5-2021

Development of a Two-Level Warping Algorithm and Its Application to Speech Signal Processing

Al-Waled H. Al-Dulaimi

Utah State University

Recommended Citation

Al-Dulaimi, Al-Waled H., "Development of a Two-Level Warping Algorithm and Its Application to Speech Signal Processing" (2021). *All Graduate Theses and Dissertations*. 8030.
https://digitalcommons.usu.edu/etd/8030

DEVELOPMENT OF A TWO-LEVEL WARPING ALGORITHM AND ITS APPLICATION TO SPEECH SIGNAL PROCESSING

by

Al-Waled H. Al-Dulaimi

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

DOCTOR OF PHILOSOPHY

in

Electrical Engineering

Approved:

Todd K. Moon, Ph.D. Major Professor	Jacob H. Gunther, Ph.D. Committee Member
Scott E. Budge, Ph.D. Committee Member	Stephanie A. Borrie, Ph.D. Committee Member
Kevin R. Moon, Ph.D. Committee Member	D. Richard Cutler, Ph.D. Interim Vice Provost of Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah

2021

ABSTRACT

Development of a Two-Level Warping Algorithm and Its Application to Speech Signal Processing

by

Al-Waled H. Al-Dulaimi, Doctor of Philosophy
Utah State University, 2021

Major Professor: Todd K. Moon, Ph.D.
Department: Electrical and Computer Engineering

The objective of this dissertation is twofold. First, we develop a new two-level dynamic warping algorithm from regular dynamic time. An outer-level warping process, which does temporal warping alignment (dynamic time warping), invokes an inner-level warping process, which achieves spectral warping alignment (dynamic frequency warping).

The second direction of this dissertation is to apply this algorithm to two different kinds of speech processing applications. In one application, the two-level dynamic warping algorithm used in a computer-based tool to help train listeners to learn to imitate dysarthric speech. The tool could eventually be used to provide the learner with feedback regarding their speech imitation accuracy during training. The study reported in this work is to see whether the processing can distinguish between habitual and imitation attempts. Another application is to achieve voice transformation, for example, transforming from a male speaker to a female speaker. For this problem, the mapping function produced by inner warping (dynamic time warping) is used to move spectral information from a source speaker to a target speaker. This process of transformation involves only spectral magnitudes, and has been found to introduce significant deleterious signal processing artifacts with the transformed speech. It has been found that reconstruction of phase information significantly improves the quality of the transformed speech. Information obtained by dynamic frequency warping is used to train an artificial neural network to produce spectral warping output information based on spectral input data to assist in voice transformation. Objective evaluation measure of spectral features and warping paths was applied to evaluate the quality of the transformed speech.

(128 pages)

PUBLIC ABSTRACT

Development of a Two-Level Warping Algorithm and Its Application to Speech Signal Processing

Al-Waled H. Al-Dulaimi

In many different fields there are signals that need to be aligned or "warped" in order to measure the similarity between them. When two time signals are compared, or when a pattern is sought in a larger stream of data, it may be necessary to warp one of the signals in a nonlinear way by compressing or stretching it to fit the other. Simple point-to-point comparison may give inadequate results, because one part of the signal might be comparing different relative parts of the other signal/pattern. Such cases need some sort of alignment to do the comparison. Dynamic Time Warping (DTW) is a powerful and widely used technique of time series analysis which performs such nonlinear warping in temporal domain. The work in this dissertation develops in two directions. The first direction is to extend the this dynamic time warping to produce a two-level dynamic warping algorithm, with warping in both temporal and spectral domains. While there have been hundreds of research efforts in the last two decades that have applied and used the one-dimensional warping process idea between time series, extending DTW method to two or more dimensions poses a more involved problem. The two-dimensional dynamic warping algorithm developed here for a variety of speech signal processing is ideally suited.

The second direction is focused on two speech signal applications. The First application is the evaluation of dysarthric speech. Dysarthria is a neurological motor speech disorder, which characterized by spectral and temporal degradation in speech production. Dysarthria management has focused primarily teaching patients to improve their ability to produce speech or strategies to compensate for their deficits. However, many individuals with dysarthria are not well-suited for traditional speaker-oriented intervention. Recent studies have shown that speech intelligibility can be improved by training the listener to better understand the degraded speech signal. A computer-based training tool was developed using a two-level dynamic warping algorithm to eventually be incorporated into a program that trains listeners to learn to imitate dysarthric speech by providing subjects with feedback about the accuracy of their imitation attempts during training.

The second application is voice transformation. Voice transformation techniques aims to modify a subject's voice characteristics to make them sound like someone else, for example from a male speaker to female speaker. The approach taken here avoids the need to find acoustic parameters as many voice transformation methods do, and instead deals directly with spectral information. Based on the two-Level DW it is straightforward to map the source speech to target speech when both are available. The resulted spectral warping signal produced as described above introduces significant processing artifacts. Phase reconstruction was applied to the transformed signal to improve the quality of the final sound. Neural networks are trained to perform the voice transformation.

In memory of my dad. I love and miss him dearly

To my lovely wife
Ghufran Al-Juboori

To my Mother

To my kids
Hisham, Rayan, Alya and Lamar

To my siblings
Walaa, Wael, Mouloud, Ayat, Omar, and Binan

ACKNOWLEDGMENTS

There are many people whom I want to mention their names here. Those who have helped me in numerous ways not only on my path through the completion of my Ph.D. but also in the other aspects of my life. First, and foremost, I want to thank my supervisor, Dr. Todd K. Moon, for all the time and effort he invested in me throughout the course of my study. His deep insights and positive manner have always been helpful and encouraging. He has been very patient, great encourager and supporter, and sometimes though as it should be. Next, special thanks go to my committee members, Dr. Gunther, Dr. Budge, Dr. Borrie, and Dr. Moon for their support and help, particularly for their patience in reading my dissertation draft. Using this space, I should specially thank Dr. Gunther for his consistent support and encouragement. Furthermore, I would like to thank and express my gratitude to Dr. Borrie for her support and help with my research as well. Also, special thanks to Dr. Budge. I really enjoyed being a teaching assistant for three of his classes.

