems are roughly comparable. However, it is known that the LSD measure is a poor predictor of human perception.

A \ B	LSF-20	MCEP-12	VAE-12	PPT-20
NAT	+0.77*	+1.34*	+1.02*	+1.28*
LSF-20		+1.08*	-0.04	+0.26*
MCEP-12			-0.44*	-0.31*
VAE-12				+0.45*

Table 3.1: Relative quality between original and vocoded stimuli. Positive values show A is better than B. Results marked with an asterisk are significantly different ($p < 0.001$) as compared to 0 (representing no preference) in a 1-sample t-test.

To evaluate reconstruction quality perceptually, I select the comparative mean opinion score (CMOS) approach to compare the speech quality of the four vocoding systems and natural speech (NAT). At each trial, participants listen to samples A and B in sequence and are then asked: "Is A more natural than B?" Participants select a response from a 5-point scale that consisted of "definitely better" (+2), "better" (+1), "same" (0), "worse" (−1), and "definitely worse" (−2). For the test materials, I used 4 speakers, 32 sentences, and 4 systems and thus 6 condition pairs, resulting in $4 \times 32 \times 6 = 768$ unique trials. The perceived loudness differences between these stimuli were minimized using a root-mean-square A-weighted (RMSA) measure [2].[4] For the experiment, I want each listener to hear each unique sentence only once (presentation order was randomized); therefore I needed $768 \div 32 = 24$ listeners to cover all trials. This and subsequent experiments were conducted on Amazon Mechanical Turk (AMT); I required listeners to have an approval rate $\geq 90\%$ and to live in the U. S. Table 3.1 shows the pair-wise relative quality of the systems (after an appropriate transformation handling the random presentation order of A and B). All synthetic systems were statistically significantly different from NAT; the difference between MCEP-12 and LSF-20, as well as the difference between MCEP-12 and VAE-12 were also significant; the latter shows that VAE-12 was able to code the speech spectrogram more efficiently. The difference between PPT-20 and LSF-20 were statistically significantly, as well as between PPT-20 and VAE-12; it shows VAE-12 outperform PPT-20 in terms of both size and perceived speech quality. The difference between LSF-20 and VAE-12 was not significant.

From Table 3.1, it is difficult to give an ordering of the systems in terms of quality. For example, it is easy to determine that NAT is the best system, while, MCEP-12 is the worst. It is hard to determine the second best system between LSF-20 and VAE-12. Table 3.1 showed an insignificant difference (0.04) between LSF-20 and VAE-12. To overcome this, I projected the non-negative pair-wise relative quality matrix to a single dimension, using multiple dimensional scaling (MDS),

[4]Samples for the experiment and the next two experiments are available at https://tuanad121.github.io/samples/2019-09-15-Manifold/

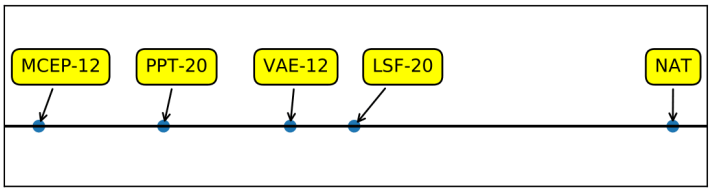


Figure 3.8: Reconstruction quality

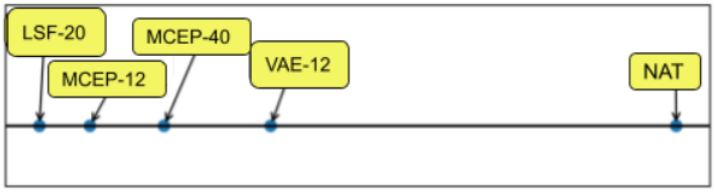


Figure 3.9: Voice conversion quality

a dimensionality reduction technique that attempts to preserve the pair-wise distances of data points. Figure 3.8 shows the result. When looking at the pair-wise relative quality matrix, the difference between LSF-20 and VAE-12 is not statistically significant. According to MDS, LSF-20 is better than VAE-12 in terms of reconstruction quality. The PPT-20 is the second worst system. Therefore, I chose not to use PPT-20 in the upcoming experiments.

3.4 Experiment: Voice Conversion

Voice conversion is the task of converting voice from a source speaker to sound like it is spoken by a target speaker. In this section, I compare manifold features to each other and to the baselines of line spectral frequency and mel-cepstrum coefficients in the task of voice conversion. I use comparative mean opinion score (CMOS) to evaluate speaker accuracy (Section 3.4.1), and speech quality (Section 3.4.2).

I perform a voice conversion experiment using four systems based on different spectral features: MCEP-40 (popular features for voice conversion), VAE-12 (my proposed manifold features), MCEP-12 (dimension-matched comparison to VAE), and LSF-20 (another classic voice conversion features). I use the same four speakers from the VCC corpus that I used in Section 3.3, and I arbitrarily arrange them into four source/target speaker pairs (two intra-gender and two inter-gender): SM2→TM1 (M2M), SM2→TF1 (M2F), SF1→TF1 (F2F), SF1→TM1 (F2M). I divided the available 163 sentences into 100 training, 30 validation, and 32 test sentences.

During training, I first align all sentences of a given source and target speakers using dynamic time warping (DTW) on 32$^{\text{nd}}$-order log filter bank features. Next, I analyze sentences with all

the systems. Finally, I train a spectral mapping from source to target features for each system. The mapping is implemented by a deep neural network (DNN) with four hidden layers of 512 nodes each. For each layer I use batch normalization, parametric ReLu [79], and dropout (at a rate of 20%). For the input vector, I add context by concatenating the current frame with the five preceding and the 5 following frames. I normalize the input and outputs of the network via standard scaling. Similar to training the VAE, I use the Adam optimizer, a mini-batch size of 256, and early stopping, for this and subsequent mapping experiments.

During conversion of the test sentences, I first analyze the source with the vocoder, then compute the desired spectral feature, and map it. In order to measure spectral mapping performance in isolation, I create output sentences from the mapped and aligned (to the target) spectral features, and the unmodified target energy, F0, and aperiodicity information. For the LSF-20 system, when necessary, we sort the mapped features per frame to satisfy the required monotonicity of LSFs. Finally, I minimized loudness differences using the RMSA measure.

3.4.1 Speaker Accuracy

Similar to Section 3.3, I use CMOS to evaluate the similarity between converted source speaker and the target speaker. In this test, listeners hear two different sentences A and B, and are then asked "is B spoken by the same speaker as A?" Listeners respond using a 5-point scale comprised of "definitely same" (+2), "same" (+1), "unsure" (0), "different" (−1), and "definitely different" (−2). One sentence is a converted sample (from source to target) and the other is an unmodified (NAT) sample. I want to determine whether converted samples sound similar to the target speaker ("same" condition), but sound different from the source speaker ("different" condition). Specifically, in half of the tests ("same" condition), listeners compare modified samples to natural speech of the target speaker. In other half ("different" condition), listeners compare modified samples to natural speech of the source speaker. In "different condition" of inter-gender conversions, however, listeners compare modified samples to natural speech of a speaker who has the same gender as the target speaker. For example, in SM2→TF1 conversion (M2F), I compare modified samples to natural speech of SF1 for the "different" condition and to natural speech of TF1 for the "same" condition.

Our test involved 32 sentences × 4 systems × 4 conversions × 2 conditions ("same" or a "different") = 1,024 unique trials. The experiment was conducted on AMT (with the same requirements as in Section 3.3) with $1024 \div 32 = 32$ participants to cover all trials. Table 3.2 shows that VAE-12 had the best average conversion performance (0.7). I used a two-tailed t-test to compare VAE-12 to other systems and found that it was significantly different ($p < 0.05$) to LSF-20.

pair \ system	LSF-20	MCEP-12	MCEP-40	VAE-12
F2M	0.6 (1.2)	0.18 (1.7)	0.0 (1.4)	0.47 (1.14)
M2F	0.4 (1.2)	0.8 (1.2)	1.0 (1.17)	0.9 (1.4)
F2F	0.5 (1.17)	1.0 (1.16)	0.6 (1.19)	0.7 (1.4)
M2M	-0.3 (1.3)	-0.5 (1.1)	0.5 (1.2)	0.8 (1.2)
average	0.3 (1.27)	0.4 (1.4)	0.5 (1.3)	*0.7* (1.4)

Table 3.2: Speaker accuracy for the same condition

A \ B	LSF-20	MCEP-12	MCEP-40	VAE-12
NAT	1.8*	1.8*	1.67*	1.7*
LSF-20		0.0	-0.54*	-1.08*
MCEP-12			-0.27*	-0.5*
MCEP-40				-0.5*

Table 3.3: Relative quality between vocoded target and mapping, results marked with an asterisk are significantly different ($p < 0.001$) in a one-sample t-test.

3.4.2 Speech Quality

I use CMOS to compare the speech quality between the four mapping systems and the NAT condition ("is A better than B?"), using a 5-point scale comprised of "definitely better" (+2), "better" (+1), "unsure" (0), "worse" (−1), and "definitely worse" (−2). The test involved 32 sentences × 4 conversions × 10 system pairs, resulting in 1,280 unique trials. I conducted the listening test on AMT with $1{,}280 \div 32 = 40$ listeners, with each listener hearing 32 unique sentence materials. The results in Table 3.3 show that VAE-12 outperformed MCEP-40 and LSF-20. The multidimensional scaling shown in Figure 3.9 shows that VAE-12 was the closest to NAT.

3.5 Experiment: Style Conversion

Different from voice conversion, style conversion is the task of converting the speech of a speaker from a source style to sound like a target style in order to improve speech intelligibility. In style conversion, one can convert spectral characteristics and duration characteristics of habitual speech to clear speech (as shown in Figure 3.10). In this section and Chapter 4, I focus on converting spectral characteristics. In Chapter 6, I present an attempt on converting duration characteristics.

First, I establish which speakers benefit from using the clear spectrum in place of habitual spectrum via a hybridization approach, as speakers use different strategies to produce clear speech. I then use a keyword recall accuracy test to measure the intelligibility of hybridized stimuli, compared with a purely-vocoded habitual condition, and a purely-vocoded clear condition (Section 3.5.2).

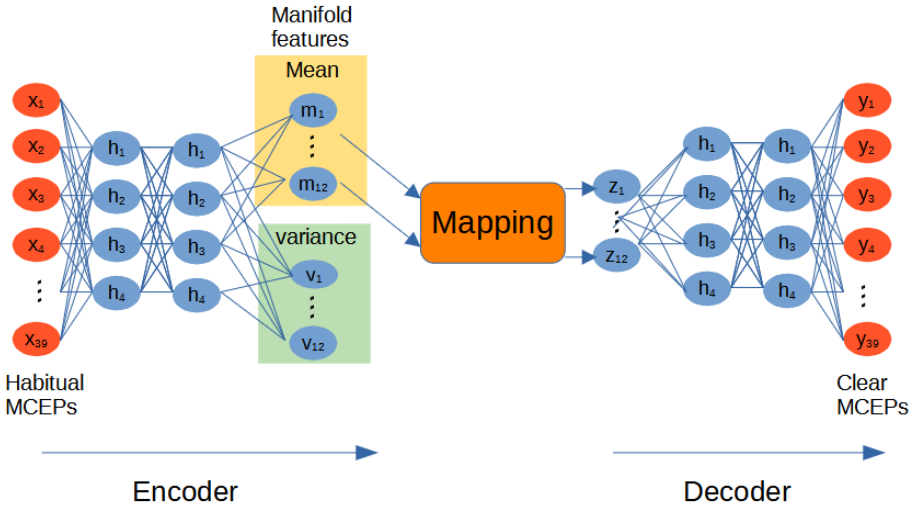


Figure 3.10: Spectral mapping for converting speaking style

Second, I evaluate the efficacy of my manifold feature for the purpose of mapping habitual style to clear style and thus improving speech intelligibility for selected speakers (Section 3.5.4).

3.5.1 Data

I use a database collected by Tjaden [204, 205]. The corpus consists of 78 speakers consisting of typical speakers (CS, N=32), speakers with multiple sclerosis (MS, N=30), and speakers with Parkinson's disease (PD, N=16). All read the same 25 Harvard sentences (Appendix A) in habitual and clear conditions (loud, slow, and fast conditions are also available). I used the speakers that Tjaden found to have the highest intelligibility difference between their clear and habitual speech. I imposed a minimum threshold of 20% absolute difference; this resulted in 11 speakers (6 CS, 2 MS, and 3 PD).

3.5.2 Hybridization

I first establish which speakers benefit from using the clear spectrum in place of the habitual spectrum (via a hybridization approach), as speakers use different strategies to produce clear speech. To this end, I measure the intelligibility of hybridized stimuli [95, 203], compared with a purely vocoded habitual condition, and a purely vocoded clear condition. The hybridized stimuli were created by combining the clear spectrum, aligned to the habitual style, with habitual F0, and habitual aperiodicity information, using the WORLD vocoder. I minimized the loudness differences of stimuli by using an RMSA measure. Finally, each utterance was mixed with babble noise at 0 dB SNR to avoid saturation effects.

	MSF7	MSF15	PDF3	PDF7	PDM6	CSM4	CSM8	CSM7	CSM6	CSF8	CSF12
vocoded HAB	68	45	5	13	30	26	42	35	30	26	18
hybrid	59	37	7	22*	55*	30	43	53*	39*	25	24
vocoded CLR	80*	53	11*	25*	56*	74*	49	65*	60*	44*	53*

Table 3.4: Average keyword accuracy. Results marked with an asterisk were significantly different ($p < 0.05$) as compared to the vocoded HAB condition in a two-tailed t-test.

The speech intelligibility test design consisted of 25 sentences $\times$ 11 speakers $\times$ 3 conditions = 825 unique trials. I performed the test on AMT, wherein 66 participants listened to 25 Harvard utterances (see appendix A), which contain five keywords each. Listeners typed out each sentence as best as they could; their responses were subsequently manually scored. I then calculated the average number of keywords correctly identified. Table 3.4 shows the average keyword accuracy. I observed that spectral hybridization led to statistically significant improvements in speakers PDF7, PDM6, CSM7, and CSM6, but also resulted in degradations for MSF7 and MSF15.

3.5.3 Mapping

I evaluate the efficacy of my proposed VAE-12 system for the purpose of mapping habitual style to clear style and thus improving speech intelligibility for speakers that have shown to benefit from the clear spectrum. I use the top three speakers PDF7, PDM6, and CSM7 that showed the most benefit in the hybridization experiment. I align each habitual utterance to its parallel clear utterance of the same speaker using DTW on 32^{nd}-order log filter-bank features. Then, I train speaker-dependent mappings from habitual VAE-12 to clear VAE-12, where I use two different DNN structures. The first structure is a typical feedforward network used in the previous voice conversion experiment (also called DNN-mapping VAE). For the second structure (also called Skip-mapping VAE), I introduce skip connections [80], as shown in Figure 3.11; I hypothesize that the latter structure is more appropriate when input and output are very similar as is the case here. The use of skip-connection is motivated by the fact that the spectral difference in style conversion is not as big as it is in voice conversion.

I create conversion stimuli consisting of the mapped VAE-12, and F0 and aperiodicity information from the original habitual speech. To create the 25 conversion sentences, I use a leave-one-out approach. Otherwise, network configurations and training parameters are identical to the voice conversion experiment.

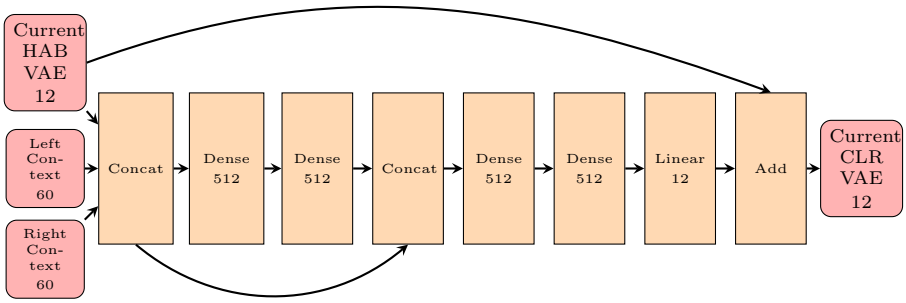


Figure 3.11: DNN architecture with skip connection

	CSM7	PDF7	PDM6
vocoded HAB	38	13	24
DNN-mapping VAE	32	13	35
Skip-mapping VAE	38	11	*46**
hybrid	56*	27*	50*
vocoded CLR	69*	23*	41*

Table 3.5: Average keyword accuracy. Results marked with an asterisk are significantly different ($p < 0.05$) as compared to the vocoded HAB condition in a two-tailed t-test.

3.5.4 Speech Intelligibility

To evaluate speech intelligibility, I designed a test consisting of 25 sentences $\times$ 3 speakers $\times$ 5 conditions (2 purely vocoded, 1 hybrid, 2 mappings) = 375 unique trails. The test was conducted similarly to the previous one in 3.5.2, except 30 listeners participated. The hybrid stimuli show an upper bound (or "oracle" mapping) on the intelligibility for the VAE-mapping. Table 3.5 shows average keyword accuracy. For PDM6, hybrid speech was significantly better than habitual speech (50 versus 24). Although DNN-mapping VAE was better than habitual speech (35 versus 24), their difference was not statistically significant. I observed that the VAE-mapping using a custom DNN with skip connection (skip-mapping VAE) led to a statistically significant improvement for speaker PDM6, but no significant differences in other cases, using a two-tailed t-test.

3.6 Conclusion

I proposed a compact and interpolable feature for spectral regression, implemented by a speaker-independent VAE. In a speech reconstruction experiment, I showed that using VAE-12 achieved significantly better perceived speech quality compared to a MCEP-12 feature. The PPT-20 did not show advantages over VAE-12 in speech reconstruction quality. In a voice conversion experiment, I

showed that mapping VAE-12 resulted in significantly better perceived speech quality compared to a MCEP-40 feature, with similar speaker accuracy, thus demonstrating the efficiency of mapping in a low-dimensional latent feature space. I also showed that VAE-12 outperformed LSF-20 in terms of similar speaker accuracy. In a *habitual* to clear style conversion experiment, I showed that VAE-12 together with a custom skip-connection deep neural network significantly improved the speech intelligibility of one of three speakers, with the average keyword recall accuracy increasing from 24% to 46%.

Chapter 4

Spectral Mapping for Style Conversion of Typical and Dysarthric Speech

In the previous chapter, I utilized a feedforward network with custom skip-connection for the style conversion task of mapping the manifold features of habitual speech to those of clear speech. In this chapter, I further improve the performance of the style conversion mapping.[1] Motivated by the success of conditional Generative Adversarial Nets (cGANs) in machine learning, I utilize cGANs for the style conversion task of mapping the manifold features of habitual speech to those of clear speech. I give an overview of cGANs, I describe the application of cGANs for mapping speaking styles, and I present my configuration for the following experiments (Section 4.1). Specifically, conditional cGANs are investigated with three mappings: one-to-one mappings (Section 4.2), many-to-one mappings (Section 4.3), and many-to-many mappings (Section 4.4). For one-to-one mappings, I compare the performance of cGANs-based one-to-one mappings to my manifold model (for more details about my manifold model, see Chapter 3). For each of the three mappings, I extensively evaluate the performance of the cGANs on both typical speakers and speakers with mild dysarthria secondary to Parkinson's disease for intelligibility improvement in a noisy environment.

4.1 Conditional Generative Adversarial Nets

4.1.1 Overview of cGANs

Traditional GANs have a generative model or a generator (G) and a discriminative model or a discriminator (D), that together play a min-max game [70] (as shown in Figure 4.1). Component G tries to fool component D by generating outputs close to the real data, while component D is

[1] This chapter is based on a paper published in Interspeech [45], *Improving Speech Intelligibility through Speaker Dependent and Independent Spectral Style Conversion*, Tuan Dinh, Alexander Kain, Kris Tjaden. This chapter gives more details.

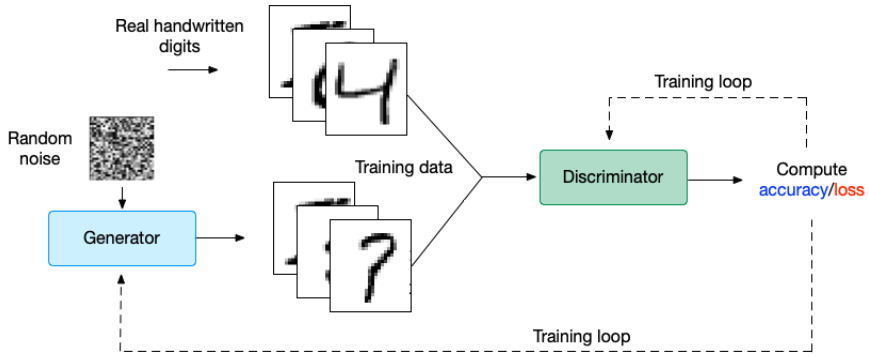


Figure 4.1: Structure of traditional GAN [5]

trained to distinguish the output of component G from real data. Component G is a mapping function from random noise z to y, $G : \{z\} \rightarrow y$. The GAN model is trained to minimize the adversarial loss $\mathcal{L}_{\text{GAN}}(D, G)$:

$$\min_G \max_D \mathcal{L}_{\text{GAN}}(D, G) = \mathbb{E}_y \left[\log D(y)\right] + \mathbb{E}_z \left[\log(1 - D(G(z)))\right] \tag{4.1}$$

In contrast, a cGAN model learns a mapping from an input x and random noise z to y, $G : \{x, z\} \rightarrow y$. The cGAN model has both G and D conditioned on input x, trained to minimize the adversarial loss $\mathcal{L}_{\text{cGAN}}(D, G)$ [93]:

$$\min_G \max_D \mathcal{L}_{\text{cGAN}}(D, G) = \mathbb{E}_{x,y} \left[\log D(x, y)\right] + \mathbb{E}_{x,z} \left[\log(1 - D(x, G(x, z)))\right] \tag{4.2}$$

4.1.2 cGANs for Style Conversion

Previously, Michelsanti applied a cGAN in speech denoising to map the spectral features of noisy speech to those of intelligible speech [143]. Thus, his work has a similar goal of improving speech intelligibility as us. In his case, he is interested in typical speakers and building hearing-aid devices for listeners to reduce noise. Michelsanti used convolution neural networks as his generator. He found that random noise input z is not useful; instead, he used spectral features of noisy speech as input.

In my case, I applied a cGAN in spectral style conversion to map the spectral features of habitual speech to those of clear speech. I am interested in atypical speakers and building the speaking-aid devices (Figure 1.4). My goal is to help the habitual speech become more resilient to noise. Different from Michelsanti, I use feedforward networks (Figure 4.2) because they work better with my manifold features. Motivated by Michelsanti, I do not use random noise z. Instead, my generator maps habitual speech features to aligned clear speech features as shown in Figure 4.3.