I also would like to express my gratitude to the Utah State University, as I enjoyed studying over there and I learned a lot.

In addition, I want to thank Dr. Mohammad Shekaramiz, Dr. Abdurazag Khalat, Dr. Mehedi Hassan, Md Munibun Billah, Sam Whiting, and many more for their friendship and encouragement.

Also, I want to thank Tricia Brandenburg, Diane Buist, Kathy Phippen, Heidi Harper, and Brady Forbush from the ECE department.

I would like to thank my gorgeous wife for having taken care of me in hard times. The constant love and affection from my family is the backbone of any successful endeavor in my life. Without their constant support and encouragement for quality education I would never have achieved the right kind of exposure to fulfill my dream of working in my area of interest.

I would like to take this space to express my gratitude to mom, and siblings for their consistent help, support, and in particular, moral support in my journey including my

academic career and other aspects of my life.

Al-Waled H. Al-Dulaimi

CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

ACRONYMS

1-DCNN	One-dimension convolutional neural network
ANN	Artificial neural network
AS	Amplitude scaling
CNN	Convolutional neural network
DFW	Dynamic frequency warping
DTW	Dynamic time warping
DW	Dynamic warping
FFT	Fast Fourier transform
GLA	Griffin-Lim algorithm
GMM	Gaussian mixture model
LPC	Linear predictive coding
MCD	Mel-cepstral distortion
ML	Machine learning
MSE	Mean square error
NN	Neural network
VT	Voice transformation

CHAPTER 1

INTRODUCTION

In this dissertation, the familiar dynamic time warping (DTW) is extended to produce a two-level dynamic warping algorithm, with mapping in both the temporal and spectral domains. This two-level dynamic warping is applied in two application areas: evaluation of dysarthric speech, and transforming speech from one speaker to another.

Since DTW is the starting point for this research, it is outlined here. More details are provided in Chapter 2. In DTW, two sequences $\mathbf{X}$ and $\mathbf{Y}$,

$$\begin{aligned}\mathbf{X} &= x_1, x_2, \ldots, x_i, \ldots, x_M \\ \mathbf{Y} &= y_1, y_2, \ldots, y_j, \ldots, y_N,\end{aligned} \tag{1.1}$$

are aligned in a nonlinear way. The idea is illustrated in Figure 1.1. There is a warping function

$$C(\mathbf{X}, \mathbf{Y}) = [c(1), c(2), \ldots, c(k), \ldots, c(K)], \tag{1.2}$$

where each $c(i)$ is a pair of indices $(i(k), j(k))$, which represents the samples being matched. K represents the length of the warping function path and may be greater than M or N. For the example in Figure 1.1, the sequence $\mathbf{X}$ have nine samples ($M = 9$) and sequence $\mathbf{Y}$ have seven samples ($N = 7$), with

$$\begin{aligned}\mathbf{X} &= 7, 9, 6, 9, 12, 6, 4, 5, 8 \\ \mathbf{Y} &= 5, 6, 4, 3, 9, 5, 6.\end{aligned} \tag{1.3}$$

And warping function path is

$$C(\mathbf{X}, \mathbf{Y}) = [(1,1), (2,2), (3,3), (3,4), (4,5), (5,5), (6,6), (7,6), (8,7), (9,7)], \tag{1.4}$$

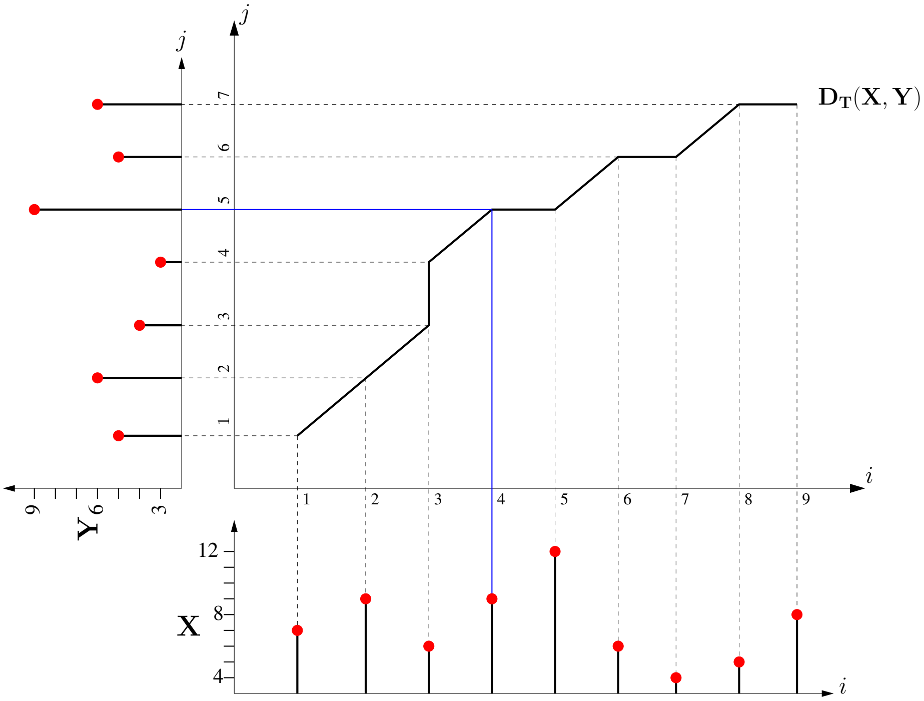


Fig. 1.1: Detail of the Dynamic time (outer) warping process.
The point at (4,5) aligns $x(4)$ with $y(5)$.

where the length of $C(\mathbf{X},\mathbf{Y})$, K, is given by the number of steps in the warping path function. For this example, we have $K = 10$ steps in the warping process. The blue line in Figure 1.1 shows the correspondence assigned between $x(4)$ and $y(5)$.

The DTW process computes the distance between each matched pair of samples of both signals and these distances are used to find the least alignment. Let $d(x_{i(k)}, y_{j(k)})$ denote a distance or cost function between the samples. A typical cost function is the square of the difference between the samples of both time series functions.

$$d(x_{i(k)}, y_{j(k)}) = (x_{i(k)} - y_{j(k)})^2 \triangleq d(c(k)). \tag{1.5}$$

The warping function path is required to minimize the overall distance function

$$D_c(\mathbf{X},\mathbf{Y}) = \sum_{k=1}^{K} d(x_{i(k)}, y_{j(k)}) = \sum_{k=1}^{K} d(c(k)), \tag{1.6}$$

and is usually subjected to some constraints (described in Chapter 2).

The dynamic time warping distance or the total cost of $(i(k), j(k))$ is recursively computed as the distance (or cost) computed in the current point plus the minimum value of the cumulative distances of the adjacent points, according to

$$D_T(x_i, y_j) = d(x_i, y_j) + \min[D_T(x_{i-1}, y_j), D_T(x_i, y_{j-1}), D_T(x_{i-1}, y_{j-1})] \tag{1.7}$$

When $i = M$ and $j = N$, corresponding to the end of the time series functions, the dynamic warping process has found the overall warped metric distance $D_T(x_M, y_N)$ between the sequence $\mathbf{X}$ and $\mathbf{Y}$, which may referred as $D_{DTW}(\mathbf{X}, \mathbf{Y})$. During this warping process, two sequences of indices $\mathbf{i} = (i(1), i(2), \ldots, i(K))$ and $\mathbf{j} = (j(1), j(2), \ldots, j(K))$ are determined, which establish the temporal warping function path. These indices describe the time alignment for both time series sequences, such that the new time aligned sequence $\mathbf{Y}_{TA}(i(k)) = \mathbf{Y}(j(k))$ $(k = 1, ..., K)$ and $\mathbf{X}$ are as similar as possible, so that, roughly, the peaks and valleys of $\mathbf{X}$ align with peaks and valleys, respectively, of $\mathbf{Y}$. The subscript $_{TA}$ refers to Temporal Alignment warping process.

Comparison of two same phrases or statements from two different sources is central to many speech recognition applications [1–6], and the DTW is frequently used in speech applications. In this dissertation, the comparison between the speech sequences from two different sources is done by mapping in both the frequency domain and time domain. Warping in the time domain overcomes temporal variability of the spoken phrase, due to differences in speech rates. The spectral information extracted from the human speech sequences are time-aligned or matched by calculating similarity between segments of these sequences. In addition, warping in the frequency domain compensates for variations in the frequency domain of speech caused by vocal tract difference among different speakers [7]. Dynamic frequency warping (DFW) used in this dissertation is reduces the effects of spectral variations of the speech due to spectral differences [8].

This combination of dynamic time mapping and dynamic frequency mapping, or two-level warping, is a new approach to speech processing. An outer warping process, which temporally aligns blocks of speech (dynamic time warp), invokes an inner warping process, which spectrally aligns based on magnitude spectra (dynamic frequency warp). This two level warping is applied in this dissertation to two applications. In the first application, we applied this algorithm to dysarthric speech in a total used to train care givers to understand dysarthric speech (Dysarthria is a neurological motor speech disorder that commonly results in reduced intelligibility because of dysarthria is a characterized by spectral and temporal degradation [9].) People can learn to understand the speech of someone with dysarthria through perceptual training [10–14]. Vocal imitation of the degraded speech during perceptual training has been shown to elevate this learning [12,14]. A tool was developed using two-level dynamic warping that provides the learner with real-time feedback regarding the accuracy of their imitation attempts during training to enhance and support this learning. This training tool compares the production of a dysarthric speech with the imitation attempt of a healthy speaker. This comparison uses a two-level dynamic warping to account for both spectral and temporal vaiabilty that may ossur due to dysarthria.

The second application of this two level dynamic warping is to voice transformation. Voice transformation, for example, from a male speaker to a female speaker or vice versa, refers to the process of changing the parameters of the human speech or changing voice personality, to achieve the conversion of a speech that uttered by one speaker (the source speaker) to sound as if other speaker (the target speaker) had spoken it [15]. Voice transformation is achieved in this dissertation using a two-level dynamic warping. The mapping function produced by the dynamic frequency warping, DFW, is used to move spectral information from a source speaker to a target speaker. This process of voice transformation involves only spectral magnitudes, and has been found to introduce significant deleterious signal processing artifacts. It has been found that the reconstruction of phase information significantly improves the quality of the transformed speech. The spectral mapping information obtained from two level dynamic warping is used as a training data to train neural network to assist in voice transformation.

The organization of this dissertation is as follows. In chapter 2, a new two-level dynamic warping algorithm is detailed. In chapter 3, some background of the dysarthria is first presented. The two-level dynamic warping algorithm is used in a computer-based training tool to help train speakers learn to imitate dysarthric speech, by providing subjects with feedback about the accuracy of their imitation attempts during training. Clinical testing was conducted within the labs of Utah State University to test this tool.

Chapter 4 begins with presenting a brief background on a conventional voice transformation. Then, voice transformation using two-level dynamic warping algorithm is described by transforming speech form one speaker (source speaker) to another (target speaker). After that, the importance and effects of the phase on the warped speech is discussed. Then, applying a specific phase reconstruction algorithm on the output of the dynamic warping algorithm is achieved. A second phase of this chapter is to train a neural network to assist in voice transformation. The time aligned source spectral feature information are used as an input to the artificial neural network (ANN). The warping function paths ($\mathbf{a}_F$ and $\mathbf{b}_F$) computed by the two-level DW are used as training data. because of the length of the warping paths ($\mathbf{a}_F$ and $\mathbf{b}_F$) may vary form phrase to phrase, an interpolation was applied on the warping paths to produce fixed length of warping paths. After that, An ANN is trained to map a sequence of time aligned source speaker's spectral feature information to the interpolated warping path information. This experiment of estimating the warping paths ($\mathbf{a}_F$ and $\mathbf{b}_F$) using the ANN was contacted in four phases. Phase one was contacted to decide which architecture for the ANN that achieves good quality for the process of voice transformation. Phase two, the selected network from phase one was trained for 600 phrases. In phase three, the spectral information of the 600 phrases was clustered into six clusters and each cluster was trained with the selected network form phase one. Phase Four was contacted using convolutional neural network (CNN).