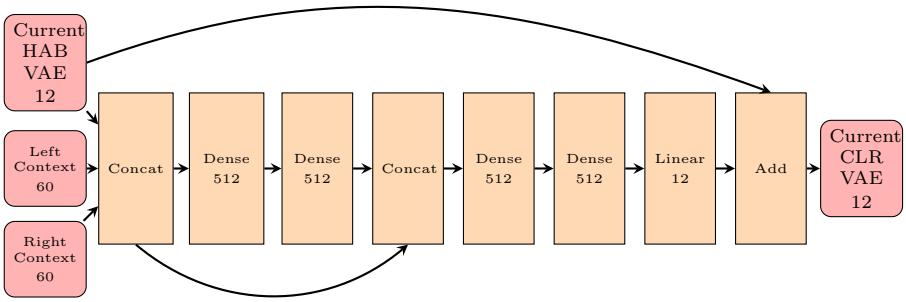


Figure 4.2: Generator architecture with skip connection

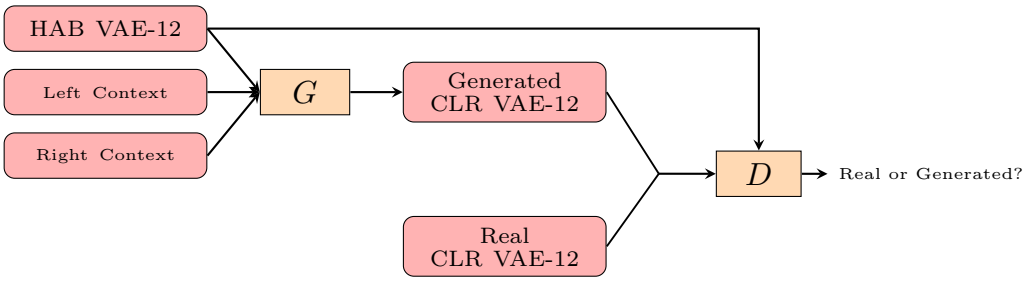


Figure 4.3: cGAN framework for converting habitual manifold features (VAE-12) to clear manifold features (VAE-12)

For the input vector of G, I add context by concatenating the current habitual frame in my VAE-12 representation with five preceding and five following frames, which helps capture the temporal dependency of speech frames. I normalize the input and outputs of the network via standard scaling. The input of D consists of either the output of G or aligned clear feature frames, combined with the current habitual feature frame (what I want the output to be conditioned on). Thus, both G and D are conditioned on the current habitual feature. In addition to the adversarial loss function $\mathcal{L}_{\text{cGAN}}(D, G)$ in Equation 4.2, I also minimize the L1 loss between the output of G and the ground truth; this addition is demonstrated to generate less blurry output compared to a root-mean-squared reconstruction loss [93]. I add the L1 loss with a scaling factor of 100 to $\mathcal{L}(D, G)$.

4.1.3 Configuration of cGANs

In this section, I present the configuration of cGANs that is used in the three mapping experiments. The structure of G is shown in Figure 4.2. For the purpose of comparison, the generator structure is the same as my previous manifold network (Section 3.2). By adding the input of G to the output of its last layer, I expect the network to focus on the difference between the habitual and clear

manifold features (VAE-12). The discriminator is a DNN with two hidden layers of 256 nodes each, and a single node output layer with sigmoidal activation function. To help stabilize the training process, I used 1) a leaky ReLU activation function with a slope of -0.2 for negative inputs for both G and D, 2) a dropout layer following each hidden layer of D with a dropout rate of 0.5, 3) the Adam optimizer with a batch size of 128, and 4) weights initialized from a zero-centered normal distribution with standard deviation 0.02 [37]. I used a momentum of 0.5, a learning rate decay of 0.00001, and learning rate of 0.0001 for D, and 0.0002 for G.

4.2 One-to-One Mapping

In this section, I apply my proposed cGAN conversion method to convert between two styles of a speaker; specifically, I aim to convert the spectral aspects of the habitual style to those of the clear style. I compare the intelligibility of mapped speech using my cGANs-based mappings to a baseline of a feedforward network with custom skip-connections (see Section 3.5). I conduct a keyword recall accuracy test (see Section 2.5) to calculate speech intelligibility.[2]

4.2.1 Data

I use a database with 78 speakers consisting of typical speakers (CS, N=32), speakers with multiple sclerosis (MS, N=30), and speakers with Parkinson's disease (PD, N=16) [204, 205]. All read the same 25 Harvard sentences (see appendix A) in habitual and clear conditions (loud, slow, and fast conditions were also available). Speaker's names consist of group, gender, and number, e. g. PDF7 is the seventh female speaker with PD.

4.2.2 Method

I apply my proposed cGAN conversion method to convert between two styles of a speaker; specifically, I aim to convert the spectral aspects of a the habitual style to those of the clear style, in an effort to improve the speech intelligibility of the former. For analysis and synthesis, I used manifold features VAE-12 (see Section 3.2). I extract fundamental frequency (F0), aperiodicity, and VAE-12 from each utterance.

I selected three speakers: CSM7, PDF7, and PDM6, who have been shown to benefit the most from the clear spectrum. I aligned each habitual utterance to its parallel clear utterance of the same speaker using dynamic time warping (DTW) on 32^{nd}-order log filter bank features. Then, I

[2]This one-to-one mapping is an attempt to improve the performance of the DNN mapping described in Section 3.5

mapping \ speaker	PDF7	PDM6	CSM7
DNN	16.8	16.67	16.44
GAN	12.85	12.58	12.67

Table 4.1: Average LSD (in dB)

pre-trained the generator that maps habitual VAE-12 to clear VAE-12, minimizing mean-squared-error loss. Pre-training stops when there is no progress in a validation set; the maximum number of epochs was 100. Finally, I trained my proposed cGAN structure up to 300 epochs.

I create conversion stimuli using the mapped VAE-12, and F0 and aperiodicity information from the source habitual speech. To create the 25 conversion sentences, I use a leave-one-out approach, using 22 sentences for training and two for validation. Hybrid stimuli are created by replacing the habitual spectra with their aligned clear spectra [95].[3]

4.2.3 Objective Evaluation

I compare the performance of my proposed cGAN to my previous DNN. Table 4.1 shows the average log spectral distortion (LSD) between mapped VAE-12 and clear VAE-12. The cGAN mapping has typically smaller average LSD than its DNN counterpart. Specifically, Figure 4.4 shows the LSD of 25 test sentences from my two mappings. For most sentences, the LSD of the GAN mapping is lower than the LSD of the DNN mapping. Moreover, Figure 4.5 shows the variance ratio $\frac{\sigma^2_{\text{CLR}}}{\sigma^2_{\text{MAP}}}$ between clear VAE-12 and mapped VAE-12 for each feature component. The smaller variance ratio of the cGAN mapping method reported that the over-smoothing effect is reduced compared to the DNN-based method. Figure 4.6 shows a comparison of various spectrum.

4.2.4 Subjective Evaluation

LSD is not a good predictor for human perception; therefore, to evaluate speech intelligibility, I designed a test consisting of 25 sentences $\times$ 3 speakers (CSM7, PDF7, PDM6) $\times$ 5 conditions (2 purely vocoded, 1 hybrid, 2 mappings) = 375 unique trials in a Latin-square design. I performed the test on the Amazon Mechanical Turk, where 60 participants listened to 25 Harvard utterances containing five keywords each (Appendix A). Listeners typed out each sentence as best as they could, and I calculated the average number of keywords correctly identified. The hybrid stimuli show an upper bound (or "oracle" mapping) on the intelligibility improvement. The vocoded HAB

[3]Samples for the one-to-one method and the next two methods are available at https://tuanad121.github.io/samples/2020-10-25-Dysarthria/

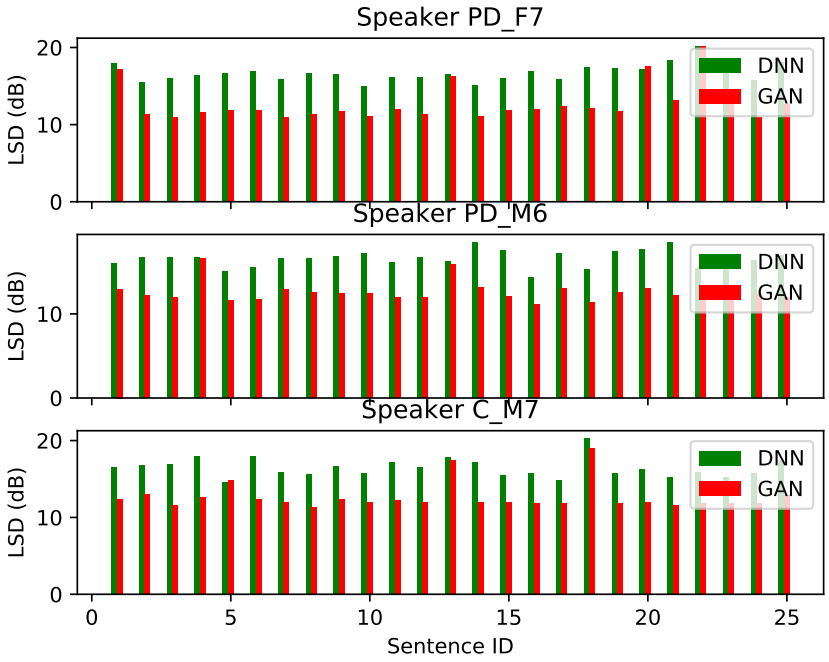


Figure 4.4: Log spectral distortion (LSD) of 25 test sentences for three speakers.

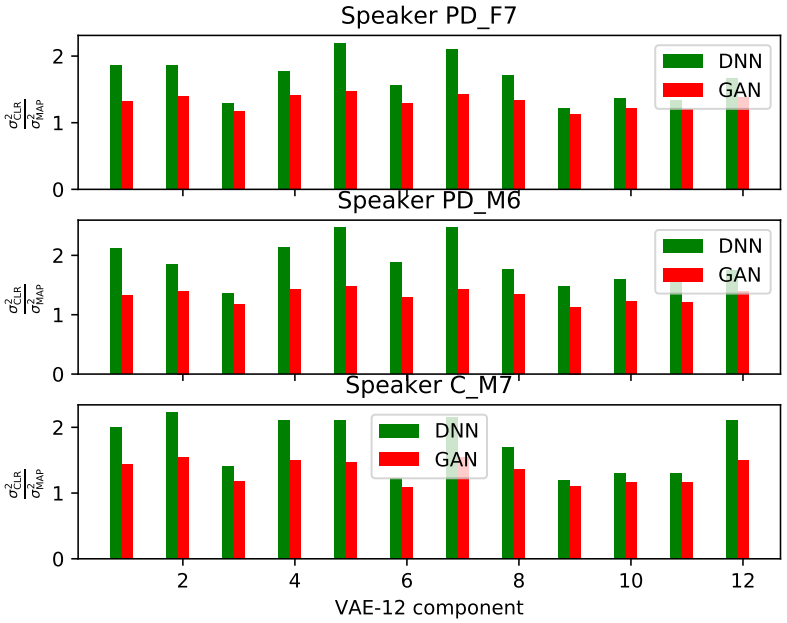


Figure 4.5: Variance ratios between clear VAE-12 (CLR) and mapped VAE-12 (MAP) features (smaller is better).

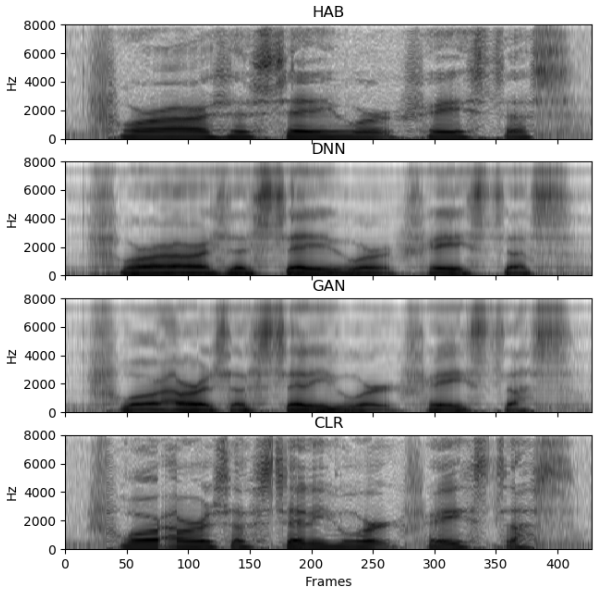


Figure 4.6: Spectrum of habitual speech (HAB), DNN mapping (DNN), cGAN mapping (GAN), and clear speech (CLR). Note the difference in formants between 2–4 kHz from the 50^{th}–100^{th} frame between the DNN and cGAN methods.

and vocoded CLR are obtained through analysis and resynthesis with unchanged parameters, using the manifold vocoder. I minimized the loudness differences between stimuli by normalizing gains in accordance with a RMSA measure. Finally, each utterance was mixed with babble noise at 0 dB SNR to avoid response saturation effects. Figure 4.7 shows average keyword accuracy. I observed that the cGAN mapping led to a statistically significant improvement ($p < 0.001$) for two speakers: PDM6 and CSM7, using a two-tailed t-test. In both cases, the cGAN mapping significantly outperformed the DNN-mapping, improving the intelligibility of two of three speakers, compared to my previous work where only one speaker improved.

4.3 Many-to-One Mappings

A disadvantage of one-to-one mappings is its requirement of specific training on the source speaker's clear speech. However, clear speech is typically unavailable for new source speakers in real life applications. Therefore, I study a many-to-one mapping approach where I map habitual speech of multiple speakers to the clear speech of a target speaker with the best sentence-level intelligibility. I trained two gender-dependent mappings that mapped all habitual VAE-12 features of all male (or female) speakers to clear VAE-12 of a male (or female) target speaker.

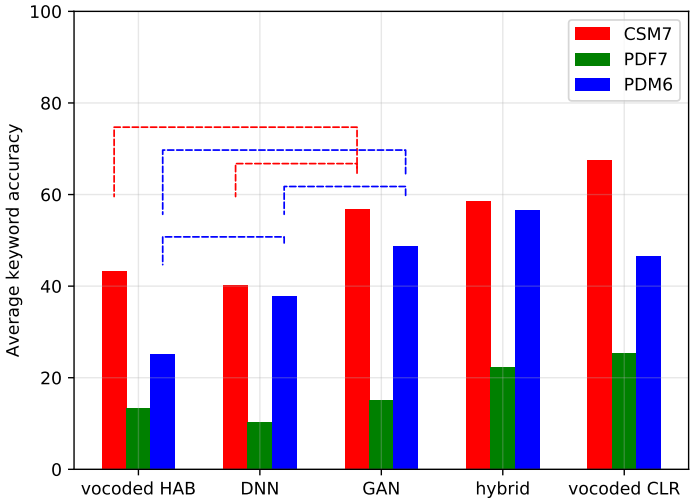


Figure 4.7: Keyword recall accuracy of three speakers. The dashed lines show statistically significant differences.

I use the same data as the previous experiment (Section 4.2.1). I select the clear speech of the two speakers CSM10 and CSF15, which has the highest sentence-level intelligibility [204, 205] for the male and female case, respectively. I train two gender-dependent mappings that mapped all habitual VAE-12 features of all male (or female) speakers (except one of three speaker CSM7, PDF7, PDM6) to clear VAE-12 of CSM10 or CSF15, respectively. The habitual speech of the three speakers CSM7, PDF7, and PDM6 is used for testing. The mapped VAE-12, in combination with the original F0 and aperiodicity of the source speaker, are used to create conversion stimuli. Hybrid stimuli are created by replacing habitual spectra of the source speaker with aligned clear spectra of the target speaker using hybridization (Section 3.5.2).

4.3.1 Objective Evaluation

The average LSD between mapped and clear spectrum, with LSD between the input habitual and clear spectrum in parentheses, were: 17.32 (21.57) dB for CSM7, 22.7 (27.62) dB for PDF7, and 18.8 (23.28) dB for PDM6, confirming that the mapped speech is closer to clear speech than the input habitual speech.

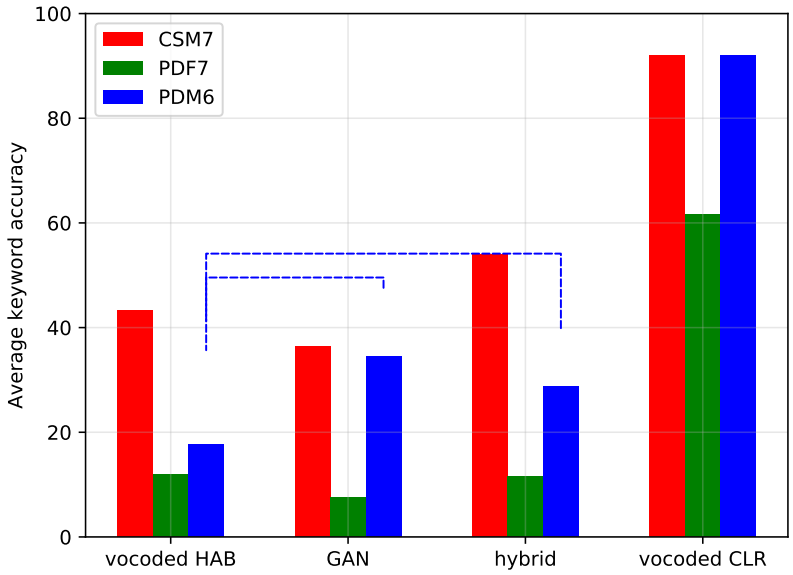


Figure 4.8: Keyword recall accuracy of three speakers. The 'vocoded CLR' condition denotes clear speech of target speakers CSM10 and CSF15 for male and female cases, respectively. The dashed lines show statistically significant differences.

4.3.2 Subjective Evaluation

To evaluate the efficacy of the method in terms of intelligibility, I designed a test consisting of 25 sentences $\times$ 3 source speakers (CSM7, PDF7, PDM6) $\times$ 3 conditions (vocoded HAB, cGAN-mapping, hybrid) + 25 sentences $\times$ 2 target speakers (CSM10, CSF15) $\times$ 1 condition (vocoded CLR) = 275 unique trials. I conducted the listening experiment similarly to the previous one, except the number of listeners was 44.

Figure 4.8 shows the resulting keyword accuracy. I found that my many-to-one style conversion significantly improved the intelligibility of one speaker of three test speakers from 17.6% to 34.4%, using a two-tailed t-test ($p < 0.01$) , while there is no improvement in other cases.

4.4 Many-to-Many Mapping

The disadvantage of the previous many-to-one mapping (Section 4.3) is that speaker characteristics cannot be preserved. In this section, I investigate a many-to-many mapping that aims to learn solely the style differences, while preserving the linguistic message and speaker characteristics. This task is the hardest among my three experiments because not all speakers' spectral changes

condition \ speaker	CSM7	PDF7	PDM6
vocoded HAB	36.8	10	28.8
GAN	39.6	15.6	26.8
hybrid	62	22.8	57.6
vocoded CLR	66.8	22.4	48

Table 4.2: Average keyword accuracy

across styles have been shown to benefit speech intelligibility. I used the same data as the previous experiments. I aligned each habitual utterance to its parallel clear utterance of the same speaker using DTW on 32^{nd}-order log filter-bank features. Then I trained all one-to-one mappings from habitual VAE-12 to clear VAE-12 simultaneously.

4.4.1 Objective Evaluation

The average LSD between the cGAN mapping and clear spectrum for the three test speakers, with the LSD between habitual and clear spectrum in parentheses, are: 16.36 (17.0) dB for CSM7, 16.66 (17.53) dB for PDF7, and 16.42 (18.06) dB for PDM6, confirming that the mapped speech is closer to clear speech than the input habitual speech.

4.4.2 Subjective Evaluation

I designed a test consisting of 25 sentences $\times$ 3 speakers (CSM7, PDF7, PDM6) $\times$ 4 conditions (vocoded HAB, GAN, hybrid, vocoded CLR) = 300 unique trials. I conducted the experiment similarly to the previous ones, except the number of listeners was 24. Table 4.2 shows average keyword recall accuracy. The cGAN resulted in improvements for two speakers: CSM7 and PDF7. However, the results were not statistically significant using a two-tailed t-test ($p > 0.05$).

4.5 Conclusion

I explored a cGAN architecture for speaker dependent and independent style conversion. In the speaker-dependent one-to-one mapping case, I showed that the cGAN outperformed a DNN in terms of average keyword recall accuracy in all cases. Moreover, the cGAN significantly improved speech intelligibility of two of three speakers, compared to one speaker when using the DNN. In the speaker-independent many-to-one mapping case, I significantly improved speech intelligibility of one of three speakers, with average keyword recall accuracy increasing from 17.6% to 34.4%. In the speaker-independent many-to-many mapping case, the cGAN can improve average keyword

accuracy over that of vocoded habitual speech for two speakers: CSM7 and PDF7, but without statistical significance.

Chapter 5

Voice Conversion and F0 Synthesis of Alaryngeal Speech

The main focus of Chapter 3 and 4 is to improve the intelligibility of people who have typical speech or mildly dysarthric speech using style conversion, where I learn how to map one speaking style to another, such as habitual to clear. In this chapter, I address the alaryngeal speech spoken by people who underwent laryngectomy, which is more difficult to understand.[1] The goal of this chapter is to convert alaryngeal speech (LAR) into intelligibility speech (INT), which relates to style conversion. Specifically, I propose an approach that has two parts. The first part predicts voicing/unvoicing and the degree of voicing using feed-forward networks. The second part is for LAR-to-INT spectral mappings using a cGANs. Figure 5.1 shows a flowchart of the approach; the first part is in the boxes labelled 'VUV model' and 'AP model' and the second part is in the box 'MCEP model'.[2]

In the rest of this chapter, I do the following. In Section 5.1, I provide the motivation for the research in this chapter. In Section 5.2, I review the related works for increasing intelligibility and naturalness of alaryngeal speech. In Section 5.3, I discuss my data, including how to create my target data of intelligible speech, which is different from the clear speech of Chapter 3 and Chapter 4. In Section 5.4, I predict voicing/unvoicing and the degree of voicing. In Section 5.5, I predict the spectrum of intelligible speech from that of alaryngeal speech. In Section 5.6, I create a synthetic fundamental frequency trajectory with an intonation model consisting of phrase and accent curves to address the unusable fundamental frequency (F0) information of alaryngeal speech. In Section 5.7, I evaluate the LAR-to-INT conversion methods on data.

[1] This chapter is based on a paper published in Interspeech [43], *Increasing the Intelligibility and Naturalness of Alaryngeal Speech Using Voice Conversion and Synthetic Fundamental Frequency*, Tuan Dinh, Alexander Kain, Robin Samlan, Beiming Cao, Jun Wang. This chapter gives more details.

[2] In this chapter, I do not use manifold features. The features are extracted using a VAE that is trained on a TIMIT database of normal speakers. I suspect that the VAE might not work well on alaryngeal speech

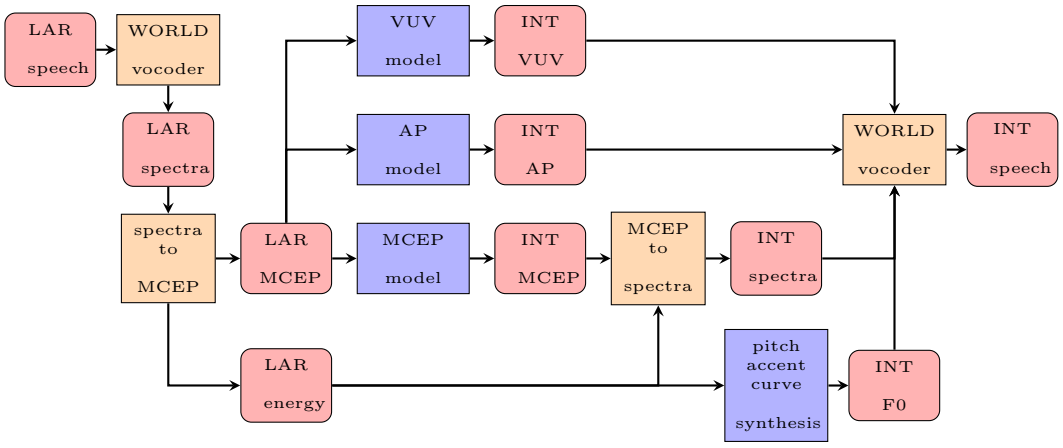


Figure 5.1: Flowchart of approach during prediction

5.1 Alaryngeal Speech

Speech is probably the most important biosignal for human communication. Pressure from the lungs drive typical laryngeal voice and speech. The pharynx, tongue, and lips shape exhaled air to produce voiceless sounds, and quasiperiodic vocal fold vibration creates the sound wave that vocal tract constrictions shape into vowels and voiced consonants. Individuals who undergo total laryngectomy lose their ability to produce speech sounds normally because their vocal folds and lungs are disconnected from the vocal tract. Laryngectomy is performed as surgical treatment for advanced laryngeal and hypopharyngeal cancers. These patients experience a lower quality of life because of their atypical speech (I term this as LAR speech) during social interactions, as they believe that other people perceive them as abnormal, or they directly experience symbolic violence [142]. In 2020, an estimated 12,370 new cases of laryngeal cancers are expected in the U. S. [1]. Although the incidence of these cancers is decreasing due to the decreasing number of smokers, there is still a large projected number in the next decades because the rate of decrease is only 2–3% [1].

There are currently a limited number of alternative communication options for people who have Laryngectomy. Typing-based alternative and augmentative communication (AAC) devices are slow and limited by the speed of typing. The main speech options for individuals after laryngectomy are (1) *esophageal* speech (pushing air from the mouth to the pharyngo-esophageal segment (PES) and using the PES for vibration), (2) tracheo-esophageal puncture (TEP) speech wherein speakers use lung air to power PES vibration for voiced speech where an one-way valve was surgically placed between the esophagus and trachea to provide airflow from lungs, and (3) use of an artificial larynx, in the form of either external electrolarynx (ELX) placed on the neck or with an intraoral

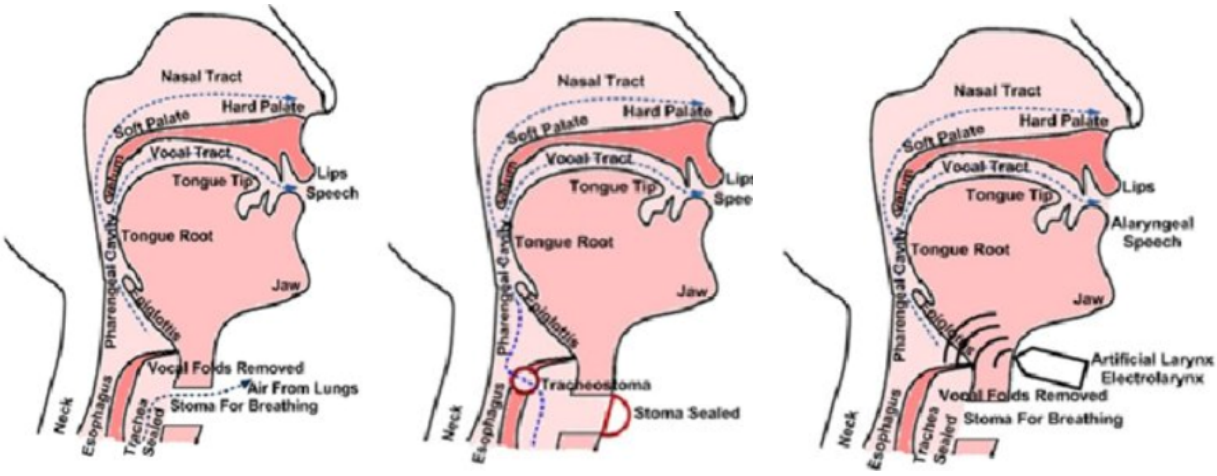


Figure 5.2: From left to right, esophageal, tracheo-esophageal, and electrolarynx speech [17]

tube, resulting in always-voiced speech at a constant pitch (see Figure 5.2). The ELX generates an electronic sound source that is shaped by the lips and tongue into always-voiced speech at a constant pitch [132]. These options are suboptimal because they are either invasive (requiring additional surgery) or are difficult to use: esophageal speech requires extensive training and practice and is difficult to learn [15], TEP is a surgical operation and requires talkers to place their thumb over their stoma during the speech act, which is not hands-free and poses certain risks, and the artificial larynx produces a very robotic-like sounding voice. All options produce unnatural sounding and difficult to understand speech for several reasons, including poor voice quality, voiced/voiceless differentiation, and articulatory precision [18, 112].

5.2 Related Work: Increasing Intelligibility of Alaryngeal Speech

For two decades, researchers have attempted to create natural-sounding speech for people who have had a laryngectomy. Using rule-based spectral voice conversion approaches, some differences between alaryngeal speech and normal speech can be compensated for by modifying the properties of speech formants. For example, in an early work, researchers made esophageal speech more intelligible by expanding the formant bandwidths [6]. Another approach is to decrease formants' frequencies using formant shifting methods [135, 179], since it was found that speakers who underwent partial laryngectomy shift their formants to higher frequencies due to the shortened vocal tract length [108]. Similarly, TEP speech is reported to have a spectral tilt which favors the high frequency band; thus, del Pozo used a 6 dB/octave roll-off filter to de-emphasize the high frequency

band [40]. These approaches led to limited improvement in intelligibility or naturalness.

For statistical voice conversion, previous approaches included the use of Gaussian mixture models [110, 111] and deep neural networks [109, 16] for mapping spectral features. These models are limited because of over-smoothing of the converted spectra, leading to muffled speech [207, 103]. Recently, the generative adversarial network (GAN) [70] has been shown to be effective in addressing the over-smoothing problem in voice conversion [103] and speech synthesis [104]. The LAR-to-INT spectral mapping can be viewed as an image-to-image translation task, in which the image is a window of the time-frequency representation of speech. In image-to-image translation, a conditional GAN (cGAN) [93] is probably effective in generating less blurry images by combining a traditional adversarial loss and a mean absolute reconstruction loss (or L1 loss). In this chapter, I leverage the cGAN architecture for mapping LAR speech to INT speech.

Generally, LAR speech lacks reliable voicing and fundamental frequency (F0) information, as well as meaningful F0 variability. One approach used to produce a more natural sounding speech signal is therefore to create converted or synthetic voicing and F0 trajectories. In related work on reconstructing normal speech from whispered speech, F0 values were estimated from filtered gain parameters [155] or using the first formant frequency and its magnitude [141].

For F0 conversion or synthesis, a variety of approaches have been proposed. The introduction of jitter (small pitch perturbations) was found to reduce the artificiality of ELX speech [35]. Another approach created an artificial pitch contour by means of filtering, scaling and offsetting the energy envelope [135]. Alternatively, the F0 trajectory can be generated using formant frequencies as well as gains from a linear prediction model [182]. In a third approach, TEP speech is first converted to whispered speech, and then an F0 trajectory was synthesized by adding a normalized short-term energy contour of the whispered speech to an average pitch when the energy value is greater than a threshold [158]. In this dissertation, I directly use the normalized short-term energy contour of LAR speech in combination with a simple intonation model to synthesize the final F0 trajectory.

5.3 Data

For the *source* LAR speech, I use a database of 4 male speakers consisting of 3 LAR-TEP speakers (L001, L002, L006) and 1 LAR-ELX speaker (L004); the average age was 61.75±8.77.[3] The speakers underwent total laryngectomy. The pitch of the LAR-TEP speech is low and highly variable, and voicing depends largely on energy. Speech analysis using standard voicing and fundamental frequency (F0) analysis algorithms fail for this type of speech the majority of the

[3]This data was provided by Dr. Jun Wang, who is one of the authors of the paper that the chapter is based on.

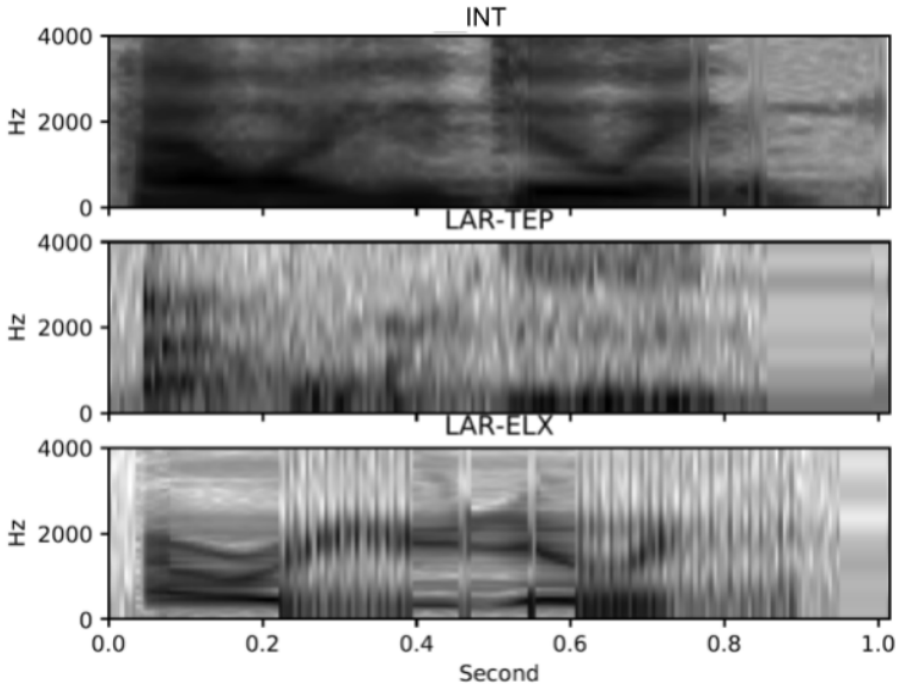


Figure 5.3: INT, LAR-TEP, and LAR-ELX spectrogram

time. The LAR-ELX speech is always voiced with a constant F0, which was about 80 Hz for L004. All speakers read all 132 sentences in the AAC132 list [30]. Figure 5.3 shows an example of INT, LAR-TEP, and LAR-ELX spectrogram.

For the *target* INT speech, the ideal option is natural voice, such as habitual speech or clear speech (Chapter 3 and Chapter 4). Although natural speech is preferred, I use a synthetic male voice due to expediency. Moreover, the advantage of using a synthetic voice is the capability of creating a lot of data and arbitrary voices. Specifically, I create a synthetic male voice using Tacotron 2 [183] with the Waveglow vocoder [176]. I create audio for all sentences in the AAC132 list. This database was then divided into 100/16/16 sentences for training/validation/testing. All waveforms are re-sampled from 22.05 kHz to 16 kHz.

I also use a multi-speaker TIMIT database [41] for pre-training. To simulate the characteristics of LAR-TEP speech, I create a fully unvoiced version of TIMIT (FU-TIMIT); and for LAR-ELX speech, I create a fully voiced version of TIMIT (FV-TIMIT). I use a process of first analyzing TIMIT using the WORLD vocoder [153, 151], then setting all frames to either unvoiced, or voiced with all F0 values set at a constant value of 80 Hz to create FU-TIMIT and FV-TIMIT, respectively. Of the 630 available speakers I use all 462/144/24 speakers designated for training, validation, and

testing, respectively. By convention, I eliminate the spoken dialect samples (SA sentences) for all speakers.

5.4 Predicting Voicing and Degree of Voicing

In this section, I propose a method for predicting when speech should be voiced and the degree of voicing from a spectrogram. Specifically, I predict a binary voicing value (VUV) and continuous 2-band aperiodicity (AP) [151] values from mel-cepstral coefficients (MCEP), using deep neural networks (DNN). This prediction will be used in re-synthesizing speech so that speech is voiced/unvoiced in appropriate places.

Due to limited amount of LAR training data, I use pre-training to leverage the general relationship between spectral characteristics and voicing. Given that FU-TIMIT should be similar to the unvoiced nature of LAR-TEP speech, and FV-TIMIT should be similar to the always voiced LAR-ELX speech, I compare three types of pre-trained models based on TIMIT, FU-TIMIT, and FV-TIMIT.[4] The second factor I investigate is the amount of context.

5.4.1 Pre-training

I *pre-train* three kinds of speaker-independent DNNs using either TIMIT, FU-TIMIT, or FV-TIMIT as input. For each utterance in the input database, I use VUV and AP from corresponding utterances in TIMIT as the target. Obviously, using TIMIT to predict voicing in TIMIT is easier than using FU-TIMIT or FV-TIMIT to predict voicing in TIMIT. But the latter two should do better when tested on LAR speech without adaptation.

I use the WORLD vocoder to obtain the spectrogram for each utterance in the input database. I further extract MCEP-32 from the spectrogram. I exclude the zeroth coefficient (representing energy), and I use MCEP coefficients 1–31 as input for training. I normalize inputs of the network with standard scaling. I add context by concatenating the current frame with the preceding and following frames. I consider context lengths of 5, 11 and 25 frames, which are 25, 55, and 105 ms, respectively.

For the target, I use the WORLD vocoder to extract voicing and 2-band aperiodicity parameters for each utterance in TIMIT. The voicing is a binary voiced/unvoiced flag (VUV). The 2-band aperiodicity (AP) is a single scalar representing the degree of voicing at 3000 Hz, which is the boundary frequency of the two frequency bands: $[0, 3000]$ and $[3000, 8000]$ Hz. I use the VUV and AP as targets for training.

[4] I use the terms voiced and unvoiced from a signal analysis point of view, not a production point of view.

The DNN has three hidden layers with 256 nodes each. The activation function is parametric ReLU. Each hidden layer is preceded by batch normalization (except the first layer), and followed by dropout with a dropout rate of 0.2 (except the last layer). I train using the Adam optimizer, a mini-batch size of 256, and early stopping. The binary cross entropy and mean-squared error loss functions are used for voicing classification and 2-band aperiodicity regression, respectively. In total, there are 9 pairs of pre-trained models: 3 pre-training datasets $\times$ 3 context sizes and 2 output types (VUV and AP).

I objectively evaluate the performance of each pair of models using balanced accuracy (BAC, defined as average recall) for VUV classification (since the classes were imbalanced), and r^2 for AP regression. For both of the measures, the closer to one the better. On average, I obtained a BAC/r^2 of 0.94/0.75, as shown in column TIMIT in Table 5.1. For each of the three training sets, the different context lengths did not result in noticeable differences in BAC and r^2. This might be because I use the same model architecture for the different context lengths. For the rest of the discussion, I just focus on the 55ms results.

The results showed that it is possible to predict voicing and the degree of voicing from spectral shape alone. As expected, the model that uses TIMIT as input (rows 5) works the best because the training data contains the voicing and degree of voicing that we want to predict. The model obtained a BAC/r^2 of 0.99/0.87. The other models based on FU-TIMIT and FV-TIMIT are not far behind. FU-TIMIT models obtained a BAC/r^2 of 0.89/0.72, and FV-TIMIT models obtained a BAC/r^2 of 0.93/0.84.

I then tested the pre-trained networks, without any adaptation, to predict target VUV or AP from LAR-TEP and LAR-ELX speech with 16 test sentences. I found that the BAC and r^2 drastically decrease across all three pairs of models. I have an approximate BAC/r^2 of 0.60/$-$0.44 for L001, 0.58/$-$0.68 for L002, 0.49/$-$0.4 for L004, 0.48/$-$0.6 for L006 as shown in columns labelled Pretrain in Table 5.1. Our expectation is that using FU-TIMIT would work best for LAR-TEP (L001, L002, and L006), and using FV-TIMIT would work best for LAR-ELX (L004). For L001, the three databases do not result in a noticeable difference in BAC. For L002, using FU-TIMIT works best in BAC as expected. For L004, normal TIMIT works best in BAC. For L006, FV-TIMIT unexpectedly works slightly better than FU-TIMIT. Although the results do not match our expectation entirely, we still need to adapt our models with LAR speech.

5.4.2 Adaptation

The previous section shows that pre-trained models using no LAR speech do not do that well, as expected. Hence, I adapt the pre-trained models with LAR-TEP and LAR-ELX speech. I use

		source →	TIMIT	FU-TIMIT	FV-TIMIT	L001 (TEP)		L002 (TEP)		L004 (ELX)		L006 (TEP)	
pre-training set ↓	context size ↓	measure ↓				Pretrain	Adapt	Pretrain	Adapt	Pretrain	Adapt	Pretrain	Adapt
TIMIT	25ms	r^2	0.88			-0.55	0.26	-0.72	0.41	-0.41	0.29	-0.87	0.04
		BAC	0.99			0.62	0.69	0.45	0.73	0.63	0.70	0.52	0.64
	55ms	r^2	0.87			-0.51	0.22	-0.70	0.43	–0.44	0.29	-0.84	0.04
		BAC	0.99			0.64	0.70	0.56	0.73	0.63	0.72	0.53	0.65
	105ms	r^2	0.86			-0.57	0.24	-0.64	0.45	-0.38	0.29	-0.86	0.05
		BAC	0.99			0.62	0.70	0.55	0.73	0.62	0.70	0.52	0.66
FU-TIMIT	25ms	r^2		0.70		-0.27	0.26	0.06	0.42	-0.30	0.24	-0.37	0.05
		BAC		0.89		0.58	0.69	0.64	0.74	0.49	0.72	0.43	0.64
	55ms	r^2		0.72		-0.17	0.21	0.017	0.43	-1.00	0.27	-0.5	0.05
		BAC		0.89		0.60	0.67	0.67	0.75	0.49	0.71	0.48	0.67
	105ms	r^2		0.72		-0.2	0.24	0.00	0.43	-1.18	0.29	-0.57	0.05
		BAC		0.87		0.62	0.74	0.69	0.74	0.49	0.71	0.49	0.66
FV-TIMIT	25ms	r^2			0.83	-0.62	0.25	-0.75	0.41	-0.30	0.30	-0.84	0.04
		BAC			0.94	0.60	0.71	0.54	0.73	0.40	0.70	0.54	0.64
	55ms	r^2			0.84	-0.58	0.23	-0.7	0.43	-0.28	0.29	-0.84	0.05
		BAC			0.93	0.58	0.72	0.55	0.73	0.48	0.70	0.55	0.64
	105ms	r^2			0.83	-0.53	0.23	-0.62	0.45	-0.25	0.29	-0.81	0.06
		BAC			0.92	0.61	0.72	0.54	0.74	0.47	0.69	0.52	0.67

Table 5.1: $r^2$and balanced accuracy (BAC), gray color indicates mismatch between source speaker and pre-training set

speaker specific adaptation due to the limited number of speakers.

I align each LAR utterance to its corresponding target intelligible synthetic utterance using dynamic time warping on 32^{nd}-order log filter bank features. There are 36 pairs of adapted models: 9 pairs of pre-trained models (3 pre-training datasets and 3 context sizes) $\times$ 4 speakers. All training settings for the DNNs are the same as those of pre-training. For adaptation, I first tried to only adapt the first one or two hidden layers. Those layers are expected to extract meaningful low-level features coming from the input data. But the results are not as good as adapting the whole models. Because there are many more voiced frames than unvoiced frames, I over-sample the unvoiced frames to balance the classes.

Columns labeled "Adapt" in Table 5.1 show the results. The average BAC of VUV and r^2 of AP are 0.69/0.22 for L001, 0.73/0.43 for L002, 0.71/0.28 for L004, and 0.65/0.05 for L006; it appears that some speakers' AP is much easier to predict than others', whereas VUV prediction performance is similar. As expected, adaptation always improved performance of pre-trained models. Varying context size resulted in a relatively narrow BAC range from 0.65 to 0.73, and thus I use 55 ms from this point forward. Surprisingly, pre-training with FU and FV-TIMIT as opposed to TIMIT did not show improved performance. Figure 5.4 shows predicted VUV and AP of a LAR test sentence with and without adaptation.

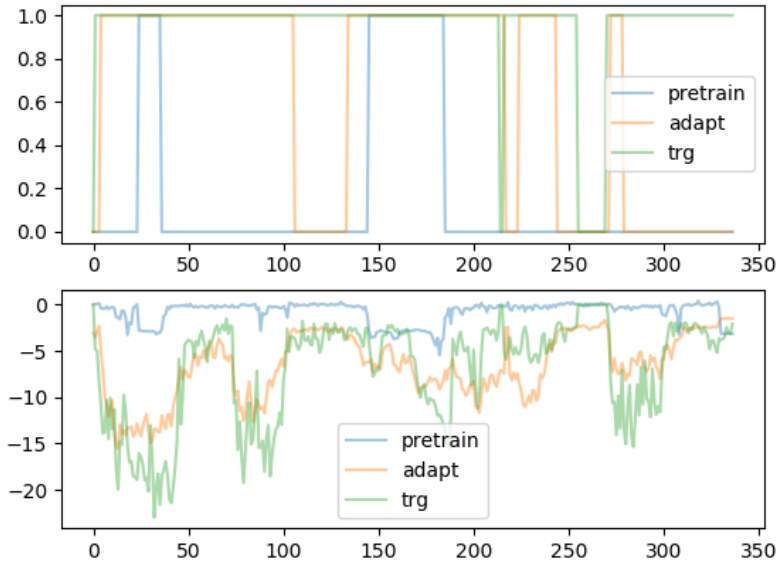
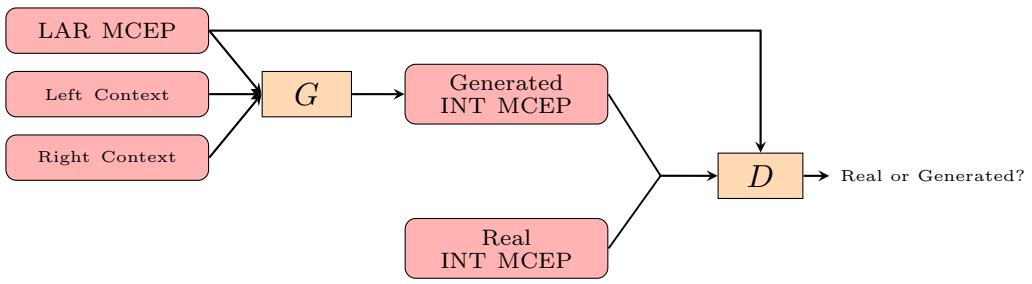


Figure 5.4: Example predictions (VUV in top panel, AP in bottom panel)



you

Figure 5.5: cGAN framework for style conversion for predicting INT spectrum from LAR spectrum

5.5 Predicting Spectrum

In this section, I propose a method for predicting normal spectral features from distorted spectral features using conditional Generative Adversarial Network. This is a parallel step to predicting voicing (see Section 5.4).