CHAPTER 2

TWO-LEVEL DYNAMIC WARPING ALGORITHM

This chapter presents a new two-level dynamic warping algorithm, in which an outer-level warping process does temporal alignment (Dynamic Time Warping, DTW), which temporally aligns block of features to compensate for temporal differences, invokes an inner-level warping process (Dynamic Frequency Warping, DFW) to achieve alignment of spectral features to account for spectral shift.

2.1 Background and Previous Work

Dynamic time warping method (DTW) is a dynamic programming technique that has long been used to find an optimal alignment between the sequences of feature vectors from two different sources [16, 17]. The DTW method calculates the distance between each pair of points of the two signals, then uses these distances to calculate the cumulative distance matrix. The least expensive path through this matrix is then found. That least expensive path is called the warping path [18]. This warping path used to synchronize signals, which causes the distance between their synchronised feature vectors to be minimized [1].

DTW was first developed to be used for speech and word recognition in 1970s with sound waves as the source [17]. DTW has since been used for a wide range of applications such as human motion animation, human activity recognition, processor cache analysis, classifying handwritten text, fingerprint indexing, time series classification and data mining [18–24]. DTW is widely used in finance, science, medicine, chemistry, astronomy, robotics, and industry [25]. Also, DTW is widely used as a time series distance measure, across a host of domain applications [26, 27].

In previous work in this field, DTW research focused on speeding up the algorithm and improving the efficiency of processing [1,27,28]. Applying constraints to the DTW process [29], approximation method of the DTW algorithm [30], lower bounding techniques [21] and using spectral feature information instead of time feature information to do the alignment [31] were considered as an examples for speeding up the DTW algorithm and improving the efficiency.

Time series similarity search under the Euclidean metric is heavily I/O bound; however, similarity search under DTW is also very demanding in terms of CPU time. One way to address this problem is to use a fast lower bounding function, based on the warping window, to help prune sequences that could not possibly be the best match. The authors in [21] proposed new lower bounding distance measure technique and they showed that this new technique speeding up DTW algorithm.

[29] introduced a form of modified DTW called Derivative DTW ($DDTW$). With the regular DTW, the cumulative distance matrix contains distances has been calculated between the feature values of the sample points. In this proposed DDTW, the distance measurement elements of the cumulative distance matrix has been calculated not between the feature values of the sample points, but between their associated estimated first order derivatives. Therefore, alignment is done based on the characteristics of the shape (slopes, peaks, valleys) of the the sequences rather than simple values and they showed that this modification to the DTW modification produces a significant alignments between time series.

The authors in [30] proposed Fast DTW, which is able to find an accurate approximation of the optimal warping path between two time series. This approximation method has linear time and space complexity, while the regular DTW has quadratic time and space complexity. This method started by down sampling the time series into a smaller time series representing low resolution. The warping path is calculated for theses lowest resolutions and projected onto an incrementally higher resolution time series. This projected warping path is refined form a lower resolution and projected again to get a higher resolution. The final warping path for the full length (or resolution) time series is calculated by repeating the above process of refining and projecting as much as the whole process is needed.

Because differences in vocal tract lengths among different speakers, there is a lot of variations in frequency domain. These variations in frequency domain is an analogy to the variations in the time domain [30, 32]. For this reason, the authors in [31] have used Mel-frequency cepstral coefficient ($MFCC$) features to reduce the degradation in a gender-independence isolated word recognition. They showed that this way of using frequency features instead of time valued features lead to decrease word error rate in isolated word recognition system.

While there have been hundreds of research efforts in the last two decades that applied and used the one-dimensional warping process idea between time series, extending DTW method to two or more dimensions poses a more involved problem [4]. The input data is no longer a one-dimensional feature vector but rather a two-dimensional or multi-dimensional feature vectors. These two-dimensional features can represent images or any other two-dimensional data. A two-dimensional dynamic warping algorithm could be used for a variety of applications such as text, gesture, facial recognition. [26] proposed the two most commonly used multi-dimensional DTW methods. We provide a description of the work of [26] to demonstrate that it differs from the two-level dynamic warping. The authors refer to their two methods as $\mathrm{DTW_D}$ and $\mathrm{DTW_I}$, where "D" and "I" refers to dependent warping process and independent warping process, respectively. The input for both methods are the multi-dimensional time series functions. A data set $\mathbb{Q} = Q_1, Q_2, \ldots, Q_M$ is a collection of M individual time series ($M \geq 2$), where each time series is a sequence of data points ordered in time to be as a set of real values and the total number of real values is equal to the length of the time series:

$$\begin{aligned} Q_1 &= q_{1,1}, q_{2,1}, \ldots, q_{n,1} \\ Q_2 &= q_{1,2}, q_{2,2}, \ldots, q_{n,2} \\ Q_3 &= q_{1,3}, q_{2,3}, \ldots, q_{n,3} \\ \vdots \quad & \quad \vdots \quad\;\; \vdots \qquad\;\; \vdots \\ Q_M &= q_{1,M}, q_{2,M}, \ldots, q_{n,M} \end{aligned} \tag{2.1}$$

where n represents the length of each time series in the data set $\mathbb{Q}$

Two input sets ($\mathbb{Q}$ and $\mathbb{C}$) were proposed as a two data sets of M-dimensional time series.