5.5.1 Conditional Generative Adversarial Network

Traditional GANs have a generative model or a generator (G) and a discriminative model or a discriminator (D), that together play a min-max game. Component G tries to fool component D by generating outputs similar to the real data, while component D is trained to distinguish the

output of component G from real data. Component G is a mapping function from random noise z to y, $G : \{z\} \rightarrow y$ [70]. In contrast, a cGAN model learns a mapping from an input x and random noise z to y, $G : \{x, z\} \rightarrow y$. The cGAN model has both G and D conditioned on input x [93], trained with the objective function $\mathcal{L}(D, G)$:

$$\min_G \max_D \mathcal{L}(D, G) = \mathbb{E}_{x,y} [\log D(x, y)] + \mathbb{E}_{x,z} [\log(1 - D(x, G(x, z)))] \tag{5.1}$$

Similar to Section 4.2, I do not use random noise z; and my generator maps LAR speech features to aligned intelligible speech features as shown in Figure 5.5. For the input vector of G, I add context by concatenating the current LAR MCEP-32 frame with five preceding and five following frames, which is the same as the 55ms context in Section 5.4. I normalize the inputs and outputs of the network via standard scaling. The input of D consists of a single frame of either the output of G or an aligned target feature frame, combined with the current LAR feature frame (what I wanted the output to be conditioned on). Thus, both G and D are conditioned on the current LAR feature frame. In addition to the adversarial loss function $\mathcal{L}(D, G)$ in Equation 5.1, I also minimize the L1 loss between the output of G and the ground truth; this addition was demonstrated to generate less blurry output compared to a root-mean-squared reconstruction loss in an image task [93]. I add the L1 loss with a weighting factor of 100 to $\mathcal{L}(D, G)$.

The structure of the generator G, shown in Figure 5.6, is similar to my previous work (see Chapter 4); however, there is no skip connection that adds the input of G to the output of its final dense layer, because performance worsened when using the skip connection. The discriminator D is a DNN with two hidden layers with 256 nodes each, and a single-node output layer with sigmoidal activation function. To help stabilize the training process, I use (1) a leaky ReLU activation function with a slope of 0.2 for negative inputs for both G and D, (2) a dropout layer following each hidden layer of D with a dropout rate of 0.5, (3) the Adam optimizer with a batch size of 128, and (4) weights initialized from a zero-centered normal distribution with standard deviation 0.02 [37]. I use a momentum of 0.5, a learning rate decay of 10^{-5}, and learning rate of 10^{-4} for D, and $2 \cdot 10^{-4}$ for G.

5.5.2 Predicting Spectrum

I first pre-train the cGAN with the methods described in Section 5.5.1 to convert FU-TIMIT and FV-TIMIT MCEPs to TIMIT MCEPs, excluding the zeroth (energy) coefficient (similar to Section 5.4.1).[5] This is a data-rich proxy for the eventual mapping of LAR to INT speech. There

[5] Unlike in Section 5.4, I do not pre-train a TIMIT MCEP to TIMIT MCEP model.

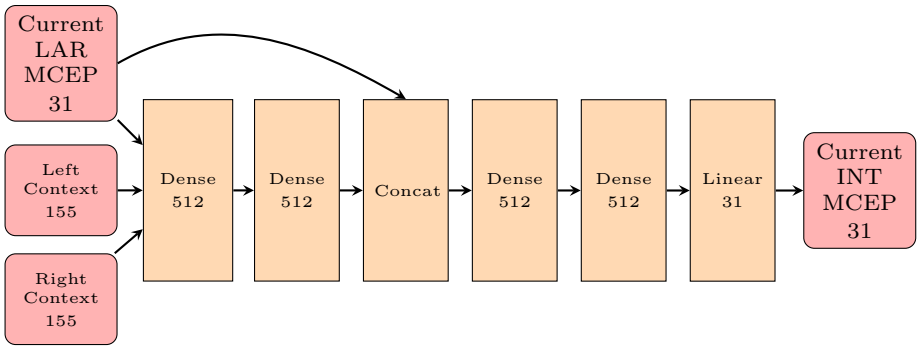


Figure 5.6: Generator architecture

are two pre-trained models, one for each pre-training database. I pass the source energy unmodified to the target features. I then calculate the predicted spectra and compared them to target spectra in terms of log spectral distortion (LSD). Table 5.2 shows the LSD between estimated spectra and target spectra. On average, the LSDs were 7.64 and 6.46 dB for FU and FV-TIMIT, respectively (TIMIT row). The values in parentheses: 11.33 and 11 dB show the LSDs before modification.

I then tested the pre-trained models on LAR speech, obtaining 60 dB for L001, 45 dB for L002, 51 dB for L004, and 62 dB for L006 (rows labelled Pretrain). I expected that FU-TIMIT would work best for LAR-TEP (L001, L002, and L006), and FV-TIMIT would work best for LAR-ELX (L004). As expected, the LSDs of FU-TIMIT are smaller than FV-TIMIT for L001, L002, and L006. The LSD of FU-TIMIT is unexpectedly smaller than FV-TIMIT for L004. Furthermore, by comparing the LSD before modification (the numbers in parentheses) and after modification, the pre-trained models do not help reduce the LSDs. For example, the LSD is 60.6 dB before modification, and it is 60 dB after modification with a pre-trained model for L001. The lack of improvement is disappointing but it is not entirely unexpected because the FU-TIMIT and FV-TIMIT models do not know about LAR speech.

I then adapt these pre-trained models to convert LAR MCEP to INT MCEP (similar to Section 5.4.2). There are eight adapted models (2 pre-trained models $\times$ 4 speakers). I adapt them in two ways: adapting only the generator or adapting both the generator and the discriminator. I observed that the latter yields lower LSD scores. The average LSD was 32 dB for L001, 33 dB for L002, 31.6 dB for L004, and 37.4 dB for L006 (rows labelled Adapt). As expected, the adaptation always improved performance. Pre-training with FU-TIMIT versus FV-TIMIT does not have noticeable effect on adaptation. This shows that our results can be improved by better modelling the differences between TEP and ELX in pre-training.

mapping (↓)	pre-train set (→)	FU-TIMIT	FV-TIMIT
FU-TIMIT → TIMIT		7.64 (11.33)	
FV-TIMIT → TIMIT			6.46 (11.00)
L001 (TEP) → INT	Pretrain	60 (60.6)	61.9 (60.6)
	Adapt	32 (60.6)	32 (60.6)
L002 (TEP) → INT	Pretrain	45 (46)	46.5 (46)
	Adapt	33 (46)	33 (46)
L004 (ELX) → INT	Pretrain	51.1 (51.5)	52.8 (51.5)
	Adapt	31.5 (51.5)	32 (51.5)
L006 (TEP) → INT	Pretrain	61.6 (61.2)	63 (61.2)
	Adapt	37.8 (61.2)	37 (61.2)

Table 5.2: LSD of predicted spectrum in dB (LSD of source spectrum in parentheses) with (or without) adaptation. 'FU-TIMIT → TIMIT' indicates predicting TIMIT voicing from FU-TIMIT spectrum. 'L001 (TEP) → INT' indicates predicting INT voicing from L001 spectrum. Gray color indicates a mismatch between pre-train set and source speaker (e.g., FV-TIMIT and L001 (TEP)).

5.6 Synthesizing Pitch

In this section, I present a method to synthesize an INT F0 from a LAR energy. A LAR F0 is unusable; therefore, I replace the LAR F0 with the INT F0. This is a parallel step to predicting voicing (see Section 5.4) and predicting spectrum (see Section 5.5).

There has been a lot of work on modelling pitch. Van Santen utilized the principle of *superpositional prosody transplant* to generate natural prosody contours and superimpose these contours on recorded speech for text-to-speech synthesis [220]. Langarani proposed a data-driven foot-based method for generating intonation contours for text-to-speech synthesis using a *phrase curve* for the entire utterance and *accent curve* for each foot in the utterance [126]. These methods, however, require text as well as phoneme labels as input.

In contrast, I want to generate pitch directly from LAR speech without any text and the focus of this dissertation is spectral conversion. Therefore, I use a phrase curve and a single accent curve to model intonation for each utterance. The phrase curve p is defined as

$$p(t) = p_{\min} + (p_{\max} - p_{\min}) \left(1 - \frac{t}{T}\right)^b \tag{5.2}$$

where I empirically set $p_{\max}$=140, $p_{\min}$=60, and b=0.5; t is a time index between 0 and T. The accent curve relates to LAR energy. To set accent curve α, I use

$$a(t) = A \cdot e(t) \tag{5.3}$$

where I empirically set A=40, and e is the max-normalized energy. Specifically, I analyze LAR speech into spectrogram using WORLD. I then derive MCEP-32 from the spectrogram. The zeroth

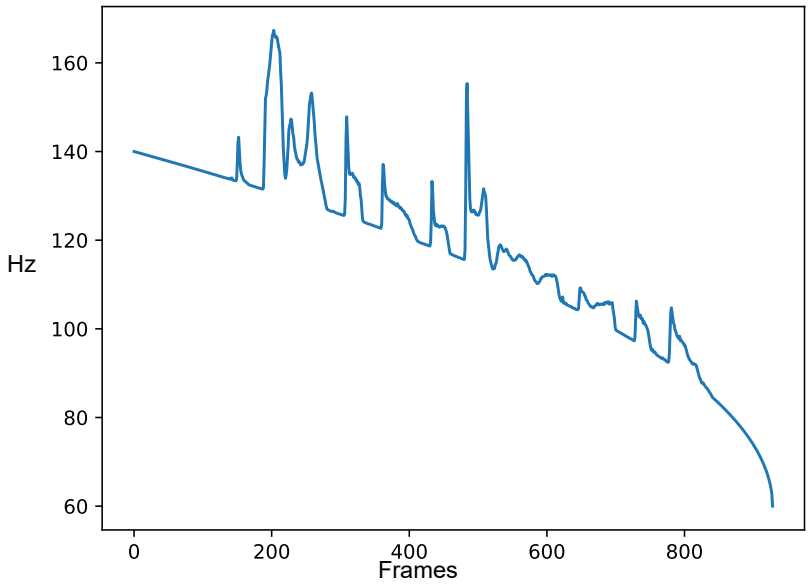


Figure 5.7: Example synthetic F0 trajectory

coefficient of MCEP-32 is energy $e(t)$, which is used to calculate accent curve in Equation 5.3. The final F0 trajectory is calculated as

$$f_0(t) = p(t) + a(t) \tag{5.4}$$

Figure 5.7 shows an example of the synthetic F0 trajectory. Informal perceptual experiments confirmed that replacing a natural F0 trajectory with a synthetic one did not reduce the naturalness of normal speech.

5.7 Experiment

In this section, I evaluate the efficacy of using predicted VUV, AP (Section 5.4), spectrum (Section 5.5), and F0 (Section 5.6) to transform LAR speech. I use two comparative mean opinion score (CMOS) tests: one for naturalness and one for intelligibility. To better understanding the efficacy of our approaches, I have three types of INT speech. The INT-spectrum denotes LAR speech with predicted MCEPs. The INT-intonation denotes LAR speech with predicted VUV, AP and F0. The INT-all denotes predicting all vocoder parameters except energy. I compare LAR to the three conditions: INT-spectrum, INT-intonation, and INT-all. Note: I compare INT speech to vocoded LAR speech: I use the WORLD vocoder to extract the F0 and spectrogram from LAR speech, then I re-synthesize LAR speech from F0 and spectrogram using the WORLD vocoder.

Speakers \ Systems	INT-spectrum	INT-intonation	INT-all
L001 (TEP)	−0.0	−0.3*	*0.4**
L002 (TEP)	−0.1	−0.0	0.1
L004 (ELX)	−0.56*	−0.25	0.22
L006 (TEP)	−0.3*	−0.2*	*0.7**

Table 5.3: Perceptual naturalness CMOS comparing modified conditions against the vocoded LAR speech condition. INT-spectrum, INT-intonation, INT-all denote predicting INT spectrum, INT VUV/AP/F0, or a combination of these, respectively. Scores marked with an asterisk are significantly different.

Speakers \ Systems	INT-spectrum	INT-intonation	INT-all
L001 (TEP)	−0.1	−0.1	0.1
L002 (TEP)	0.1	0.2	−0.3*
L004 (ELX)	−0.34*	*0.34**	−0.2
L006 (TEP)	0.2	−0.1	−0.0

Table 5.4: Perceptual intelligibility CMOS comparing modified conditions against the vocoded LAR speech condition. INT-spectrum, INT-intonation, INT-all denote predicting INT spectrum, INT VUV/AP/F0, or a combination of these, respectively. Scores marked with an asterisk are significantly different.

I am aware that vocoded LAR speech is probably inferior compared to the original LAR speech, because LAR speech is difficult to analyze; however for a fair comparison of my method I used it as a baseline. Informal listening tests have shown that vocoded LAR speech is indistinguishable from the original LAR speech in terms of intelligibility and naturalness.

Each of the two CMOS consisted of 16 sentences × 4 speakers × 3 pairs of conditions = 192 unique trials.[6] I limit each listener to hear each unique sentence once (presentation order was randomized); therefore I need blocks of $192 \div 16 = 12$ listeners to cover all trials. Both experiments were conducted on Amazon Mechanical Turk (AMT); I required listeners to have an approval rate $\geq 90\%$ and to live in the U. S. Each test had 48 listeners, for a total of *96* listeners. In each trial, participants listen to samples A and B in sequence and were then asked: "Is A more natural than B?" or "Is A more intelligible than B?" for the naturalness and intelligibility tests, respectively. Responses are selected from a 5-point scale that consisted of "definitely better" (+2), "better" (+1), "same" (0), "worse" (−1), and "definitely worse" (−2).

Table 5.3 shows the naturalness preference scores between modified speech and LAR speech.

[6]Samples for the two experiments are available at: https://tuanad121.github.io/samples/2020-10-25-Alaryngeal/

Positive scores with asterisk show an statistically significant improvement over LAR speech. INT-all (VUV, AP, F0 and MCEPs) significantly improved the naturalness of LAR speech for L001 and L006 ($p < 0.01$ in a one-sample t-test).

Table 5.4 shows the intelligibility preference scores between modified speech and LAR speech. The INT-intonation (VUV, AP, and using a synthetic F0) significantly improved intelligibility ($p < 0.01$ in a one-sample t-test) only for L004 (LAR-ELX speech). This is probably because there was no sufficient voicing information in TEP speech. Further studies with a larger number of patients should be conducted to verify these preliminary findings.

According to the results of naturalness CMOS, modifying an individual factor (spectrum or intonation) is not as effective as modifying all factors. For example, INT-all significantly improve the naturalness of LAR speech; INT-spectrum and INT-intonation, however, significantly reduce the naturalness for speaker L006. This is probably because modifying one factor might cause mismatch between the factor and other factors of LAR speech, which decreases the naturalness. In contrast, I significantly improve the intelligibility of L004 by modifying the intonation alone. This is probably because replacing a robotic F0 with a more natural synthetic F0 as well as modifying voicing and degree of voicing improve intelligibility of L004.

5.8 Conclusion

I proposed an approach that has two parts to improve naturalness and intelligibility of LAR speech: 1) predicting INT voicing/unvoicing and degree of voicing from LAR spectrum using a DNN, and 2) predicting INT MCEPs from LAR spectrum using cGANs. I also created a synthetic F0 trajectory with an intonation model consisting of phrase and accent curves. For predicting INT voicing/unvoicing and INT degree of voicing, using different context lengths did not have a noticeable impact. Additionally, pre-training the prediction networks on FU-TIMIT, and FV-TIMIT as opposed to TIMIT did not have improved performance. Similarly, pre-training with FU-TIMIT, and FV-TIMIT did not have improved performance for predicting MCEP of INT speech.

Adaptation always improved performance. In my subjective tests with four LAR speakers, I significantly improved the naturalness of two speakers, and I significantly improved the intelligibility of one speaker. The results are promising for a challenging task with a lot of individual variability among four LAR speakers.

Chapter 6

Towards Duration Style Conversion

In this chapter, I report preliminary results of improving speech intelligibility using duration conversion.[1] In previous chapters, I conducted spectral style conversion to improve speech intelligibility (see Chapter 3 and Chapter 4). In addition to spectral features, phoneme duration is thought to be important for improving intelligibility during style conversion (see Section 2.1.2). Therefore, my goal is to improve intelligibility using duration style conversion. To do so, I first analyze the phoneme duration changes between five speaking styles (habitual, slow, fast, clear, loud) to confirm that all phonemes in a sentence are not changed uniformly. For this preliminary work, I focus on converting from fast speech and habitual speech to slow speech.

In the rest of this chapter, I present the motivation for non-uniform duration conversion (Section 6.1). I present my work on predicting target duration information (Section 6.2); and I review time-scale modification procedures (Section 6.3). I present the speech data that I use for training and testing (Section 6.4). I analyze how different speaking styles affect sentence and phoneme durations (Section 6.5). Then I present a preliminary attempt to predict phoneme-level scaling factors (Section 6.6). Finally, I evaluate the performance of uniform and non-uniform duration conversion in terms of intelligibility using oracle scaling factors (Section 6.7).

6.1 Motivation for Non-uniform Duration Conversion

Changing sentence duration is necessary when desiring to change the overall speaking rate of an utterance. It is also required as part of the needed prosodic changes during voice conversion (Section 2.3).

The simplest approach to duration conversion is to apply a scaling factor to all phoneme segments equally, which is known as *uniform duration conversion*. However, it is well-known that

[1] The results in this chapter have not been published in any paper.

a change in speaking rate affects different phonemes differently [131, 64, 65, 125, 105]. Thus, duration conversion based on sentence-level (or *uniform*) scaling is likely to result in unnatural-sounding, and possibly less intelligible speech. Moreover, it has been observed that phoneme duration differ somewhat arbitrarily between source and target speakers [10]. Consequently, a simple sentence-level scaling is also likely to not mimic the target speaker's individual duration characteristics very well.

To address the issues of uniform scaling, researchers have started using phoneme-level approaches, which is known as *non-uniform duration conversion*. Covell proposed a non-uniform speech compression, which utilized local energy to determine the amount of compression [38]. In another approach, Kupryjanow classified speech signal into vowel, consonant, and silence regions with a phoneme classifier; then he applied different scaling factors on the three regions [123]. The study showed that non-uniform duration conversion outperformed uniform conversion in terms of naturalness. However, it is unclear that non-uniform conversion is better than uniform conversion in improving intelligibility in noise.

6.2 Predicting Target Durations

An automatic duration conversion consists of two steps: 1) predicting target duration information (e.g sentence duration, scaling factor), and 2) time-scale modification. In this section, I present my work on the first step, and Section 6.3 addresses the second step. I present three different methods for predicting target durations: sentence level, phoneme level, and frame level.

For sentence-level duration conversion, I predict a scaling factor for each sentence. The scaling factors are used by time-scale modification to alter the duration of a source sentence. In the sentence-level duration conversion, all phonemes in the source sentence are uniformly modified. This duration conversion is also known as uniform duration conversion.

For phoneme-level (or segmental) duration conversion, I predict a phoneme-level scaling factor for each phoneme in a sentence. The scaling factors in combination with source phoneme durations are used by time-scale modification to alter the duration of a source sentence. In the phoneme-level duration conversion, phonemes are non-uniformly modified. This duration conversion is known as non-uniform duration conversion. However, source phoneme labels are not available in real-life applications. During testing, I should use a phoneme predictor to obtain the phoneme identities (Section 6.6). Finally, the predicted scaling factor show how to shift source frames to obtain the target sentence duration. As a result, the additional phoneme prediction adds extra processing to the approach as well as longer delay.

The finest-grained duration conversion is frame-level (or sub-segmental) duration conversion. In this approach, I predict frame-level scaling factors for each sentence. The frame-level scaling factor are used by time-scale modification to alter the duration of a source sentence. The advantage of the frame-level approach over phoneme-level approach is that phonetic segmentation is not necessary; thus, there is less processing and the delay is shorter. However, frame-level duration conversion is outside of the scope of this dissertation.

For each of the three methods, one needs to predict the scaling factors from acoustic features (e.g., spectrum) using supervised training methods. One can use a corpus that has many speakers who read sentences in multiple speaking styles (see Section 6.4). One can compute scaling factors between sentences in two speaking styles (such as fast as the source and slow as the target) at the sentence, phoneme or frame level. One can use supervised training to map the acoustics of the source to the desired scaling factors. During testing, the scaling factors can be predicted from the acoustic features of the source sentence, and then applied to the source sentence giving the modified sentence.

In this preliminary work, rather than predicting the scaling factors, I use an oracle that should give us the optimal scaling factors. This will allow me to determine whether non-uniform duration conversion is better than uniform conversion without dealing with errors in prediction. During testing (Section 6.7), for an individual sentence I know how the speaker changes the duration between fast speech and slow speech. I calculate the oracle scaling factors for the sentence between its fast speech and slow speech. Then, I apply the oracle scaling factor on the fast speech to obtain a modified speech, using the time-scale modification that I describe next.

6.3 Time-scale Modification

A time-scale modification is a procedure for stretching or compressing the duration of an input audio signal. The basic steps of a time-scale modification procedure are presented in Figure 6.1. The first step is to decompose the original signal into short overlapping *analysis frames* with an *analysis hopsize* of H_a. In the second step, the frames are relocated on the time axis according to a *synthesis hopsize* of H_s. The frames are then adapted to address the artifacts introduced by the frame relocation. Finally, the adapted frames, also known as *synthesis frames*, are superimposed to obtain an output signal. In other words, the input signal is altered in length by a scaling factor of $\alpha = \frac{H_s}{H_a}$ to obtain the output signal. The main differences between the procedures are how the analysis frames are selected and how they are adapted. In this section, I review several procedures for time-scale modification.

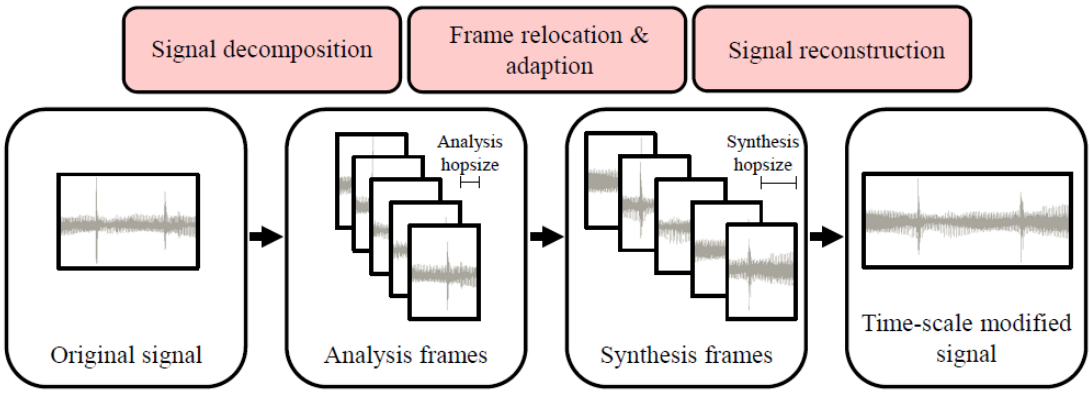


Figure 6.1: Time-scale modification procedure [49]

6.3.1 Overlap-Add

In the *overlap-add* procedure, the synthesis frames are obtained by windowing the analysis frames without any further processing. Although the procedure is efficient, superimposing the unmodified synthesis frames usually produces phase discontinuities in the output signal [48]. As a result, periodic or harmonic structures in the input signal are not preserved. This causes strong harmonic artifacts in the output signal, which is specially harmful for vowels. In contrast, with a very short length of analysis frames, this procedure is successful in preserving percussive sounds, which is helpful for stop consonants.