The first proposed method ($\mathrm{DTW_I}$) in [26] for doing multi-dimensional DTW ($MDTW$) was based on computing the cumulative distances of all dimensions independently. These distances were measured under regular DTW. For a specific m^{th} dimension, the cumulative distance of the DTW of m^{th} dimension can be computed as the distance found in the current location ($d(q_{i,m}, c_{j,m})$) and the minimum value of the DTW cumulative distances of the adjacent elements:

$$\begin{aligned} DTW(Q_m, C_m) = d(q_{i,m}, c_{j,m}) + \min[&DTW(i-1, j)_m, \\ &DTW(i, j-1)_m, \\ &DTW(i-1, j-1)_m] \\ &i, j = 1, 2, \ldots, n \end{aligned} \tag{2.2}$$

where Q_m and C_m are the m^{th} dimension of the data set $\mathbb{Q}$ and $\mathbb{C}$, respectively, and $d(q_{i,m}, c_{j,m})$ be proposed to be $d(q_{i,m}, c_{j,m}) = (q_{i,m} - c_{j,m})^2$. Therefore, the MDTW can be calculated as

$$DTW_I(\mathbb{Q}, \mathbb{C}) = \sum_{m=1}^{M} DTW(Q_m, C_m) \tag{2.3}$$

In this method, each dimension is considered to be independent on each others to find the DTW between two data set $\mathbb{Q}$ and $\mathbb{C}$. This way of computing DTW gives each dimension the freedom to warp itself independently of the others.

The authors in [26] proposed another way of computing MDTW, DTW_D, by ignoring the dimensions independent and forcing all the dimensions to warp identically for each time index. In this way, they consider MDTW to be calculated in a similar way of regular DTW algorithm used for one-dimensional, (2.2), except that they redefine $d(q_i, c_j)$ as the squared Euclidean distances of M-dimensional of two or multiple data sets of time series of information instead of the single data point used in the more familiar one-dimensional case. They proposed that for the same data set ($\mathbb{Q}$ and $\mathbb{C}$) of multi-dimensional time series, the new equation for computing distance can be defined as:

$$d(q_i, c_j) = \sum_{k=1}^{M} (q_{i,k} - c_{j,k})^2 \tag{2.4}$$

where $q_{i,k}$ is the i^{th} element in the k^{th} dimension of $\mathbb{Q}$ and $c_{j,k}$ is the j^{th} element in the k^{th} dimension of $\mathbb{C}$. Now, MDTW is calculated by redefining (2.3) as:

$$DTW_D(\mathbb{Q}, \mathbb{C}) = \sum_{l=1}^{n} DTW(Q_l, C_l) \tag{2.5}$$

where n represents the length of the time series.

2.2 Two-Level Dynamic Warping

In this work, dynamic time warping (DTW) is extended to a two-level model operating on both temporal and spectral domains. The outer-level, DTW, temporally aligns block of spectral features to compensate for tempo differences (such us different speech rates). Dynamic frequency (spectrum) warping (DFW) or “inner-level warping” is used to perform spectral alignment based on spectral features of blocks of speech data, such as aligning spectral features of male speaker to spectral features of female speaker. In this work, we applied the combination of inner and outer warping, simply referring as ”dynamic warping,” or DW.

2.2.1 Outer Level Dynamic Warping

Our description of DW begins with a review of dynamic time warping. The basic concept was presented in Chapter 1. Here we expand this description. The language used in this dissertation suggests application to speech data, but it could be used for other signals as well. As suggested by the name “outer-level dynamic *time* warping” or “dynamic time warping”, DTW seeks for temporal alignment to align block of feature vectors of speech data to compensate for different speech rates.

For two speakers i $(i = 1, 2)$, let $s_i(t, f)$ refer to the feature information at a timeslice with integer time index t and integer frequency index f. The vector $\mathbf{s}_i(t,:)$, has k elements, which represent the spectral feature vector information at timeslice with integer time index t of speaker i. For speaker i, the speech data signal can be represented as a sequence of T_i block of feature vectors computed from overlapping block of features of speech signal information

$$\mathcal{S}_i = \{\mathbf{s}_i(1,:), \mathbf{s}_i(2,:), \ldots, \mathbf{s}_i(T_i,:)\}, \qquad i = 1, 2. \tag{2.6}$$

Let $d(\mathbf{s}_1(t_1,:), \mathbf{s}_2(t_2,:))$ denote a metric distance between spectral feature vectors $\mathbf{s}_1(t_1,:)$ and $\mathbf{s}_2(t_2,:)$ at time-slices t_1 and t_2. Let $d_T(t_1, t_2)$ represent the minimum cost distance between the sequences $\mathcal{S}_1$ and $\mathcal{S}_2$, up to blocks at times t_1 and t_2. The subscript $_T$ refers to time warping process. The goal of DTW is to find a warping function

$$C(\mathcal{S}_1, \mathcal{S}_2) = [c(1), c(2), \ldots, c(k), \ldots, c(K)], \tag{2.7}$$

where each $c(i)$ is a pair of indices $(a_T(k), b_T(k))$, such that the cost distance on this warping function is minimized, with the following constraints [33]:

- **Monotonicity**:

  $$a_T(k-1) \leq a_T(k) \text{ and } b_T(k-1) \leq b_T(k).$$

  This constraint prevents the alignment path from moving back in time.

- **Continuity**:

  $$a_T(k) - a_T(k-1) \leq 1 \text{ and } b_T(k) - b_T(k-1) \leq 1.$$

  This constraint prevents skipping. This constraint ensures that the alignment process will not neglect or hide any samples.

- **Boundary**:

  $$a_T(1) = b_T(1) = 1,\ a_T(K) = T_1 \text{ and } b_T(K) = T_2.$$

  The warping function must match the endpoints of the two time series functions, the alignment path must start and end at the bottom left and at the top right, respectively (diagonally opposite corner).

- **Warping Window**:

  $$|t_1 - t_2| < w_T,$$

  where $w_T > 0$ is the window length. DTW constrains the alignment path to not to go far from the diagonal.

With these constraints, the path may go several cells horizontally along the function with the x-axis or vertically along the function with the y-axis.