6.3.2 Waveform Similarity Overlap-Add

The performance of overlap-add procedure is limited by phase discontinuities. To address the issue, Verhelst selected the frames such that successive synthesis frames better fit together when superimposing them [221]. Specifically, he introduced a frame position tolerance $\Delta_{\max}$ in his Waveform Similarity Overlap-Add algorithm. After relocating the frames according to a synthesis hopsize of H_s, he maximized the similarity of two overlapping frames in the overlapping regions by shifting the frames on the time-axis by an amount of $\Delta \in [-\Delta_{\max}; \Delta_{\max}]$. Note that, the *waveform similarity overlap-add* procedure becomes the overlap-add procedure when $\Delta_{\max} = 0$. The introduced tolerance is successful in reducing the phase discontinuity artifacts. However, the procedure introduces perceivable *stuttering artifacts*, because the shifted frame positions tend to cluster around transients (e.g., p and b), which duplicate the transients several times.

6.3.3 Phase Vocoder

While Verhelst addressed the phase discontinuities in time domain [221], *phase vocoder* targeted the problem in frequency domain [62, 175]. The researchers considered each analysis frame as a weighted sum of sinusoids with given frequency and phase. Then, they computed the synthesis frames by adapting the phases of the sinusoids to avoid phase discontinuities due to adding up the synthesis frames.

First, Fourier transform is applied to analysis frames to obtain speech spectrum. Each frequency bin of the spectrum represents a sinusoid contributing to the original signal. Second, the *instantaneous frequencies* of the frequency bins are obtained from the differences of successive spectra [46]. Third, the phases of spectra are adapted using the instantaneous frequencies and the synthesis hopsize H_s [175]. Finally, inverse Fourier transform is applied to the speech spectrum to obtain synthesis frames.

The advantage of the procedure is to ensure the phase continuity of all sinusoidal contributing to output signal, which is known as *horizontal phase coherence*. However, the phase continuity of sinusoidal within one frame, known as *vertical phase coherence*, tends to be destroyed during the phase adaptation. As a result, transients, which are highly dependent on maintaining the vertical phase coherence, are smeared in output signal [48]. Moreover, the loss of vertical phase coherence causes a distinct sound coloration of phase vocoder modification, which is known as *phasiness* [127].

6.3.4 Phase Vocoder with Identity Phase Locking

Laroche reduced the loss of vertical phase coherence using *identity phase locking* [128]. In the procedure, they grouped the frequency bins; then, they updated the phases of the frequency bins in the same group, simultaneously. Specifically, frequency bins surrounding a peak of magnitude spectrum are grouped. In phase adaptation, the frequency bins that contain spectral peaks are updated in the usual phase vocoder fashion. The phases of the remaining frequency bins are *locked* to the phases of the nearest spectral peak, which ensures vertical phase coherence.

6.3.5 Combination of Time-Scale Modification procedures

Previous sections showed that different time-scale modification procedures are well-suited for particular types of input signal. Overlap-add works particularly well for percussive signals (see Section 6.3.1 and Section 6.3.2); while, phase vocoder works particularly well for signals with harmonic

content (see Section 6.3.3 and Section 6.3.4). Driedger proposed a combined time-scale modification approach using harmonic-percussive source separation techniques [50]. After decomposing the input signal, he applied the phase vocoder with identity phase locking to harmonic component and OLA to percussive component.

For this chapter, I decided to use the *phase vocoder with identity phase locking* procedure for my experiments. The procedure achieved the best quality of modified speech in an informal listening test that I conducted.

6.4 Data

Our speech corpus consists of 32 speakers × 24 identical Harvard sentences (Appendix A) × 5 speaking styles (*conditions*): habitual (H, synonymous with conversational), clear (C), loud (L), slow (S), and fast (F), for a total of 3,840 utterances recorded at 22,050 Hz [204, 205]. The slow speech is obtained by instructing speakers to speak slowly. The fast speech is obtained by instructing speakers to speak fast. These speaking styles are relevant to intelligibility study and clinical management of dysarthria.

For the purposes of phonetic labeling, a phonetic expert did phoneme transcription for 25 sentences of on arbitrary speaker in clear style (since labeling in this condition is likely to be the most accurate). The speaker is known as "template speaker". I then find the phoneme boundary for the unlabelled utterance in the remaining conditions of the template speaker using dynamic time warping (DTW). Specifically, the acoustic stream of an unlabelled (*query*) utterance is aligned with the template utterance with the same linguistic content. I transfer the phoneme boundary and phoneme labels from the template utterance to the query utterance.[2] I then manually edit the automatically-generated phoneme boundary of the query utterance, as well as inserting and deleting phonemes (e.g., pause), but not substituting phonemes. I choose not to modify phoneme identities to allow for a direct comparison, even though these types of changes are typical when changing speaking rates (e. g. a plosive may become flapped, or a vowel may become centralized). After obtaining verified label files for one speaker, I align the unlabelled speakers' sentence in all conditions, and then edit these in the same manner as previously. Example spectrum are shown in Figures 6.2 and 6.3.

[2]One can use Montreal Forced Aligner [139] as a better way to obtain the phoneme boundary.

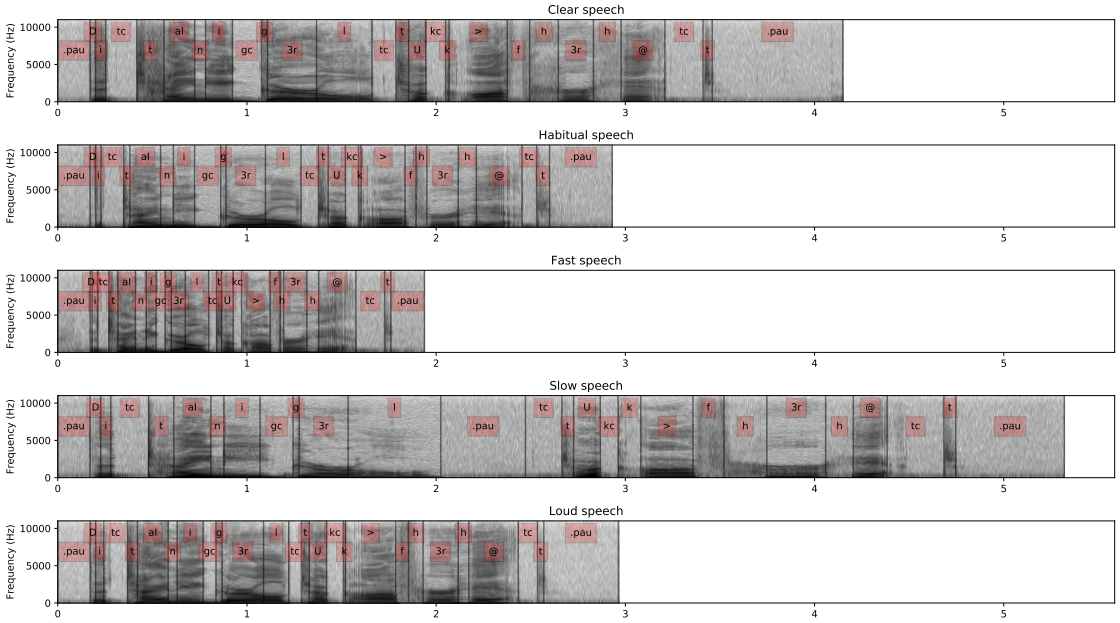


Figure 6.2: Spectrum and phonetic labels of one sentence in different conditions for the template speaker.

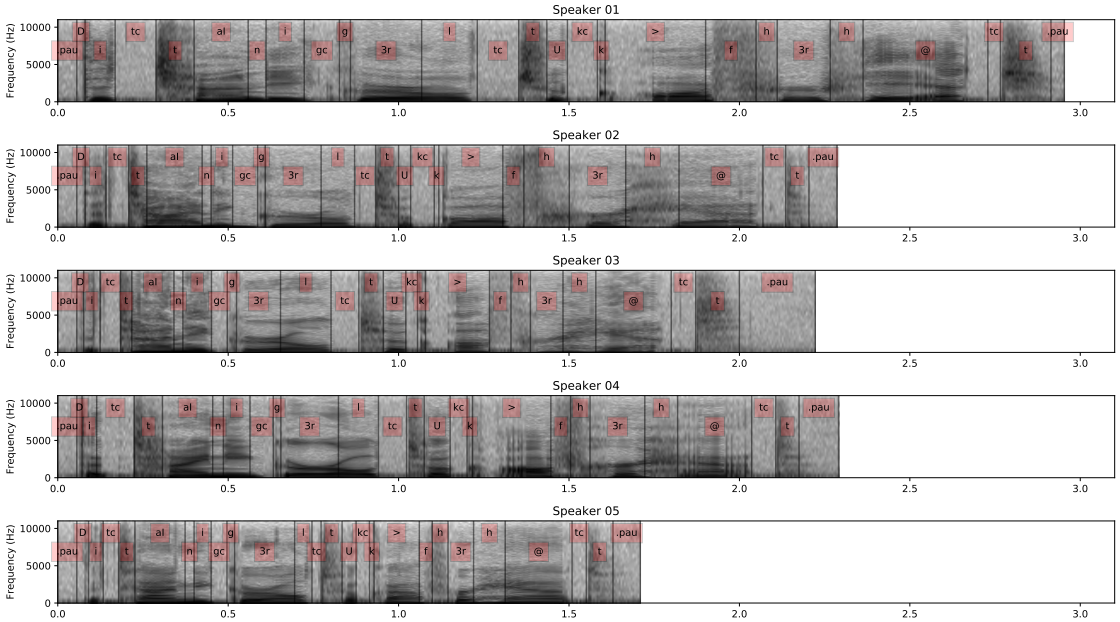


Figure 6.3: Spectrum and phonetic labels of one sentence in the habitual condition produced by five speakers.

6.5 Duration Analysis

In this section, I investigate how consistent speakers are when varying sentence and phoneme durations in different speaking styles (Section 6.5.1). I also explore the duration variation between speaking style pairs at sentence and phoneme level (Section 6.5.2). Finally, I determine whether a linear relation between phoneme and sentence-level scaling factors is sufficient (Section 6.5.3).

6.5.1 Consistency of Duration of Speakers in Each Speaking Style

The goal of the experiment is to learn if speakers have a common understanding of each speaking style. I examine how consistent speakers are in their sentence and phoneme durations in each speaking style. I do this by examining the standard deviation of the scaling factors.

Sentence-level

I calculate the scaling factors of identical sentences between two different speakers in an identical speaking style. For each speaking style (condition) $c \in [\mathrm{H}, \mathrm{C}, \mathrm{L}, \mathrm{S}, \mathrm{F}]$ and sentence u, I calculate the log-transformed scaling factor going from speaker s_i to speaker s_j as

$$f^c_{s_i \to s_j, u} = \log_2 \frac{d^c_{s_j,u}}{d^c_{s_i,u}} \tag{6.1}$$

where d denotes the duration of the sentence u without initial and final pauses.[3] Each speaker pair (s_i, s_j) only appears once in the calculation. Additionally, I arbitrarily named a speaker with *lower* speaker-id as s_i, and a speaker with *higher* speaker-id as s_j. For each speaking style and each speaker pair, I calculate the average of log-transformed scaling factors $\overline{f^c_{s_i \to s_j}}$ by averaging over all available sentences.[4] Due to my limited data, I assume the distribution of $\overline{f^c_{s_i \to s_j}}$ is a *normal* distribution.[5] I estimate the distribution using *kernel density estimation*. Figure 6.4 shows the distribution of these terms for each speaking style; the standard deviations are $\sigma^{\mathrm{H}} = 0.25$, $\sigma^{\mathrm{C}} = 0.4$, $\sigma^{L} = 0.3$, $\sigma^{S} = 0.5$, $\sigma^{\mathrm{F}} = 0.22$. By comparing kurtosis (a measure of the tailedness of a distribution), I realize that the duration is most consistent between speakers in fast style, while the the duration is less consistent between speakers in slow style. There appears to be a common speaking rate when speaking fast. In contrast, people have different speaking rates

[3] Calculating the average of the log-transformed scaling factors is equivalent to calculating the log of the geographic mean.

[4] Note that, instead of using Equation 6.1, I can obtain the same values of $\overline{f^c_{s_i \to s_j}}$ using $\overline{\log_2 d^c_{s_j}} - \overline{\log_2 d^c_{s_i}}$ where $\overline{\log_2 d^c_{s_i}}$ and $\overline{\log_2 d^c_{s_j}}$ are average of log-transformed sentence durations d, in speaking style c, of speakers s_i and s_j, respectively.

[5] I realize that for each speaker pair (s_i, s_j), one could order them randomly or also include (s_j, s_i). All three approaches should result in the same standard deviation.

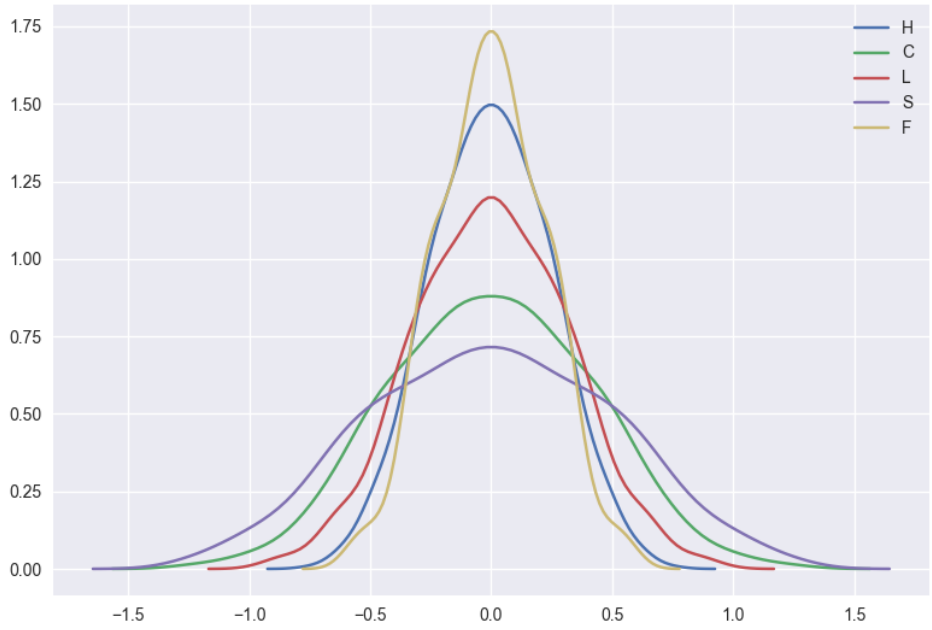


Figure 6.4: Kernel-density-estimate of the average of log-transformed scaling factors $\overline{f^c_{s_i \to s_j}}$ between speakers, averaged over all sentences, for each matched condition.

when speaking slowly. As a result, predicting scaling factors of a speaker-independent fast-to-slow duration conversion might be challenging due to the inconsistency in speakers.

Phoneme-level

To calculate the phoneme-level scaling factors, I perform a DTW-alignment of the phoneme labels to handle the potential differences in phoneme sequences when comparing speakers. Specifically, two phoneme sequences of the same sentence from two speakers are matched up using a simple edit distance. I do not consider pause labels in the phoneme alignment. For each speaking style c, and sentence u, I calculate the log-transformed scaling factor for the n^{th} phoneme token from speaker s_i to s_j as

$$f^c_{s_i \to s_j, u}(n) = \log_2 \frac{d^c_{s_j,u}(n)}{d^c_{s_i,u}(n)} \tag{6.2}$$

where $d(n)$ represents the n^{th} phoneme's duration, with $n = 1, \ldots, N^c_{s_i \to s_j, p}$, the number of available phonemes p for a particular condition c, sentence, and source and target speakers. Finally, I calculate the average scaling factor for each *type* of phoneme p via

$$\overline{f_{s_i \to s_j, p}} = \frac{1}{N_{s_i \to s_j, p}} \sum_{c \in [\text{H,C,L,S,F}]} \sum_{u=1}^{24} \sum_{n \in P^c_{s_i \to s_j, u}} f^c_{s_i \to s_j, s}(n) \tag{6.3}$$

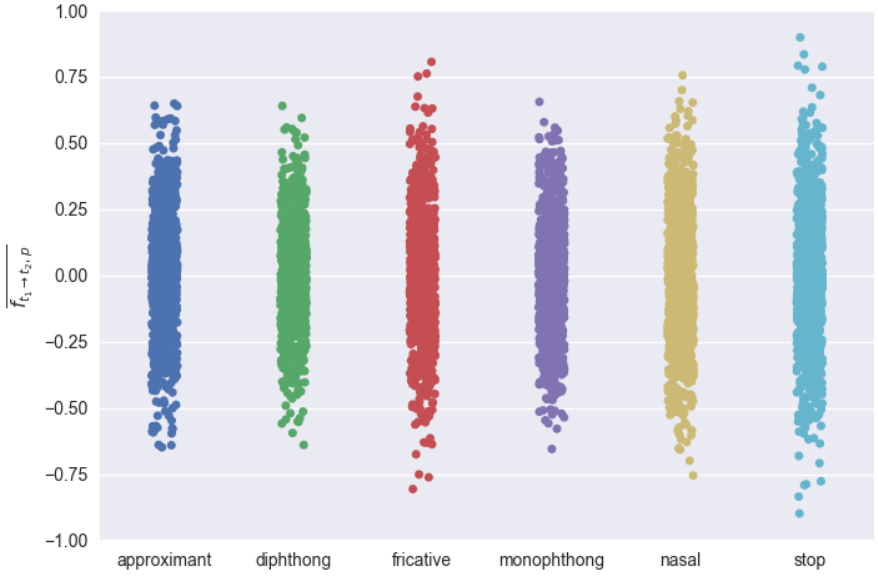


Figure 6.5: Distribution of average phoneme-level scaling factors

where $P^{c}_{s_i \to s_j, u}$ are the set of indices where p occurs, and $N_{s_i \to s_j, p}$ is the number of total occurrences of the phoneme over all conditions and all sentences, given particular source and target speakers.[6] Because the durations of closure phonemes following pauses are ambiguous, I did not considered those phonemes. Figure 6.5 shows distribution of average phoneme scaling factors in different phoneme categories. We see that diphthong and monophthong have more variation than fricative and nasal. Since we pooled all the speaking styles and we know that slow has the most variation, so the variation in fricative and nasal is probably in the slow speech condition. In future work, it may be helpful to examine the phoneme variation in each speaking style.

6.5.2 Duration Variation between Speaking Styles

The goal of the experiment is to learn how different speakers understand speaking styles. I examine how speakers vary sentence-level and phoneme-level durations in multiple speaking styles.

[6] Note that (instead of using Equation 6.3) we can obtain the same values of $\overline{f_{s_i \to s_j, p}}$ using $\overline{\log_2 d_{s_j, p}} - \overline{\log_2 d_{s_i, p}}$ where $\overline{\log_2 d_{s_j, p}}$ and $\overline{\log_2 d_{s_i, p}}$ are average of log-transformed sentence durations d of phoneme p, of speakers s_j and s_i, respectively.

from \ to	(C)lear	(L)oud	(S)low	(F)ast
(H)abitual	0.86 (0.32) [0.22, 1.46]	0.19 (0.2) [−0.24, 0.67]	1.09 (.36) [0.52, 2.10]	−0.45 (0.14) [−0.67, −0.13]
(C)lear		−0.68 (0.33) [−1.50, −0.15]	0.23 (0.29) [−0.55, 0.92]	−1.31 (0.3) [−1.94, −0.74]
(L)oud			0.9 (0.37) [0.30, 1.89]	−0.64 (0.23) [−1.19, −0.20]
(S)low				−1.54 (0.38) [−2.44, −0.86]

Table 6.1: Log-transformed scaling factor $\overline{f_s^{\text{from}\to\text{to}}}$ means, standard deviations, minima, and maxima between conditions. Positive values show slowing down, negative values show speeding up. Numbers change sign when reversing "to" and "from".

Sentence-level

I calculate the scaling factors of identical sentences between two different speaking style conditions. Specifically, for each speaker $s = 1, \ldots, 32$ and utterance $u = 1, \ldots, 24$, I calculate the log-transformed scaling factor going from a "from" to a "to" condition

$$f_{s,u}^{\text{from}\to\text{to}} = \log_2 \frac{d_{s,u}^{\text{to}}}{d_{s,u}^{\text{from}}} \tag{6.4}$$

where d represents the duration of the sentence (to improve accuracy, I discard the initial and final pauses). For each speaker, I calculate the mean scaling factor $\overline{f_s^{\text{from}\to\text{to}}}$ by averaging over all available sentences.[7] Table 6.1 gives us the statistics of this terms over all speakers. The pairs of (fast, slow), (loud, slow), and (habitual, slow) have standard deviations of 0.38, 0.37, and 0.36, respectively, which is the highest standard deviations. The high standard deviations can be attributed to the fact that people have different speaking rates when speaking slowly (see Section 6.5.1). For visualization purposes, I also show a graph of the average $\overline{f^{\text{from}\to\text{to}}}$ in Figure 6.6.[8] These last scaling factors represent the values that can be used in a simple linear scaling approach of duration conversion, i. e. all phoneme durations are scaled by the same factor, likely resulting in unnatural-sounding speech.

I observe that the greatest change took place between the slow and the fast speaking style conditions, as expected. I also observe that the loud condition was nearly of the same duration as the habitual condition. In terms of duration variation between speakers, the greatest variance

[7] Note that (instead of using Equation 6.4) we can obtain the same values of $\overline{f_s^{\text{from}\to\text{to}}}$ using $\overline{\log_2 d_s^{\text{to}}} - \overline{\log_2 d_s^{\text{from}}}$ where $\overline{\log_2 d_s^{\text{from}}}$ and $\overline{\log_2 d_s^{\text{to}}}$ are average of log-transformed sentence durations d, of speaker s, of speaking styles from and to, respectively.

[8] As a result of using logarithm, the average log-transformed scaling factors are additive. For example, The H-to-C value (0.86) equals to the sum of H-to-L (0.19) value and L-to-C value (0.68).

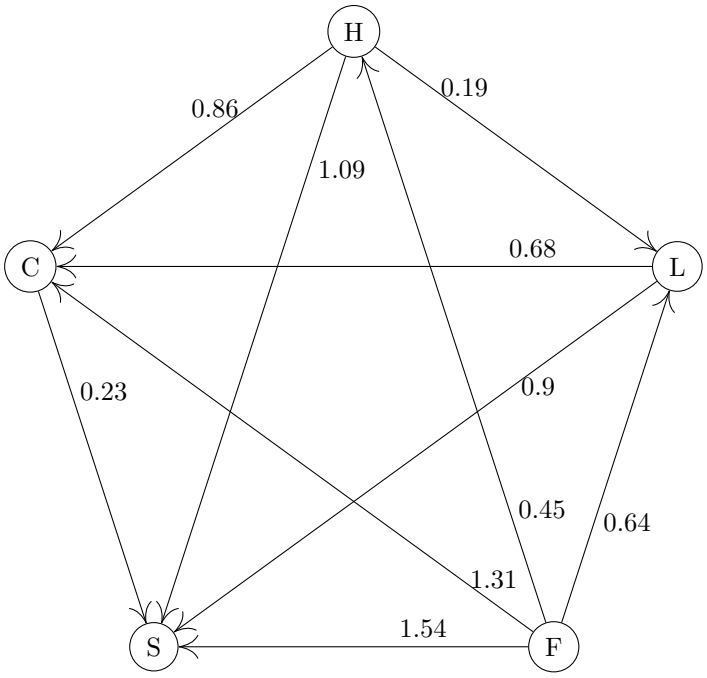


Figure 6.6: Average log-transformed scaling factors $\overline{f^{\text{from}\to\text{to}}}$ between conditions, averaged over all speakers. Arrow direction is the direction of slowing down speech. Note that arrows are additive and numbers change sign when reversing.

was found when going from slow to fast; the minimum scaling factor was $2^{-2.44} = 0.18$ and the maximum was $2^{-0.86} = 0.55$. Thus, I select the pair of fast and slow for duration conversion (Section 6.7). I also select the pair of habitual and slow for duration conversion because habitual speech is commonly used in real-life applications. As a result, slow speech is selected as target speaking style instead of *clear* speech as in previous chapters (see Chapter 3 and Chapter 4).

Phoneme-level

I consider the phoneme-level effects of different speaking styles on duration. To handle the potential differences in phoneme sequences when comparing speaking styles, I perform the same phoneme alignment as in Section 6.5.1, phoneme-level. Then, for each speaker t, and for each sentence s, I calculate the log-transformed scaling factors for the n^{th} phoneme *token*

$$f_{s,u}^{\text{from}\to\text{to}}(n) = \log_2 \frac{d_{s,u}^{\text{to}}(n)}{d_{s,u}^{\text{from}}(n)} \tag{6.5}$$

where $d(n)$ represents the n^{th} phoneme's duration, with $n = 1, \ldots, N_{t,s}^{\text{from}\to\text{to}}$, the number of available phonemes for a particular speaker, sentence, and source and target conditions. Finally, I

Phone p	$\overline{d_p^{\mathbf{S}}}$	$\overline{d_p^{\mathbf{F}}}$	$\overline{d_p^{\mathbf{F}} - d_p^{\mathbf{S}}}$	$\overline{f_p^{\mathbf{S}\to\mathbf{F}}}$ (↑)	$\overline{d_p^{\mathbf{F}} - d_p^{\mathbf{S}} \cdot 2^{\overline{f^{\mathbf{S}\to\mathbf{F}}}}}$
monophthong	255	81	-174.0	-1.6	-7.0
approximant	161	58	-104.0	-1.5	2.0
fricative	178	67	-112.0	-1.5	5.0
nasal	162	59	-104.0	-1.5	3.0
stop	85	29	-56.0	-1.3	0.0
diphthong	356	132	-225.0	-1.3	9.0

Table 6.2: Average phoneme-level durations of slow ($\overline{d_p^{\mathrm{S}}}$) and fast ($\overline{d_p^{\mathrm{F}}}$) speaking styles, their difference $\overline{d_p^{\mathrm{F}} - d_p^{\mathrm{S}}}$ and the associated scaling factor $\overline{f_p^{\mathrm{S}\to\mathrm{F}}}$ (the table is sorted on this)

calculate the average scaling factor for each *type* of phoneme p via

$$\overline{f_p^{\text{from}\to\text{to}}} = \frac{1}{N_p^{\text{from}\to\text{to}}} \sum_{s=1}^{32} \sum_{u=1}^{24} \sum_{n \in P_{s,u}^{\text{from}\to\text{to}}} f_{s,u}^{\text{from}\to\text{to}}(n) \tag{6.6}$$

where $P_{s,u}^{\text{from}\to\text{to}}$ are the set of indices where p occurs, and $N_p^{\text{from}\to\text{to}}$ is the number of total occurrences of the phoneme over all speakers and all sentences, given particular source and target conditions.[9]

I report on the most extreme case, namely when going from the slow to the fast condition (or vice versa). The results are shown in Table 6.2. Note that I ignore stop closures that follow pauses, as their duration is ambiguous. I observe that vowels have the most variation when changing speaking rate. Moreover, a linear and uniform scaling approach has significant errors compared to real phoneme duration. Therefore, a non-uniform scaling approach should be used for changing speaking rate.

6.5.3 Sentence and Phoneme-level Scaling Factors

In this section, I determine whether duration variation is phoneme specific. I examine the linear relationship between sentence-level scaling factors and phoneme-level scaling factors.

A sentence-level scaling factor is defined as a ratio between a source sentence's duration and a target sentence's duration. To calculate the duration of each sentence, I first align the phoneme sequence of the sentence in habitual speaking style to its corresponding phoneme sequences in the other four styles (clear, slow, fast, loud) to determine the common phonemes across all styles same as in Section 6.5.1. I then add up the durations of the common phonemes to obtain the sentence durations in the five styles. Finally, I calculate the sentence-level scaling factors from habitual to

[9]Note that (instead of using Equation 6.6) we can obtain the same values of $\overline{f_p^{\text{from}\to\text{to}}}$ using $\overline{\log_2 d_p^{\text{to}}} - \overline{\log_2 d_p^{\text{from}}}$ where $\overline{\log_2 d_p^{\text{to}}}$ and $\overline{\log_2 d_p^{\text{from}}}$ are average of log-transformed sentence durations d of phoneme p, of speaking styles to and from, respectively.

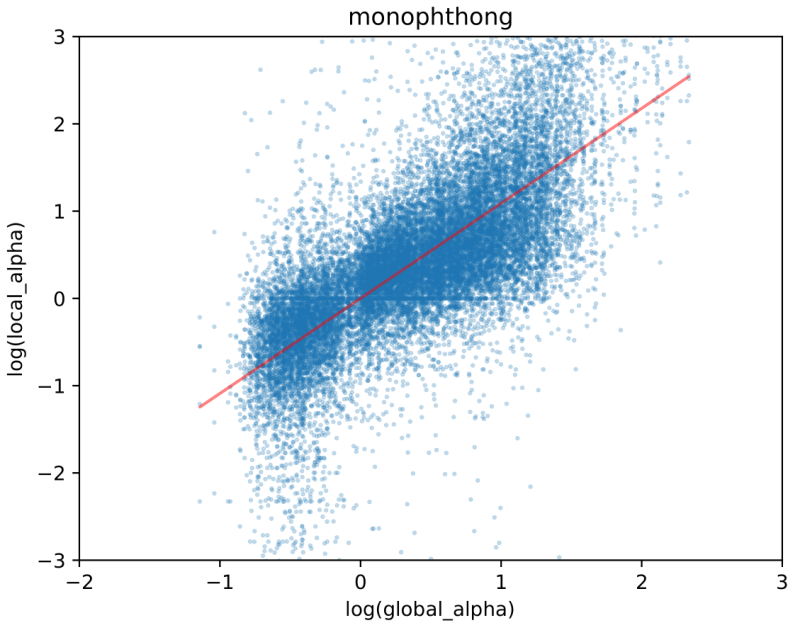


Figure 6.7: least squared regression between sentence (global) and phoneme-level (local) scaling factors of monophthong

the other styles by dividing the sentence duration in target style to that in habitual. Moreover, I calculate the phoneme-level scaling factor from habitual to the other four styles on the common phonemes by dividing the target phoneme durations to the aligned habitual phonemes' durations.

I examine the linear relationship between sentence-level scaling factors and phoneme-level scaling factors in log-transformed scaling factors. Figure 6.7 shows a least squared regression between sentence and sentence-level scaling factors with the intercept at zero. Table 6.3 shows the slope of the linear regression lines. The stops g, b, d has a slope less than this of stops p, t, k because these stops p, t, k includes burst and aspiration, and the aspiration has a lot of temporal flexibility. A slope of 1.00 reports that an uniform time-modification is sufficient for the phoneme category. The slope less than 1.00 of some phoneme categories such as stops: g, b, d confirm my hypothesis that uniform time scaling is not sufficient for some phoneme categories.

6.6 Predict phoneme-level scaling factor

In this section, I present my preliminary method to predict phoneme-level scaling factors from a sentence-level scaling factor and phoneme-level slopes (Table 6.3). My goal is to build a non-uniform duration conversion procedure that takes as input a source sentence and a sentence-level

Category c	Slope	#datapoints (%)
closures	0.96	18.6
monophthong	1.09	24.2
diphthong	0.86	6.4
stops:g,b,d	0.63	6.3
stops:p,t,k	0.95	9.8
fricative	1.00	16.8
nasal	1.02	5.8
approximant	1.12	9.3
other	0.84	2.8

Table 6.3: Slope of regression lines between sentence and phoneme-level scaling factors for each phoneme category

scaling factor.

In order to utilize the phoneme-level slopes (Table 6.3), the method requires phoneme labels and phoneme boundaries of the source sentence. For a given phoneme p with duration d_p in the source sentence, the phoneme duration d'_p is determined as follows

$$d'_p = f_c \times d_p \tag{6.7}$$

where f_c denotes a phoneme-level scaling factor for phoneme category c, c is phoneme category of phoneme p (e.g., monophthong). The phoneme-level scaling factor is defined as follows

$$f_c = (\alpha - 1) \times s_c + 1 \tag{6.8}$$

where α denotes the sentence-level scaling factor, and s_c denotes phoneme-level slope for phoneme category c (as in Table 6.3).[10] Note that $\alpha = 1$, which means no change, results in $d'_p = d_p$ as should be the case.

Using this approach, one can obtain a piece-wise phoneme-level scaling factor trajectory over time using phoneme labels and phoneme boundaries. Within a phoneme segment, the phoneme-level scaling factor f_c is a constant number.

In real applications, however, phoneme labels are not available. Instead, one can use a *pre-trained phoneme classifier* to obtain the likelihoods of each phoneme p. One can calculate the phoneme-level scaling factor trajectory as follows

$$f_c(t) = (\alpha - 1) \times \sum_p s_c \times \text{Prob}(p, t) + 1 \tag{6.9}$$

[10] According to Equation 6.8, these phoneme-level slopes are different from those in Table 6.3. One can use a similar procedure as in Section 6.6 to obtain phoneme-level scaling factors f_c and sentence-level scaling factors α. However, these slopes s_c are linear coefficients of linear regression lines that satisfy Equation 6.8. Note that those slopes s_c in Table 6.3 satisfy $\log_2 f_c = s_c \times \log_2 \alpha$. As we can see, those slopes in Table 6.3 are in logarithm. but it is not clear that we can use those slopes here

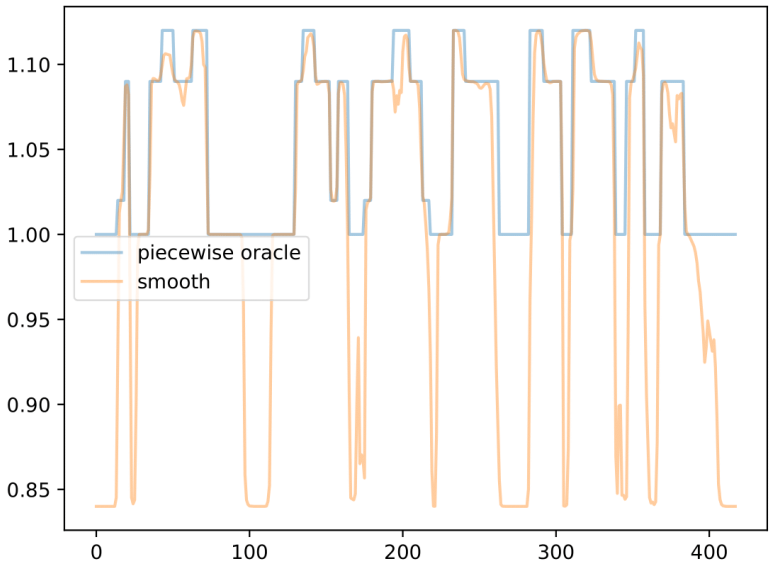


Figure 6.8: Piece-wise vs smooth scaling factor trajectory on a TIMIT sentence. The piece-wise trajectory requires phoneme labels and phoneme boundaries; While, the smooth trajectory requires a pre-trained phoneme classification.

where $f_c(t)$ denotes phoneme-level scaling factor at time t, c is the phoneme category of phoneme p, α denotes the sentence-level scaling factor, s_c denotes phone-level slope, and $\text{Prob}(p, t)$ denotes the probability of seeing phoneme p at time t. The probability is calculated using the pre-trained phoneme classifier. Because of the probability of phoneme labels (e.g. at the transition area of two phonemes), the phoneme-level scaling factor trajectory becomes a smooth trajectory. Specifically, at the transition area of two phonemes, the probability of past phonemes gradually reduces; while, the probability of current phoneme gradually increases. Thus, the phoneme-level slope of past phoneme also reduces; while, the slope of current phoneme increases. The fading-out of past phoneme's slope and the fading-in of current phoneme's slope create a smooth transition of phoneme-level slopes.

Note that the scaling factors in Figure 6.8 are not calculated using equations than Equation 6.8 and Equation 6.9. Instead, for piece-wise, I calculated the scaling factor as $\log_2 f_c = s_c \times \log_2 \alpha$. For smooth case, I calculated the scaling factor as $\log_2 f_c(t) = \log_2 \alpha \times \sum_p s_c \times \text{Prob}(p, t)$. By looking at the smooth trajectory of phoneme-level scaling factors, I determined that 85% of the time the error is less than 0.01.

6.7 Duration Conversion with Oracle Scaling Factors

In Section 6.6, I ran some preliminary tests for predicting scaling factors, now I only focus on time-scale modification; thus, I use oracle scaling factors. In this section, I compare the performance of uniform (sentence-level) and non-uniform (phoneme-level) duration conversion in terms of intelligibility and naturalness. Recall that duration conversion consists of two main steps: 1) predicting target scaling factors (see Section 6.2), and 2) time-scale modification (see Section 6.3). As discussed in Section 6.2, I only focus on the second task of time-scale modification. For target scaling factors, I use an oracle instead of predicting them. Specifically, for an individual sentence, I know how the speaker changes the sentence between its fast style and slow style. I calculate the oracle scaling factors for the sentence between the two speaking styles. Then I apply the oracle scaling factor on the fast speech to obtain the modified speech, using the time-scale modification procedure. In this way, the final intelligibility and naturalness of the modified speech is only affected by the interaction between time-scale modifications and oracle phoneme and sentence-level scaling factors.

I explore two different style conversions: habitual-to-slow and fast-to-slow. I include habitual-to-slow conversion because habitual speech is commonly used in real-life applications. I evaluate the naturalness and intelligibility of the modified speech. I include fast-to-slow conversion because the duration differences between fast speech and slow speech are greater than those between habitual speech and slow speech (see Section 6.5). For fast-to-slow conversion, I evaluate the intelligibility of the modified speech but not naturalness.

6.7.1 Habitual-to-Slow Duration Conversion

In this section, I convert habitual speech to slow speech using a time-scale modification procedure, *phase vocoder with identity phase lock*, given oracle sentence or phoneme-level scaling factors. For an individual habitual sentence, we create two sentences. For the first, we apply an oracle sentence-level scaling factor to uniformly convert sentence duration. For the second, we apply oracle phoneme-level scaling factors and habitual phoneme labels to non-uniformly convert phoneme durations.

I selected 14 speakers to manually correct the phoneme labels of habitual speech and slow speech. I selected 10 out of 25 Harvard sentences with the highest difference in sentence duration. In total, I obtained 14 (speakers) $\times$ 10 (sentences) $\times$ 2 (speaking styles) = 280 label files. Because of the variation of phoneme sequences between habitual speech and target slow speech, I first align the phoneme sequences of habitual speech and target slow speech. Then I only kept the common

Speaker ID \ Preference test	Naturalness	Intelligibility
1	-0.3	-0.45*
4	0.125	-0.1
6	-0.25	-0.325
12	*0.475**	*0.425**
19	-0.25	-0.375*
20	-0.75*	-0.56*
22	-0.2	-0.125
24	0	-0.375*
25	-0.525*	-0.525*
26	-0.375	-0.175
28	0.225	0.1
30	*0.45**	0.225
31	-0.075	-0.45*
33	0.15	-0.175

Table 6.4: Naturalness and intelligibility preference test for habitual-to-slow duration conversion. Positive scores means non-uniform (phoneme-level) is better. Asterisk shows significant difference from zeros in a two-tailed t-test

phonemes for each pair of habitual speech and target slow speech. I modify the *aligned* habitual speech using phase vocoder to achieve the duration of *aligned* slow speech.

There are 10 (sentences) $\times$ 14 (speakers) = 140 trials. Each trial consists of uniformly modified speech and non-uniformly modified one. In total, there are 280 stimuli.[11] I mixed the stimuli with babble noise at 0dB SNR. I recruited participants on Amazon Mechanical Turk. All participants are living in the U. S., and they are required to have an acceptance rate of 95% on Amazon Mechanical Turk. I needed 14 participants to cover all 140 trials.

I conduct the first preference test to compare the intelligibility between uniformly modified speech and non-uniformly modified speech. For each trial, a participant listened to a pair of stimuli; then, they answer "if the second one is more intelligible than the first one?" with a five-point scale: "much worse" (-2), "worse" (-1), "same" (0), "better" (1), "much better" (2). There were 56 participants. Thus, each pair was listened to four times. Table 6.4 showed the result of this intelligibility test. There is no case in which non-uniform modification significantly outperforms uniform modification in terms of intelligibility.

I did the second preference test to the evaluate naturalness of time-scale modification. The configuration of this test is similar to that of the intelligibility preference test. For each trial of the naturalness preference test, a participant listen to a pair of stimuli; then, they answer "if the second one is more natural than the first one?" with a five-point scale: "much worse" (-2), "worse"

[11] Samples for the experiment are available at https://tuanad121.github.io/samples/2021-07-10-Duration/

Preference test / Speaker ID	Intelligibility
1	-0.15
4	0.125
6	-0.375*
12	0.325
19	0.1
20	-0.6*
22	-0.4*
24	-0.2
25	-0.55*
26	-0.475*
28	-0.275
30	0.05
31	-1
33	-0.125

Table 6.5: Intelligibility preference test for fast-to-slow duration conversion. Positive scores means non-uniform (phoneme-level) is better. Asterisk means significantly different from zeros in a two-tailed t-test

(-1), "same" (0), "better" (1), "much better" (2). There were 56 participants. Table 6.4 showed the results of this naturalness test. There are two speakers (12, 30) with a significantly better performance of non-uniform modification; while, there are two speakers (20, 25) with significantly better performance of uniform modification.

6.7.2 Fast-to-Slow Duration Conversion

In this section, I convert fast speech to slow speech using a time-scale modification procedure, *phase vocoder*, given oracle sentence or phoneme-level scaling factors. I conduct the third preference test to evaluate the intelligibility of the time-scale modification. The configuration of this intelligibility preference test is the same as previous intelligibility preference test (Section 6.7.1). For each trial, a participant listen to a pair of stimuli; then, they answer "if the second one is more intelligible than the first one?" with a five-point scale: "much worse" (-2), "worse" (-1), "same" (0), "better" (1), "much better" (2). There were 56 participants. Table 6.5 showed the results of this test. There is no case in which non-uniform modification significantly outperforms uniform modification in terms of intelligibility.

6.8 Conclusion

In my naturalness preference test for habitual-to-slow conversion, non-uniform conversion significantly outperformed uniform conversion for two speakers; however, the uniform conversion significantly outperformed the non-uniform conversion for other two speakers. My remaining experiments on intelligibility did not show that non-uniform conversion was significantly better than uniform conversion. I attribute the under-performance of non-uniform conversion to artifacts in the output signal. I discovered that the output signals of non-uniform conversion had more artifacts than those of uniform conversion, probably because non-uniform conversion is more complex than uniform conversion. I suggest improving the quality of the time-scale modification procedure before conducting further studies on non-uniform duration conversion.

Chapter 7

Conclusion

In this dissertation, I explored different approaches to improve the intelligibility of habitual speech, (mild) dysarthric speech and alaryngeal speech on the speaker side using well-established machine learning methods. To make intelligibility of habitual and dysarthric speech become more resilient to noise, I converted the speech into a special clear speaking style. To increase the quality and intelligibility of alaryngeal speech, which is harder to understand than dysarthric speech, I converted the alaryngeal speech into intelligible speech. In Section 7.1, I summarize my main contributions. In Section 7.2, I discuss future directions of the research.

7.1 Contributions

As outlined in Chapter 1, this dissertation has four objectives:

1. To determine effective spectral features for spectral voice and style conversion.
2. To develop effective HAB-to-CLR spectral mappings using well-established machine learning algorithms.
3. To develop effective conversion methods from alaryngeal speech to intelligible speech, using well-established machine learning algorithms.
4. To investigate the performance of uniform and non-uniform duration style conversion.

The first objective was to determine effective spectral features for spectral voice and style conversion (Chapter 3). I contrasted two new sets of spectral mapping features: 1) probabilistic peak tracking features (PPT), which are formant-like hand-crafted features, and 2) manifold features (VAE), which are machine learnable by a Variational Autoencoder. The two sets of features are integrated into a high quality vocoder, WORLD. I extensively evaluated the two sets of features by comparing them to each other and to baselines, which are two commonly-used

spectral representations: line spectral frequency (LSF) and mel-cepstrum coefficients (MCEP) in speech reconstruction, voice conversion, and style conversion. For each of the four types of features (VAE, PPT, LSF, MCEP), I specified the number of features (e.g., VAE-12 denotes VAE with 12 features).

In a speech reconstruction experiment, I showed that using VAE-12 achieved significantly better perceived speech quality compared to MCEP-12. VAE-12 also outperformed PPT-20 in terms of perceived speech quality. In a voice conversion experiment, I showed that mapping VAE-12 resulted in significantly better perceived speech quality compared to MCEP-40, with similar speaker accuracy, thus demonstrating the efficiency of mapping in a low-dimensional latent feature space. I also showed that VAE-12 outperformed LSF-20 in terms of similar speaker accuracy. In a *habitual* to clear style conversion experiment, I showed that VAE-12 together with a custom skip-connection deep neural network significantly improved the speech intelligibility of one of three speakers, with the average keyword recall accuracy increasing from 24% to 46%. The results was reported in Tuan Dinh, Alexander Kain, Kris Tjaden, *Using a Manifold Vocoder for Spectral Voice and Style Conversion*, Interspeech, 2019.

The second objective was to develop effective HAB-to-CLR spectral mappings using well-established machine learning algorithms (Chapter 4). I proposed a cGAN-based style conversion for mapping the manifold features of habitual speech to those of clear speech. I gave an overview of cGANs, I described the application of cGANs for mapping speaking styles, and I presented my configuration for the following experiments. Specifically, conditional Generative Adversarial nets (cGANs) were investigated with three mappings: one-to-one mappings, many-to-one mappings, and many-to-many mappings. For one-to-one mappings, I compared the performance of cGANs-based one-to-one mappings to my manifold model. For each of the three mappings, I extensively evaluated the performance of the cGANs on both typical speakers and speakers with mild dysarthria secondary to Parkinson's disease for intelligibility improvement in a noisy environment.

In the speaker-dependent one-to-one mapping case, I showed that the cGAN outperformed a DNN in terms of average keyword recall accuracy in all cases. Moreover, the cGAN significantly improved speech intelligibility of two of three speakers, compared to one speaker when using the DNN. In the speaker-independent many-to-one mapping case, I could significantly improve speech intelligibility of one of three speakers, with average keyword recall accuracy increasing from 17.6% to 34.4%. In the speaker-independent many-to-many mapping case, the cGAN improved average keyword accuracy over that of vocoded habitual speech for two speakers: CSM7 and PDF7, but without statistical significance. The result was reported in Tuan Dinh, Alexander Kain, Kris Tjaden, *Improving Speech Intelligibility through Speaker Dependent and Independent Spectral Style*

Conversion, Interspeech, 2020.

The third objective was to develop effective conversion methods from alaryngeal speech to intelligible speech, using well-established machine learning algorithms (see Chapter 5). I proposed an approach that has two parts and a F0 synthesis method for transforming alaryngeal speech (LAR speech) to intelligible speech (INT speech). The first part predicts voicing/unvoicing and the degree of voicing using feed-forward networks. The second part is for LAR-to-INT spectral mappings using a cGANs.