The dynamic time warping distance or the total cost of $(a_T(k), b_T(k))$ is computed as the distance (or cost) computed in the current point plus the cost of the cheapest path to it. For any given node (t_1, t_2) in the path, as shown in Figure 2.1, and by the monotonicity and continuity constraints mentioned above, three backward neighbors nodes $(t_1, t_2 - 1)$, $(t_1 - 1, t_2)$, and $(t_1 - 1, t_2 - 2)$, will be checked. However, these nodes are used to compute the metric distance between the sequence $\mathcal{S}_1$ and the sequence $\mathcal{S}_2$, and to find $d_T(t_1, t_2)$, recursively by

$$d_T(t_1, t_2) = d(\mathbf{s}_1(t_1, :), \mathbf{s}_2(t_2, :)) + \min[d_T(t_1 - 1, t_2), d_T(t_1, t_2 - 1), d_T(t_1 - 1, t_2 - 1)] \tag{2.8}$$

This is essentially a statement of Bellman principle of optimality. When $t_1 = T_1$ and $t_2 = T_2$, corresponding to the end of the speech sequences, the outer (time) warping process determines an overall warped metric distance $d_T(T_1, T_2)$ between the sequences $\mathcal{S}_1$ and $\mathcal{S}_2$, which may be referred as $d_T(\mathcal{S}_1, \mathcal{S}_2)$. The DTW also computes a sequence of indices $\mathbf{a}_T = (a_T(1), a_T(2), \ldots, a_T(N))$ and $\mathbf{b}_T = (b_T(1), b_T(2), \ldots, b_T(N))$, which are called the temporal warping function paths. The length N of the warping function paths may vary from one time sequence to another, depending on how many turns the temporal warping paths have, and N may be different form T_1 and T_2. These indices describe the temporal alignment for both source speech signal and target speech signal, such that the new temporal aligned spectral information vectors $\mathbf{s}_{1_{TA}}(a_T(i), :) = \mathbf{s}_1(b_T(i), :)$ $(i = 1, \ldots, N)$ and $\mathbf{s}_2$ are as similar as possible, where peaks and valleys of $\mathbf{s}_1$ align with peaks and valleys, respectively,

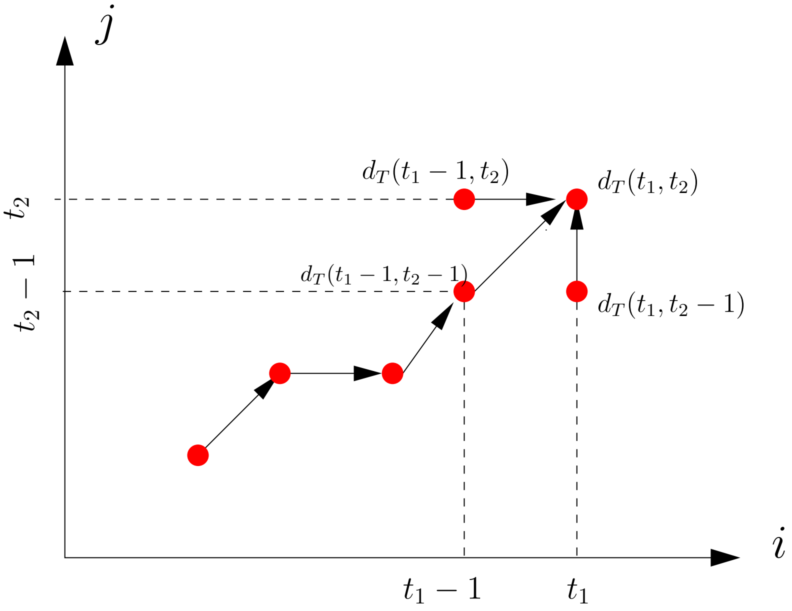


Fig. 2.1: Recursive way to find current cost distance (d_T).

of $\mathbf{s}_2$. This is portrayed in Figure 2.2. The subscript $_{TA}$ refers to Temporal Alignment warping process.

2.2.2 Inner Level Dynamic Warping

Analogous to the warping in the time performed by dynamic time warping (DTW), dynamic *frequency* warping (DFW) can warp the frequency dynamically to align spectral information.

The metric distance between frequency vectors $\mathbf{s}_1(t_1,:)$ and $\mathbf{s}_2(t_2,:)$ in (2.8), $d(\mathbf{s}1(t_1,:),\mathbf{s}2(t_2,:)$, can itself be computed using warping between these vectors. Because these vectors typically represent frequency information, this is called dynamic *frequency* warping (DFW), which is considered as the second level of warping. This warping is also referred to as the inner warping. DFW can be computed in a way similar to the outer warping. Let $\mathbf{s}_i(:) = \mathbf{s}_i(t_i,:)$, $i = 1,2$ denote the spectral information vectors at time block number t_i that is passed to the inner warping function (DFW) from the outer warping function (DTW). DFW is applied to calculate the distance between $\mathbf{s}_1(:)$ and $\mathbf{s}_2(:)$ as

$$d_F(k_1,k_2) = \mathrm{dist}(\mathbf{s}_1(k_1),\mathbf{s}_2(k_2)) + \min d_F(k_1-1,k_2), d_F(k_1,k_2-1), d_F(k_1-1,k_2-1) \qquad (2.9)$$