I provided the motivation for the research; and I reviewed the related works for increasing intelligibility and naturalness of alaryngeal speech. I discussed my data, including how to create my target data of intelligible speech, which is different from the clear speech of Chapter 3 and Chapter 4. I predicted voicing/unvoicing and the degree of voicing. I predicted the spectrum of intelligible speech from that of alaryngeal speech. I created a synthetic fundamental frequency trajectory with an intonation model consisting of phrase and accent curves to address the unusable fundamental frequency (F0) information of alaryngeal speech. We evaluated the LAR-to-INT conversion methods on data.

For predicting INT voicing/unvoicing and degree of voicing, using different context lengths did not have a noticeable impact. Moreover, pre-training the prediction networks on FU-, and FV-TIMIT as opposed to TIMIT did not have improved performance. Similarly, pre-training with FU-, and FV-TIMIT did not have improved performance for predicting MCEP of intelligible speech. Adaptation always improved performance. In my subjective tests with four LAR speakers, I significantly improved the naturalness of two speakers, and I significantly improved the intelligibility of one speaker. The results are promising for a challenging task with a lot of individual variability among four LAR speakers. This is reported in Tuan Dinh, Alexander Kain, Robin Samlan, Beiming Cao, Jun Wang, *Increasing the Intelligibility and Naturalness of Alaryngeal Speech Using Voice Conversion and Synthetic Fundamental Frequency*, Interspeech, 2020.

The fourth objective was to investigate the performance of uniform and non-uniform duration style conversion (Chapter 6). I showed preliminary results of improving speech intelligibility using duration conversion. In previous objectives, I conducted spectral style conversion to improve speech intelligibility. In addition to spectral features, phoneme duration also proved to be important in improving intelligibility during style conversion. Therefore, my goal was to improve intelligibility using duration style conversion. Specifically, I explored the non-uniform duration style conversion. My intuition is that a non-uniform duration conversion is better than a uniform one. However, my effort of conducting non-uniform style duration conversion did not outperform the uniform style duration conversion. I attribute the failure to the artifacts created by duration

conversion algorithms. Further work should be done to reduce the artifacts in order to examine the performance of duration conversion on speech intelligibility.

I presented an overview of duration conversion for improving speech intelligibility. I presented my duration conversion method; and I reviewed time-scale modification procedures. I presented the experimental data. Then, I analyzed how changing speaking style affect sentence and phoneme duration. Finally, we evaluated the performance of uniform and non-uniform duration conversion in terms of intelligibility in an ideal case when (oracle) sentence- and phoneme-level scaling factors were given.

My experiments did not show that non-uniform time-scale modification was significantly better than uniform one in terms of naturalness and intelligibility. The under-performance of non-uniform modifications can be attributed to artifacts in output signal. I discovered that the output signals of non-uniform modification had more artifacts than those of uniform modification. I suggest to improve the quality of time-scale modification procedure, phase vocoder, before conducting further study on non-uniform duration conversion.

7.2 Future Direction

One area to improve is the HAB-to-CLR style conversion. Brenk showed that phoneme duration of slow speech is not a contributing factor to the intelligibility of slow speech [27]. However, the combination of phoneme duration and spectrum of slow speech is a contributing factor to intelligibility of slow speech. Therefore, combining spectral conversion and duration conversion in style conversion is a future direction for improving intelligibility of habitual and dysarthric speech. One approach is to use unified models for modelling both spectral mappings and duration conversion simultaneously. My intuition is that people change duration and spectrum at the same time when changing their speaking style. By jointly modelling the changes of duration and spectrum, I can create better models for both duration and spectrum. Currently, the sequence-to-sequence and transformer models show their potential for modelling spectrum and duration simultaneously in voice conversion [101, 99]. Thus, future studies should investigate the efficacy of these models in style conversion for improving speech intelligibility.

Another area to improve is a thorough evaluation of HAB-to-CLR style conversion. When evaluating my HAB-to-CLR style conversion methods, I considered a challenging noise condition, babble noise at 0dB SNR. However, the intelligibility of CLR speech depends on noise conditions and listeners' hearing ability [173, 60, 25, 121]. Thus, there is a need for further evaluation of the method with different noise conditions to understand more about the efficacy of my mapping.

Moreover, all listeners in my evaluation have normal hearing ability. There remains a question about the relationship between the efficacy of my mappings and listeners' hearing ability. A future direction is to investigate the efficacy of my style conversion methods with different noise conditions and listeners' hearing ability.

Another area to improve is the LAR-to-INT conversion. The small amount of LAR speech probably limits the performance of DNNs models. Thus, one approach is to pre-train the models on synthetic data that simulates LAR speech. In Section 5.4 and 5.5, pre-training did not show noticeable effect, which suggests that our pre-trained dataset did not properly simulate LAR speech. I suspect that LAR speech is different from INT speech in many aspects other than voicing. There should be further investigation on the properties of LAR speech in order to generate a better pre-trained dataset.

Appendix A

List of Speech Stimuli

The 25 Harvard sentences with keywords in bold were used in keyword recall accuracy test to evaluate sentence-level intelligibility.

1. **Glue** the sheet to the **dark blue background**
2. The **box** was **thrown beside** the **parked truck**
3. **Four hours** of **steady work faced** us
4. The **hogs** were **fed chopped corn** and **garbage**
5. The **soft cushion broke** the **man's fall**
6. The **girl** at the **booth sold fifty bonds**
7. She **blushed when** he **gave** her a **white orchid**
8. **Note closely** the **size** of the **gas tank**
9. The **square wooden crate** was **packed** to be **shipped**
10. He **sent** the **figs**, but **kept** the **ripe cherries**
11. A **cup** of **sugar makes sweet fudge**
12. **Place** a **rosebush** near the **porch steps**
13. A **saw** is a **tool used** for **making boards**
14. The **dune rose from** the **edge** of the **water**
15. The **ink stain dried** on the **finished page**
16. The **harder** he **tried** the **less** he **got done**

17. **Paste** can **cleanse** the **most dirty brass**
18. The **ancient coin** was **quite dull** and **worn**
19. The **tiny girl took off** her **hat**
20. The **pot boiled**, but the **contents failed** to **jell**
21. The **sofa cushion** is **red** and of **light weight**
22. An **abrupt start** does **not win** the **prize**
23. These **coins will** be **needed** to **pay** his **debt**
24. **Hoist** the **load** to **your left shoulder**
25. **Burn peat** after the **logs give out**

Bibliography

[1] American cancer society 2020.

[2] *International Electrotechnical Commission, Electroacoustics-sound level meters-part 1: Specifications, 61672*, 2002.

[3] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara. Voice conversion through vector quantization. In *ICASSP*, 1988.

[4] M. Airaksinen, B. Bollepalli, L. Juvela, Z. Wu, S. King, and P. Alku. Glottdnn — a full-band glottal vocoder for statistical parametric speech synthesis. In *Interspeech 2016*, pages 2473–2477, 2016.

[5] M. Alhamid. Generative adversarial networks gans: A beginner's guides. https://towardsdatascience.com/generative-adversarial-networks-gans-a-beginners-guide-f37c9f3b7817, 2020.

[6] R. H. Ali and S. B. Jebara. Esophageal speech enhancement using excitation source synthesis and formant patterns modification. In *Signal-Image Technology & Internet Based Systems*, pages 615–624, 2006.

[7] A. Amano-Kusumoto. *Relationship between acoustic features and speech intelligibility: a dissertation*. PhD thesis.

[8] A. Amano-Kusumoto and J. Hosom. A review of research on speech intelligibility and correlations with acoustic features. *Center for Spoken Language Understanding, Oregon Health and Science University (Technical Report CSLU-011-0001)*, 2011.

[9] H. AR, T. CW, and F. DA. Effects of different frequency response strategies upon recognition and preference for audible speech stimuli. *J Speech Hear Res*, 34:1185–1196, 1991.

[10] L. Arslan and D. Talkin. Speaker transformation using sentence hmm based alignments and detailed prosody modification. In *Proceedings of the 1998 IEEE International Conference*

on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181), volume 1, pages 289–292 vol.1, 1998.

[11] L. M. Arslan. Speaker transformation algorithm using segmental codebooks (stasc). *Speech Communication*, 28(3):211–226, 1999.

[12] L. M. Arslan and D. Talkin. Voice conversion by codebook mapping of line spectral frequencies and excitation spectrum. In *EUROSPEECH*, 1997.

[13] B. Atal and M. Schroeder. Predictive coding of speech. In *Conf. Communications and Proc.*, 1967.

[14] E. Azarov, M. Vashkevich, D. Likhachov, and A. Petrovsky. Real-time voice conversion using artificial neural networks with rectified linear units. In *INTERSPEECH*, 2013.

[15] B. J. Bailey, J. T. Johnson, and S. D. Newlands. *Head & Neck Surgery–Otolaryngology*, volume 1. Lippincott Williams & Wilkins, 2006.

[16] O. I. Ben, J. D. Martino, and K. Ouni. Enhancement of esophageal speech obtained by a voice conversion technique using time dilated fourier cepstra. *International Journal of Speech Technology*, 22(1):99–110, 2019.

[17] R. M. Bhat, J. Singh, and P. Lehana. Investigations of the effect of nonlinearly generated excitations on the quality of the synthesized alaryngeal speech. *Indian journal of science and technology*, 10:1–12, 2017.

[18] N. Bi and Q. Yingyong. Speech conversion and its application to alaryngeal speech enhancement. In *ICSP*, pages 1586–1589, 1996.

[19] M. Blaauw and J. Bonada. Modeling and transforming speech using variational autoencoders. *Proceedings of INTERSPEECH*, 2016.

[20] B. Blesser. Audio dynamic range compression for minimum perceived distortion. *IEEE Transactions on Audio and Electroacoustics*, 17(1):22–32, 1969.

[21] B. Bollepalli, L. Juvela, and P. Alku. Generative adversarial network-based glottal waveform model for statistical parametric speech synthesis. Proceedings of Interspeech 2017; Interspeech: Annual Conference of the International Speech Communication Association, page 5, 2017.

[22] D. Bonardo and E. Zovato. Speech synthesis enhancement in noisy environment. *Proceedings of INTERSPEECH*, pages 2853–2856, 2007.

[23] B. Bozkurt, B. Doval, , and T. Dutoit. A method for glottal formant frequency estimation. In *Proceedings of ICSLP, International Conference on Spoken LanguageProcessing*, 2004.

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

[25] A. R. Bradlow, N. Kraus, and E. Hayes. Speaking clearly for learning-impaired children: sentence perception in noise. *Journal of Speech, language, and hearing research*, 46:80–97, 2003.

[26] A. R. Bradlow, B. M. Torretta, and D. B. Pisoni. Intelligibility of normal speech i: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication*, 20:255–272, 1996.

[27] F. V. Brenk, A. Kain, and K. Tjaden. Investigating acoustic correlates of intelligibility gains and losses during slowed speech: A hybridization approach. *American Journal of Speech-Language Pathology*, 2021.

[28] H. Brouckxon, W. Verhelst, and B. Schuymer. Time and frequency depedent amplification for speech intelligibility enhancement in noisy environments. *Proceedings of INTERSPEECH*, pages 557–560, 2008.

[29] R. H. Byrd, P. Lu, and J. Nocedal. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific and Statistical Computing*, 16(5):1190–1208, 1995.

[30] B. Cao, N. Sebkhi, T. Mau, O. T. Inan, and J. Wang. Permanent magnetic articulograph (pma) vs electromagnetic articulograph (ema) in articulation-to-speech synthesis for silent speech interface. In *Proceedings of the Eighth Workshop on Speech and Language Processing for Assistive Technologies*, pages 17–23, 2019.

[31] D. T. Chappell and J. H. Hansen. Speaker-specific pitch contour modeling and modification. In *ICASSP*, 1998.

[32] J. Chen, J. Benesty, Y. Huang, and S. Doclo. New insights into the noise reduction wiener filter. *IEEE Trans. Audio Speech Lang. Process*, 14(1):1218–1234, 2006.

[33] L.-H. Chen, Z.-H. Ling, Y. Song, and L.-R. Dai. Joint spectral distribution modeling using restricted boltzmann machines for voice conversion. In *INTERSPEECH*, 2013.

[34] L.-H. Chen, L.-J. Liu, Z.-H. Ling, Y. JIan, and L.-R. Dai. The ustc system for voice conversion challenge 2016: Neural network based approaches for spectrum, aperiodicity and f0 conversion. In *INTERSPEECH*, 2016.

[35] K. Chenausky and J. MacAuslan. Utilization of microprocessors in voice quality improvement: the electrolarynx. *Current Opinion in Otolaryngology & Head and Neck Surgery*, 8(3):138–142, 2000.

[36] D. G. Childers, K. Wu, D. HIcks, and B. Yegnanarayana. Voice conversion. *Speech Communication*, 8(2):147–158, 1989.

[37] S. Chintala. How to train a gan? https://github.com/soumith/ganhacks, 2016.

[38] M. Covell, M. Withgott, and M. Slaney. Mach1: nonuniform time-scale modification of speech. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, volume 1, pages 349–352 vol.1, 1998.

[39] A. del Pozo. *Voice source and duration modelling for voice conversion and speech repair*. PhD thesis, University of Cambridge, 2008.

[40] A. del Pozo and S. Young. Continuous tracheosophageal speech repair. In *the European Signal Processing Conference*, 2006.

[41] L. Deng, X. Cui, R. Pruvenok, J. Huang, S. Momen, Y. Chen, and A. Alwan. A database of vocal tract resonance trajectories for research in speech processing. In *ICASSP*, 2006.

[42] S. Desai, A. W. Black, B. Yegnanarayana, and K. Prahallad. Spectral mapping using artificial neural networks for voice conversion. In *IEEE Transactions on Audio, Speech, and Language Processing*, 2010.

[43] T. Dinh, A. Kain, R. Samlan, B. Cao, and J. Wang. Increasing the intelligibility and naturalness of alaryngeal speech using voice conversion and synthetic fundamental frequency. *in Proceedings of INTERSPEECH*, 2020.

[44] T. Dinh, A. Kain, and K. Tjaden. Using a manifold vocoder for spectral voice and style conversion. In *INTERSPEECH*, pages 1388–1392, 2019.

[45] T. Dinh, A. Kain, and K. Tjaden. Improving speech intelligibility through speaker dependent and independent spectral style conversion. *in Proceedings of INTERSPEECH*, 2020.

[46] M. Dolson. Computer musical journal. *The phase vocoder: a tutorial*, 10(44):14–27, 1986.

[47] B. Doval, C. D'Alessandro, and N. Henrich Bernardoni. The voice source as a causal/anticausal linear filter. In *VOQUAL'03*, page 1, Genève, Switzerland, 2003.

[48] J. Driedger and M. Müller. Tsm toolbox: Matlab implementations of time-scale modification algorithms. In *DAFx*, 2014.

[49] J. Driedger and M. Müller. A review of time-scale modification of music signals. *Applied Sciences*, 6(2), 2016.

[50] J. Driedger, M. Müller, and S. Ewert. Improving time-scale modification of music signals using harmonic-percussive separation. *IEEE Signal Processing Letters*, 21(1):105–109, 2014.

[51] R. Drullman, J. M. Festen, and R. Plomp. Effect of reducing slow temporal modulations on speech reception. *Journal of the Acoustical Society of America*, 95(5):2670–2680, 1994.

[52] R. Drullman, J. M. Festen, and R. Plomp. Effect of temporal envelope smearing on speech reception. *Journal of the Acoustical Society of America*, 95(2):1053–1064, 1994.

[53] T. En-Najjary, O. Rosec, and T. Chonavel. A new method for pitch prediction from spectral envelope and its application in voice conversion. In *INTERSPEECH*, 2003.

[54] T. En-Najjary, O. Rosec, and T. Chonavel. A voice conversion method based on joint pitch and spectral envelope transformation. In *INTERSPEECH*, 2004.

[55] D. Erro and A. Moreno. Frame alignment method for cross-lingual voice conversion. In *INTERSPEECH*, 2007.

[56] D. Erro and A. Moreno. Weighted frequency warping for voice conversion. In *INTERSPEECH*, 2007.

[57] D. Erro, A. Moreno, and A. Bonafonte. Voice conversion based on weighted frequency warping. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(5):922–931, 2010.

[58] F. Fang, J. Yamagishi, I. Echizen, and J. Lorenzo-Trueba. High-quality nonparallel voice conversion based on cycle-consistent adversarial network. In *ICASSP*, 2018.

[59] S. Ferguson. Talker differences in clear and conversational speech: Vowel intelligibility for normal-hearing listeners. *Journal of the Acoustical Society of America*, 116:2365–2373, 2004.

[60] S. H. Ferguson and D. Kewley-Port. Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners. *Journal of the Acoustical Society of America*, 112:259–271, 2002.

[61] S. H. Ferguson and D. Kewley-Port. Talker differences in clear and conversational speech: Acoustic characteristics of vowels. *Journal of Speech, Language, and Hearing Research*, 50:1241–1255, 2007.

[62] J. L. Flanagan and R. M. Golden. Phase vocoder. In *Bell System Technical Journal*, volume 45, page 1493–1509, 1966.

[63] N. R. French and J. C. Steinberg. Factors governing the intelligibility of speech sounds. *Journal of the Acoustical Society of America*, 50:1241–1255, 1947.

[64] T. Gay. Effect of speaking rate on diphthong formant movements. *JASA*, 44(6):1570–1573, December 1968.

[65] T. Gay. Effect of speaking rate on vowel formant movements. *JASA*, 63(1):223–230, January 1978.

[66] E. Godoy, M. Koutsogiannaki, and Y. Stylianou. Approaching speech intelligibility enhancement with inspiration from lombard and clear speaking styles. *Computer Speech & Lang.*, 28(2):629–647, 2014.

[67] E. Godoy, O. Rosec, and T. Chonavel. Alleviating the one-to-many mapping problem in voice conversion with context-dependent modelling. In *INTERSPEECH*, 2009.

[68] E. Godoy, O. Rosec, and T. Chonavel. Spectral envelope transformation using dfw and amplitude scaling for voice conversion with parallel or nonparallel corpora. In *INTERSPEECH*, 2011.

[69] E. Godoy, O. Rosec, and T. Chonavel. Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(4):1313–1323, 2012.

[70] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014.

[71] S. Gordon-Salant. Recognition of natural and time/intensity altered cvs by young and elderly subjects with normal hearing. *ournal of the Acoustical Society of America*, 82(6):1599–1607, 1986.

[72] S. Gordon-Salant. Effects of acoustic modification on consonant recognition by elderly hearing-impaired subjects. *Journal of the Acoustical Society of America*, 81(4):1199–1202, 1987.

[73] S. Gordon-Salant and P. J. Fitzgibbons. Selected cognitive factors and speech recognition performance among young and elderly listeners. *Journal of Speech and Hearing Research*, 40:423–431, 1997.

[74] S. Gordon-Salant and P. J. Fitzgibbons. Sources of age-related recognition difficulty for time-compressed speech. *Journal of Speech, Language, and Hearing Research*, 44:709–719, 2001.

[75] Z. Hanzlicek and J. Matousek. F0 transformation within the voice conversion framework. In *INTERSPEECH*, 2007.

[76] M. Hashimoto and N. Higuchi. Spectral mapping method for voice conversion using speaker selection and vector field smoothing. In *EUROSPEECH*, 1995.

[77] V. Hazan and D. Markham. Acoustic-phonetic correlates of talker intelligibility for adults and children. *Journal of the American Academy of Audiology*, 116(5):3108–3118, 2004.

[78] V. Hazan and A. Simpson. The effect of cue-enhancement on the intelligibility of nonsense word and sentence materials presented in noise. *Speech Communication*, 24:211–226, 1998.

[79] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *Proceedings of the ICCV*, pages 1026–1034, 2015.

[80] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *Computer Vision and Pattern Recognition (CVPR)*, 2016.

[81] E. Helander, J. Schwarz, J. Nurminen, H. Silen, and M. Gabbouj. On the impact of alignment on voice conversion performance. In *INTERSPEECH*, 2008.

[82] E. Helander, H. Silén, T. Virtanen, and M. Gabbouj. Voice conversion using dynamic kernel partial least squares regression. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(3):806–817, 2012.

[83] E. E. Helander and J. Nurminen. On the importance of pure prosody in the perception of speaker identity. In *INTERSPEECH*, 2007.

[84] K. S. Helfer. Auditory and auditory-visual recognition of clear and conversational speech by older adults. *Journal of the American Academy of Audiology*, 9:234–242, 1998.

[85] J. M. Hillenbrand and M. J. Clark. Some effects of duration on vowel recognition. *Journal of the Acoustical Society of America*, 108(6):3509–3523, 2000.

[86] J. M. Hillenbrand and T. M. Nearey. Identification of resynthesized /hvd/ utterances: Effects of formant contour. *Journal of the Acoustical Society of America*, 406(6):3509–3523, 1999.

[87] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang. Voice conversion from non-parallel corpora using variational auto-encoder. In *APSIPA*, 2016.

[88] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang. Voice conversion from non-parallel corpora using variational auto-encoder. *Proceeding APSIPA ASC*, 2016.

[89] W.-N. Hsu, Y. Zhang, and J. Glass. Learning latent representations for speech generation and transformation. *Proceedings of INTERSPEECH*, 2017.

[90] S. Imai. Cepstral analysis synthesis on the mel frequency scale. In *ICASSP*, 1983.

[91] S. Imai, K. Sumita, and C. Furuichi. Mel log spectrum approximation (mlsa) filter for speech synthesis. *Electronics and Communications in Japan (Part I: Communica- tions)*, 66(2):10–18, 1983.

[92] U. Isik, R. Giri, N. Phansalkar, J.-M. Valin, K. Helwani, and A. Krishnaswamy. PoCoNet: Better Speech Enhancement with Frequency-Positional Embeddings, Semi-Supervised Conversational Data, and Biased Loss. In *Proc. Interspeech 2020*, pages 2487–2491, 2020.

[93] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *Proceedings of CVPR*, pages 5967–5976, 2016.

[94] A. Kain. *High resolution voice transformation*. PhD thesis, Oregon Health and Science University, 2001.

[95] A. Kain, A. Amano-Kusumoto, and J.-P. Hosom. Hybridizing conversational and clear speech to determine the degree of contribution of acoustic features to intelligibility. *Journal of the Acoustical Society of America*, 124(4):2308–2319, 2008.

[96] A. Kain, J.-P. Hosom, X. Niu, J. P. van Santen, M. Fried-Oken, and J. Staehely. Improving the intelligibility of dysarthric speech. *Speech Communication*, 49:743–759, 2007.

[97] A. Kain and M. W. Macon. Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction. In *ICSLP*, 1998.

[98] A. Kain and M. W. Macon. Spectral voice conversion for text-to-speech synthesis. In *ICASSP*, 1998.

[99] H. Kameoka, W.-C. Huang, K. Tanaka, T. Kaneko, N. Hojo, and T. Toda. Many-to-many voice transformer network. In *arXiv:2005.08445*, 2020.

[100] H. Kameoka, K. Tanaka, T. Kaneko, and N. Hojo. Convs2s-vc: Fully convolutional sequence-to-sequence voice conversion. In *arXiv preprint arXiv:1811.01609*, 2018.