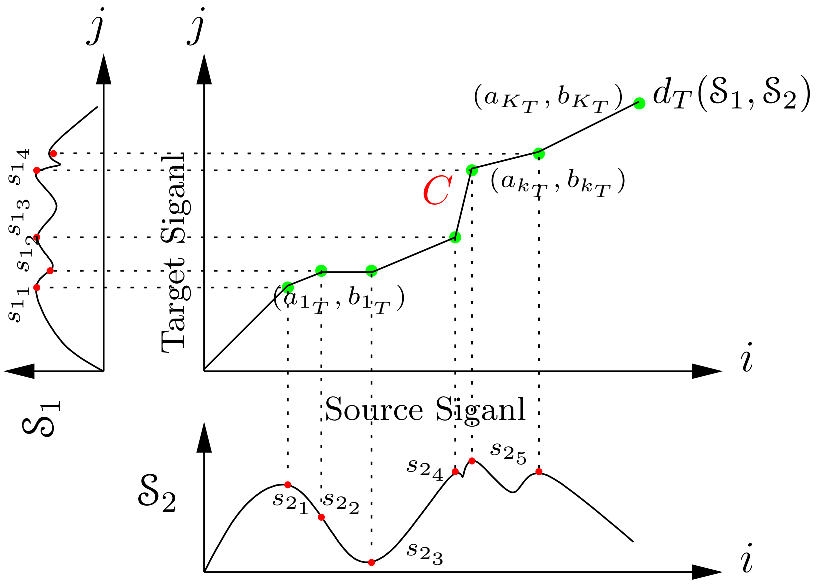


Fig. 2.2: Dynamic time (outer) warping.

The subscript $_F$ refers to a frequency (or spectral) warping process, and $\text{dist}(\mathbf{s}_1(k_1), \mathbf{s}_2(k_2))$ represents the metric distance between elements of the feature spectral vectors. At the end of frequency warping process, the distance between spectral vectors is computed as

$$d(\mathbf{s}_1(t_1, :), \mathbf{s}_2(t_2, :)) \triangleq d_F(K, K). \tag{2.10}$$

DFW produces another sequence of indices $\mathbf{a}_F = (a_F(1), a_F(2), \ldots, a_F(M))$ and $\mathbf{b}_F = (b_F(1), b_F(2),$

$\ldots, b_F(M))$, which describe the spectral warping function path. The temporal aligned source spectrum information, $\mathbf{s}_{1_{TA}}$, is warped to match the target spectrum information, $\mathbf{s}_2$, creating a modified source spectrum information $\hat{\mathbf{s}}_1$ according to

$$\hat{s}_{1_{TF}}(a_F(i)) = s_{1_{TA}}(b_F(i)), \qquad i = 1, \ldots, M \tag{2.11}$$

where the subscripts $_T$ and $_F$ refer to temporal warping process and frequency warping process, receptively. This spectral alignment mapping process drags spectral components of the temporal aligned source blocks to match the target spectrum information, $\mathbf{s}_2$. The length

M of the warping function paths may vary from one spectral feature vector to another, depending on how many turns the spectral warping paths have. The warping path in DFW is also subject to several constrains:

- **Monotonicity**:

  $$a_F(1) \leq a_F(2) \text{ and } b_F(1) \leq b_F(2).$$

  This constraint prevents the alignment path from moving back in time.

- **Continuity**:

  $$a_F(k) - a_F(k-1) \leq 1 \text{ and } b_F(k) - b_F(k-1) \leq 1.$$

  This constraint prevents skipping. This constraint ensures that the alignment process will not neglect or hide any samples.

- **Boundary**:

  $$a_F(1) = b_F(1) = 1,\ a_F(M) = K \text{ and } b_F(M) = K.$$

  The warping function must match the endpoints of the two time series functions, the alignment path must start and end at the bottom left and at the top right, respectively (diagonally opposite corner).

- **Warping Window**:

  $$|k_1 - k_2| < w_F,$$

  where $w_F > 0$ is the window length. DFW constrains the alignment path to not to go far from the diagonal.

The combination of the inner and outer warping processes applied to speech signal is portrayed in Figure 2.3. Starting at the bottom of the diagram, speech signal is split into different overlapping chunks. Spectral feature information is computed for each chunk. These spectral feature vectors information are passed through DTW for temporal alignment process, where at every stage of the temporal alignment process, spectral alignment process or warping process is achieved using DFW. The method for doing DTW is given in Algorithm 2.1