[101] H. Kameoka, K. Tanaka, D. Kwasny, T. Kaneko, and N. Hojo. Convs2s-vc: Fully convolutional sequence-to-sequence voice conversion. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2020.

[102] T. Kaneko and H. Kameoka. Parallel-data-free voice conversion using cycle-consistent adversarial networks. In *EUSIPCO*, 1017.

[103] T. Kaneko, H. Kameoka, K. Hiramatsu, and K. Kashino. Sequence-to-sequence voice conversion with similarity metric learned using generative adversarial networks. In *INTERSPEECH*, 2017.

[104] T. Kaneko, S. Takaki, H. Kameoka, and J. Yamagishi. Generative adversarial network-based postfilter for stft spectrograms. *Proceedings of INTERSPEECH*, 2017.

[105] H. Kato, M. Tsuzaki, and Y. Sagisaka. Effects of phoneme class and duration on the acceptability of temporal modifications in speech. *Journal of the Acoustical Society of America*, 111(1):387–400, January 2002.

[106] H. Kawahara, I. Masuda-Katsuse, and A. D. Cheveigne. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sound. *Speech communication*, 27(3):187–207, 1999.

[107] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno. Tandem-straight: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation. *ICASSP*, 2008.

[108] R. A. Kazi, V. M. Prasad, J. Kanagalingam, C. M. Nutting, P. Clarke, P. Rhys-Evans, and K. J. Harrington. Assessment of the formant frequencies in normal and laryngectomized indi- viduals using linear predictive coding. *Journal of Voice*, 21(6):661–668, 2007.

[109] K. Kazuhiro and T. Toda. Electrolaryngeal speech enhancement with statistical voice conversion based on cldnn. In *EUSIPCO*, pages 2115–2119, 2018.

[110] N. Keigo, T. Toda, H. Saruwatari, and K. Shikano. Electrolaryngeal speech enhancement based on statistical voice conversion. In *INTERSPEECH*, pages 1431–1434, 2009.

[111] N. Keigo, T. Toda, H. Saruwatari, and K. Shikano. Speaking-aid systems using gmm-based voice conversion for electrolaryngeal speech. *Speech Communication*, 54(1):134–146, 2012.

[112] Z. A. Khan, P. Green, S. Creer, and S. Cunningham. Reconstructing the voice of an individual following laryngectomy. *Augmentative and Alternative Communication*, 27(1):61–661, 2011.

[113] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou. An algorithm that improves speech intelligibility in noise for normal hearing listeners. *Journal of the Acoustical Society of America*, 126(3):1486–1494, 2008.

[114] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *Proceedings of 3rd ICLR*, 2015.

[115] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *Proceedings of ICLR*, 2014.

[116] B. Kollmeier and M. Wesselkam. Development and evaluation of a sentence test for objective and subjective speech intelligibility assessment. *Journal of the Acoustical Society of America*, 102(4):1085–1099, 1997.

[117] M. Koutsogiannaki, P. N. Petkov, and Y. Stylianou. Simple and artefact-free spectral modifications for enhancing the intelligibility of casual speech. *Proceedings of ICASSP*, pages 4648–4652, 2014.

[118] M. C. Koutsogiannaki. *Intelligibility Enhancement of Casual Speech based on Clear Speech Properties*. PhD thesis, University of Crete, 2015.

[119] J. Krause. *Properties of naturally produced clear speech at normal rates and implications for intelligibility enhancement*. PhD thesis, MIT, Cambridge, 2001.

[120] J. C. Krause and L. D. Braida. Investigation alternative forms of clear speech: The effects of speaking rate and speaking mode on intelligibility. *Journal of the Acoustical Society of America*, 112(5):2165–2172, 2002.

[121] J. C. Krause and L. D. Braida. Acoustic properties of naturally produced clear speech at normal speaking rates. *Journal of the Acoustical Society of America*, 15:362–378., 2004.

[122] J. C. Krause and L. D. Braida. Evaluationg the role of spectral and envelope characteristics in the intelligibility advantage of clear speech. *ournal of Acoustical Society of America*, 125(5):3346–3357, 2009.

[123] A. Kupryjanow and A. Czyżewski. A non-uniform real-time speech time-scale stretching method. In *Proceedings of the International Conference on Signal Processing and Multimedia Applications*, pages 1–7, 2011.

[124] A. Kusumoto, T. Arai, K. Kinoshita, N. HOdoshima, and N. Vaughan. Modulation enhancement of speech by a pre-processing algorithm for improving intelligibility in reverberant environments. *Speech Communication*, 45:101–113, 2005.

[125] H. Kuwabara. Acoustic and perceptual properties of phonemes in continuous speech as a function of speaking rate. In *EUROSPEECH*, 1997.

[126] M. S. E. Langarani, J. P. H. van Santen, S. H. Mohammadi, and A. Kain. Data-driven foot-based intonation generator for text-to-speech synthesis. In *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, pages 1596–1600. ISCA, 2015.

[127] J. Laroche and M. Dolson. Phase-vocoder: about this phasiness business. In *Proceedings of 1997 Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 4 pp.–, 1997.

[128] J. Laroche and M. Dolson. Improved phase vocoder time-scale modification of audio. *IEEE Transactions on Speech and Audio Processing*, 7(3):323–332, 1999.

[129] K.-S. Lee. Statistical approach for voice personality transformation. *EEE Transactions on Audio, Speech, and Language Processing*, 15(2):641–651, 2007.

[130] J. S. Lienard and M. G. Dibenedetto. Effect of vocal effort on spectral properties of vowels. *Journal of the Acoustical Society of America*, 106(1):411–422, 1999.

[131] B. E. Lindblom. Spectrographic study of vowel reduction. *JASA*, 35(11):1773–1781, November 1963.

[132] H. Liu and M. L. Ng. Electrolarynx in voice rehabilitation. *Auris Nasus Larynx*, 34(3):327–332, 2007.

[133] S. Liu, E. D. Rio, A. R. Bradlow, and F. G. Zeng. Clear speech perception in acoustic and electric hearing. *Journal of the Acoustical Society of America*, 116(4):2374–2383, 2004.

[134] S. Liu and F. Zeng. Temporal properties in clear speech perception. *Journal of the Acoustical Society of America*, 120(1):424–432, 2006.

[135] A. Loscos and J. Bonada. Esophageal voice enhancement by modeling radiated pulses in frequency domain. In *121st Convention of the Audio Engineering Society*, 2006.

[136] B. Makki, S. Seyedsalehi, N. Sadati, and M. N. Hosseini. Voice conversion using nonlinear principal component analysis. In *CIISP*, 2007.

[137] K. Maniwa, A. Jongman, and T. Wade. Perception of clear fricatives by normal-hearing and simulated hearing-impaired listeners. *Journal of the Acoustical Society of America*, 123(2):1114–1125, 2008.

[138] R. Martin. Speech enhancement based on minimum mean-square error estimation and supergaussian priors. *IEEE Transactions on Speech and Audio Processing*, 13(5):845–856, September 2005.

[139] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger. Montreal forced aligner: trainable text-speech alignment using kaldi. In *Proceedings of the 18th Conference of the International Speech Communication Association*, 2017.

[140] M. J. McAuliffe, A. R. Fletcher, S. E. Kerr, G. A. O'Beirne, and T. Anderson. Effect of dysarthria type, speaking condition, and listener age on speech intelligibility. *American Journal of Speech-Language Pathology*, 2017.

[141] McLoughlin, I. Vince, J. Li, and Y. Song. Reconstruction of continuous voiced speech from whispers. In *INTERSPEECH*, 2013.

[142] J. Mertl, E. Žáčková, and B. Řepová. Quality of life of patients after total laryngectomy: the struggle against stigmatization and social exclusion using speech synthesis. *Disability and Rehabilitation: Assistive Technology*, 13(4):342–352, 2018.

[143] D. Michelsanti. Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification. In *INTERSPEECH*, 2017.

[144] H. Mizuno and M. Abe. Voice conversion algorithm based on piecewise linear conversion rules of formant frequency and spectrum til. *Speech Communication*, 16(2):153–164, 1995.

[145] S. Mohammadi, A. Kain, and J. van Santen. Making conversational vowels more clear. *Proceedings of INTERSPEECH*, 2012.

[146] S. H. Mohammadi. Reducing one-to-many problem in voice conversion by equalizing the formant locations using dynamic frequency warping. In *ArXiv e-prints*, 2015.

[147] S. H. Mohammadi and A. Kain. Semi-supervised training of a voice conversion mapping function using a joint-autoencoder. In *INTERSPEECH*, 2015.

[148] S. H. Mohammadi and A. Kain. An overview of voice conversion systems. *Speech Communication*, 2017.

[149] S. J. Moon and B. Lindblom. Interaction between duration, context, and speaking style in english stressed vowels. *Journal of the Acoustical Society of America*, 96(1):40–55, 1994.

[150] M. Morise. D4c, a band-aperiodicity estimator for high-quality speech synthesis. *Speech Communication*, 84(57-65), 2016.

[151] M. Morise. D4c, a band-aperiodicity estimator for high-quality speech synthesis. *Speech Communication*, 84:57–65, 2016.

[152] M. Morise, F. Yokomori, and K. Ozawa. World: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE transactions on information and systems*, E99-D(77):1877–1884, 2016.

[153] M. Morise, F. Yokomori, and K. Ozawa. World: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE transactions on information and systems*, E99-D(7):1877–1884, 2016.

[154] E. Morley, E. Klabbers, J. P. van Santen, A. Kain, and S. H. Mohammadi. Synthetic f0 can effectively convey speaker id in delexicalized speech. In *INTERSPEECH*, 2012.

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

[156] A. Mouchtaris, Y. Agiomyrgiannakis, and Y. Stylianou. Conditional vector quantization for voice conversion. In *ICASSP*, 2007.

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

[158] A. R. N, A. Rao MV, G. N. Meenakshi, and P. K. Ghosh. Reconstructing neutral speech from tracheoesophageal speech. In *INTERSPEECH*, 2018.

[159] T. Nakashika, T. Takiguchi, and Y. Ariki. High-order sequence modeling using speaker-dependent recurrent temporal restricted boltzmann machines for voice conversion. In *INTERSPEECH*, 2014.

[160] T. Nakashika, T. Takiguchi, and Y. Ariki. Voice conversion using rnn pre-trained by recurrent temporal restricted boltzmann machines. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(3):580–587, 2015.

[161] M. Narendranath, H. A. Murthy, S. Rajendran, and B. Yegnanarayana. Transformation of formants for voice conversion using artificial neural networks. *Speech communication*, 16(2):207–216, 1995.

[162] V. K. Narne and C. S. Vanaja. Effect of envelop enhancement on speech perception in individuals with auditory neuropathy. *Ear and Hearing*, 29(1):45–53, 2008.

[163] R. Niederjohn and J. Grotelueschen. The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression. *IEEE Trans. Audio Speech Lang. Process*, 24:277–282, 1976.

[164] G. V. Nuffelen, M. D. Bodt, J. Vanderwegen, P. V. de Heyning, and F. Wuyts. Effect of rate control on speech production and intelligibility in dysarthria. *Folia Phoniatrica et Logopaedica*, 62(3):110–119, 2010.

[165] G. V. Nuffelen, M. D. Bodt, F. Wuyts, and P. V. de Heyning. The effect of rate control on speech rate and intelligibility of dysarthric speech. *Folia Phoniatrica et Logopaedica*, 61(2):69–75, 2009.

[166] J. Nurminen, J. Tian, and V. Popa. Voicing level control with application in voice conversion. In *INTERSPEECH*, 2007.

[167] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano. Maximum likelihood voice conversion based on gmm with straight mixed excitation. In *INTERSPEECH*, 2006.

[168] D. O'Shaughnessy. Speech communications: Human and machine, 2nd ed. *IEEE Press, New York*, 2000.

[169] K. K. Paliwal. Interpolation properties of linear prediction parametric representations. In *EUROSPEECH*, 1995.

[170] pawangfg. Variational autoencoders. https://www.geeksforgeeks.org/variational-autoencoders/, 2020.

[171] K. Payton, R. Uchanski, and L. Braida. Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing. *Journal of the Acoustical Society of America*, 95(3):1581–1592, 1994.

[172] M. Picheny, N. Durlach, and L. Braida. Speaking clearly for the hard of hearing i: Intelligibility differences between clear and conversational speech. *Journal of Speech and Hearing Research*, 28:96–103, 1985.

[173] M. A. Picheny, N. I. Durlach, and L. D. Braida. Speaking clearly for the hard of hearing ii: Acoustic characteristics of clear and conversational speech. *Journal of Speech and Hearing Research*, 29:434–446, 1986.

[174] M. K. Pichora-Fuller, B. A. Schneider, and M. Daneman. How young and old adults listen to and remember speech in noise. *Journal of the Acoustical Society of America*, 97(1):593–608, 1995.

[175] M. Portnoff. Implementation of the digital phase vocoder using the fast fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(3):243–248, 1976.

[176] R. Prenger, R. Valle, and B. Catanzaro. Waveglow: A flow-based generative network for speech synthesis. In *ICASSP*, 2019.

[177] T. Quatieri and R. McAulay. Peak-to-rms reduction of speech based on a sinusoidal model. *IEEE Trans. on Audio and Electroacoustics*, 39:273–288, 1991.

[178] S. V. Rao, N. J. Shah, and H. A. Patil. Novel pre-processing using outlier removal in voice conversion. In *SSW*, 2016.

[179] S. H. Reza, I. V. McLoughlin, and F. Ahmadi. Reconstruction of normal sounding speech for laryngectomy patients through a modified celp codec. *IEEE Transactions on Biomedical Engineering*, 57(10):2448–2458, 2010.

[180] B. Sauert and P. Vary. Near end listening enhancement: speech intelligibility improvement in noisy environments. pages 493–496, 2006.

[181] D. J. Schum. Intelligibility of clear and conversational speech of young and elderly talkers. *Journal of the American Academy of Audiology*, 7:212–218, 1996.

[182] H. R. Sharifzadeh, I. V. McLoughlin, and F. Ahmadi. Regeneration of speech in voice-loss patients. In *13th International Conference on Biomedical Engineering*, pages 1065–1068, 2009.

[183] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *ICASSP*, pages 4779–4783, 2018.

[184] K. Shikano, S. Nakamura, and M. Abe. Speaker adaptation and voice conversion by codebook mapping. *IEEE International Sympoisum on Circuits and Systems*, pages 594–597, 1991.

[185] Z. Shuang, F. Meng, and Y. Qin. Voice conversion by combining frequency warping with unit selection. In *ICASSP*, 2008.

[186] Z.-W. Shuang, Z.-X. Wang, Z.-H. Ling, and R.-H. Wang. A novel voice conversion system based on codebook mapping with phoneme-tied weighting. In *ICSLP*, 2004.

[187] M. D. Skowronski and J. G. Harris. Applied principles of clear and lombard speech for automated intelligibility enhancement in noisy environments. *Speech Communication*, 48:549–558, 2006.

[188] R. Smits, L. T. Bosch, and R. Collier. Eveluation of various sets of acoustic cues for the perception of prevocalic stop consonants. i. perception experiment. *Journal of the Acoustical Society of America*, 100(6):3852–3864, 1996.

[189] M. Sommers and J. Barcroft. Stimulus variability and the phonetic relevance hypothesis: Effects of variability in speaking style, fundamental frequency, and speaking rate on spoken word identification. *Journal of the Acoustical Society of America*, 119(4):2406–2416, 2006.

[190] M. S. Sommers, L. C. Nygaard, and D. B. Pisoni. Stimulus variability and spoken word recognition. i. effects of variability in speaking rate and overall amplitude. *Journal of the Acoustical Society of America*, 96(3):1314–1324, 94.

[191] E. L. Stine and A. Wingfield. Process and strategy in memory for speech among younger and older adults. *Psychology and Aging*, 2(3):272–279, 1987.

[192] K. L. Stipancic, K. Tjaden, and G. Wilding. Comparison of intelligibility measures for adults with parkinson's disease, adults with multiple sclerosis, and healthy controls. *Journal of Speech, Language, and Hearing Research*, 59(2), 2016.

[193] I. Stylianou. *Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification.* PhD thesis, Ecole Nationale Supérieure des Télécommunications, 1996.

[194] D. Sundermann, H. Ney, and H. Hoge. Vtln-based cross-language voice conversion. In *ASRU*, 2003.

[195] D. Sündermann, A. Bonafonte, H. Ney, and H. Hög. A study on residual prediction techniques for voice conversion. In *ICASSP*, 2005.

[196] S. Takamichi, T. Toda, A. W. Black, and S. Nakamura. Modulation spectrum-based post-filter for gmm-based voice conversion. In *APSIPA*, 2014.

[197] T. Takeuchi and Y. Tatekura. Speech intelligibility enhancement in noisy enviroment via voice conversion with glimpse proportion measure. *Proceeding of APSIPA ASC*, pages 1713–1717, 2018.

[198] M. Tamura, M. Morita, T. Kagoshima, and M. Akamine. One sentence voice adaptation using gmm-based frequency-warping and shift with a sub-band basis spectrum model. In *ICASSP*, 2011.

[199] K. Tanaka, H. Kameoka, T. Kaneko, and N. Hojo. Atts2s-vc: Sequence-to-sequence voice conversion with attention and context preservation mechanisms. 2018.

[200] Y. Tang and M. Cooke. Learning static spectral weightings for speech intelligibility enhancement in noise. *Computer Speech & Language*, 49:1–16, 2018.

[201] X. Tian, S. W. Lee, Z. Wu, E. S. Chng, H. Li, X. Tian, S. W. Lee, Z. Wu, E. S. Chng, and H. Li. An exemplar-based approach to frequency warping for voice conversion. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 2017.

[202] X. Tian, Z. Wu, S. Lee, and E. S. Chng. Correlation-based frequency warping for voice conversion. In *ISCSLP*, 2014.

[203] K. Tjaden, A. Kain, and J. Lam. Hybridizing conversational and clear speech to investigate the source of increased intelligibility in speakers with parkinson's disease. *Journal of Speech, language, and hearing research*, 57:1191–1205, 2014.

[204] K. Tjaden, J. Lam, and G. E. Wilding. Vowel acoustics in parkinson's disease and multiple sclerosis: comparison of clear, loud, and slow speaking conditions. *Journal of Speech, language, and hearing research*, 56:1485–1502, 2013.

[205] K. Tjaden, J. E. Sussman, and G. E. Wilding. Impact of clear, loud, and slow speech on scaled intelligiblity and speech serverity in parkinson's disease and multiple sclerosis. *Journal of Speech, language, and hearing research*, 57:779–792, 2014.

[206] K. Tjaden and G. Wilding. Rate and loudness manipulations in dysarthria: acoustic and perceptual findings. *Journal of Speech, Language, and Hearing Research*, 47:766–783, 2004.

[207] T. Toda, A. W. Black, and K. Tokuda. Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter. In *ICASSP*, 2005.

[208] T. Toda, A. W. Black, and K. Tokuda. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(8):2222–2235, 2007.

[209] T. Toda, L.-H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, and J. Yamagishi. The voice conversion challenge 2016. *Proceedings of INTERSPEECH*, 2016.

[210] E. Turajlic, D. Rentzos, S. Vaseghi, and C.-H. Ho. Evaluation of methods for parameteric formant transformation in voice conversion. In *ICASSP*, 2003.

[211] O. Turk and L. M. Arslan. Robust processing techniques for voice conversion. *Computer Speech and Language*, 20(4):441–467, 2006.

[212] C. W. Turner, S. J. Smith, P. L. Aldridge, and S. L. Stewart. Formant transition duration and speech recognition in normal and hearing-impaired listeners. *Journal of the Acoustical Society of America*, 101(5):2822–2825, 1997.

[213] R. Uchanski, S. Choi, L. Braida, C. Reed, and N. Durlach. Speaking clearly for the hard of hearing iv: further studies of the role of speaking rate. *J. of Speech and Hearing, 39:494–509.*, 1996.

[214] R. M. Uchanski. Clear speech. *The handbook of speech perception*, pages 207–235, 2005.

[215] A. Uriz, P. Aguero, J. Tulli, E. Gonzalez, and A. Bonafonte. Voice conversion using frame selection and warping functions. In *RPIC*, 2009.

[216] H. Valbret, E. Moulines, and J. Tubach. Voice transformation using psola technique. *Speech Communication*, 11(2):175–187, 1992. Eurospeech '91.

[217] H. Valbret, E. Moulines, and J.-P. Tubach. Voice transformation using psola technique. In *ICASSP*, 1992.

[218] J.-M. Valin, U. Isik, N. Phansalkar, R. Giri, K. Helwani, and A. Krishnaswamy. A perceptually-motivated approach for low-complexity, real-time enhancement of fullband speech. In *Proc. Interspeech 2020*, pages 2482–2486, 2020.

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

[220] J. van Santen, A. Kain, E. Klabbers, and T. Mishra. Synthesis of prosody using multi-level unit sequences. *Speech Communication*, 46(3):365–375, 2005. Quantitative Prosody Modelling for Natural Speech Description and Generation.

[221] W. Verhelst and M. Roelands. An overlap-add technique based on waveform similarity (wsola) for high quality time-scale modification of speech. In *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 554–557 vol.2, 1993.

[222] D. Vincent, O. Rosec, and T. Chonavel. A new method for speech synthesis and transformation based on an arx-lf source-filter decomposition and hnm modeling. In *ICASSP*, 2007.

[223] S. William. End-to-end deep neural network for automatic speech recognition. 2015.

[224] D. Williamson, Y. Wang, and D. Wang. Estimating nonnegative matrix model activations with deep neural networks to increase perceptual speech quality. *Journal of the Acoustical Society of America*, 138(3):1299–1407, 2015.

[225] A. Wingfield, J. S. Aberdeen, and A. L. Stine. Word onset gating and linguistic context in spoken word recognition by young and elderly adults. *Journal of Gerontology*, 46(3):127–129, 1991.

[226] C.-H. Wu, C.-C. Hsia, T.-H. Liu, and J.-F. Wang. Voice conversion using duration-embedded bi-hmms for expressive speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1109–1116, 2006.

[227] F.-L. Xie, Y. Qian, Y. Fan, F. K. Soong, and H. Li. Sequence error (se) minimization training of neural network for voice conversion. In *INTERSPEECH*, 2014.

[228] C. Zhu, R. H. Byrd, and J. Nocedal. L-bfgs-b: Algorithm 778: L-bfgs-b, fortran routines for large scale bound constrained optimization. *ACM Transactions on Mathematical Software*, 23(4):550–560, 1997.

[229] P. Zolfaghari and T. Robinson. A formant vocoder based on mixtures of gaussians. In *ICASSP*, 1997.

[230] T.-C. Zorila, D. Erro, and I. Hernaez. Improving the quality of standard gmm-based voice conversion systems by considering physically motivated linear transformations. In *Advances in Speech and Language Technologies for Iberian Languages*, 2012.

Biographical Note

Tuan Anh Dinh was born on January 12, 1989 in Hai Phong, Vietnam. He received a Bachelor of Science with a major in Software Engineering from Hanoi University of Science and Technology in July 2012. During the following two years, he was employed as a researcher in Vietnam Academy of Science and Technology. He received a Masters of Science with a major in Information Science from the Japan Advanced Institute of Science and Technology in March 2016. In the Fall of 2016, he joined the Ph.D. program at Oregon Health & Science University. His research work includes digital signal processing, machine learning, intelligibility improvement, and speech analysis/synthesis.