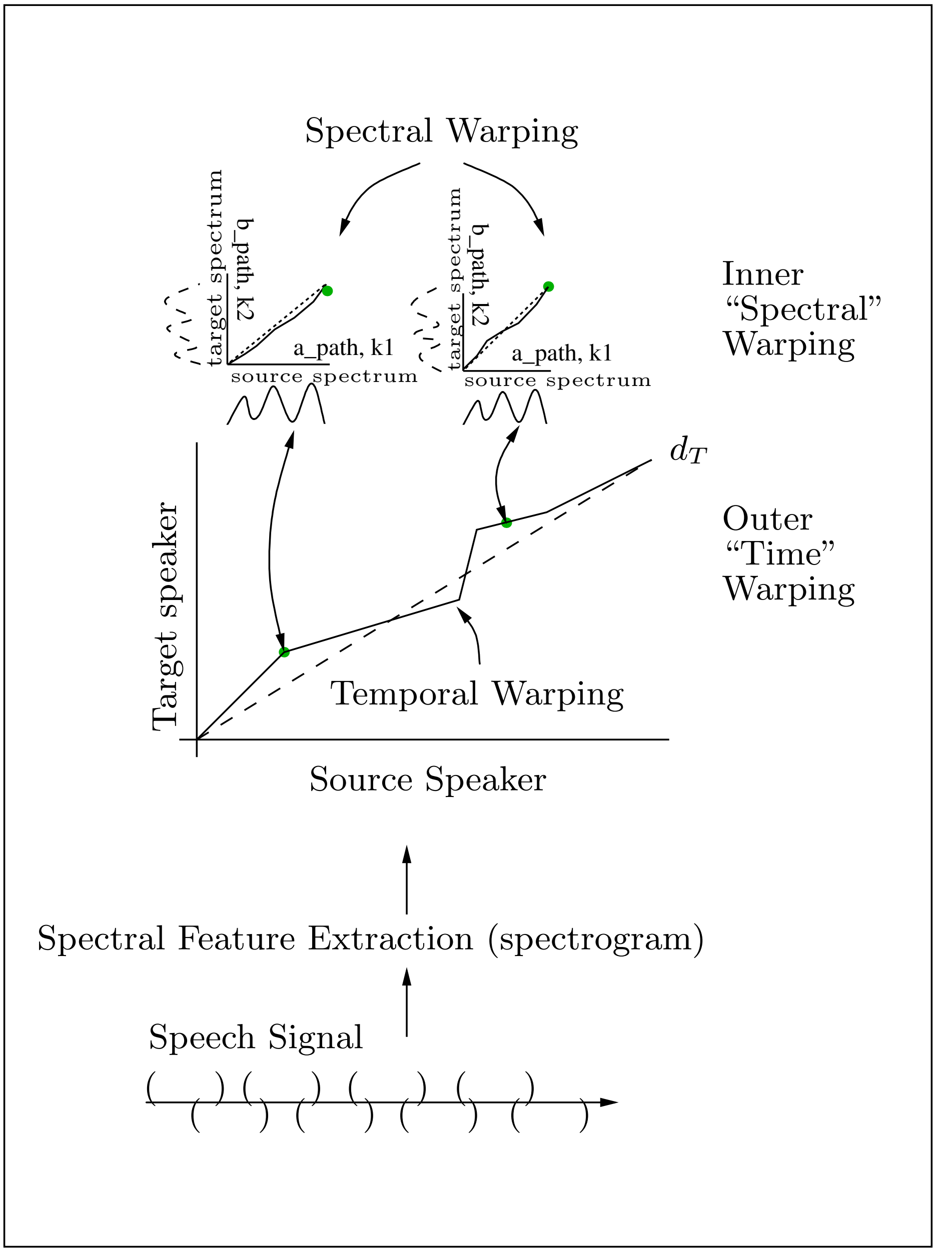


Fig. 2.3: Inner/Outer Dynamic Warping

Algorithm 2.1 Dynamic Wrapping (DW)

```
Input:
    First spectral sequence, $\mathcal{S}_1 = \{\mathbf{s}_1(1, 1 : n_T), \mathbf{s}_1(2, 1 : n_T), \ldots, \mathbf{s}_1(n_F, 1 : n_T)\}$
    Second spectral sequence, $\mathcal{S}_2 = \{\mathbf{s}_2(1, 1 : m_T), \mathbf{s}_1(2, 1 : m_T), \ldots, \mathbf{s}_1(m_F, 1 : m_T)\}$
    window size $w_T$ for DTW process
    window size $w_S$ for DFW process
Output:
    Distance between $\mathcal{S}_1$ and $\mathcal{S}_2$
    Indices $\mathbf{a}(N)$ and $\mathbf{b}(N)$ for Temporal alignment
    Indices $\mathbf{b}(M)$ and $\mathbf{b}(M)$ for spectral alignment
Begin
    Initialize DTW array, $DTW = array[0 \ldots n_T, 0 \ldots m_T]$
    Initialize DFW array, $DFW = array[0 \ldots n_F, 0 \ldots m_F]$
    Adapt window size, $w_T = max(w_T, abs(n_T - m_T))$
    For $i_T = 1$ to $n_T$
      For $j_T = 1$ to $m_T$
          $DTW[i_T, j_T] = \infty$
      End For $j_T$
    End For $i_T$
    Set $DTW[0, 0] = 0$
    For $i_F = 1$ to $n_F$
      For $j_F = 1$ to $m_F$
          $DTW[i_F, j_F] = \infty$
      End For $j_F$
    End For $i_F$
    Set $DFW[0, 0] = 0$
    For $i_T = 1$ to $n_T$
      For $j_T = max(1, i_T - w_T)$ to $min(m_T, i_T + w_T)$
          Adapt window size, $w_F = max(w_F, abs(n_F - m_F))$
          For $i_F = 1$ to $n_F$
            For $j_F = max(1, i_F - w_F)$ to $min(m_F, i_F + w_F)$
              $cost = norm(abs(log_{10}(\mathcal{S}_1(i_T, i_F))) - (log_{10}(\mathcal{S}_2(j_T, j_F)))$
             $DFW[i_F, j_F] := cost + minimum(DTW[i_F - 1, j_F],$
                                          $DFW[i_F, j_F - 1],$
                                          $DFW[i_F - 1, j_F - 1])$
           End For $j_F$
         End For $i_F$
         Searching minimum path through $DFW[i_F, j_F]$, save $\mathbf{a}_F$, $\mathbf{b}_F$
         $cost_{new} = DFW[n_F, m_F]$
           $DTW[i_T, j_T] := cost_{new} + minimum(DTW[i_T - 1, j_T],$
                                          $DTW[i_T, j_T - 1],$
                                          $DTW[i_T - 1, j_T - 1])$
    End For $j_T$
    End For $i_T$
    Searching minimum path through $DTW[i_T, j_T]$, save $\mathbf{a}_T$, $\mathbf{b}_T$
```

CHAPTER 3

TRAINING SPEECH IMITATION ACCURACY USING DYNAMIC WARPING

In this chapter, the method of two-level dynamic warping described in the previous chapter is applied to explore whether the processing can distinguish between reading and imitation attempts. Dysarthria is a neurological motor speech disorder that commonly results in reduced intelligibility. Communication partners can learn to better understand the speech of someone with dysarthria through perceptual training. Vocal imitation of the degraded speech during perceptual training has been shown to elevate this learning. This chapter presents a tool that could eventually be used to provide the learner with real-time feedback regarding the accuracy of their imitation attempts during training which may further enhance this learning. We describe a training tool that compares dysarthric speech productions with the imitation attempts of healthy subjects, using a two-level dynamic warp that accounts for both spectral and temporal degradation. Feature vectors derived from both the spectrogram and LPC are examined.

3.1 Background and Introduction

Many individuals have been born with neuro-motor speech disorders or have obtained it due to neurological injury or disease (e.g., stroke, traumatic brain injury, Parkinson's disease) which cause difficulties with speech production. These disorders are collectively called dysarthria.

Dysarthria characterized by spectral and temporal degradation in speech production, and typically results in reduced speech intelligibility. Speech intelligibility is defined as the accuracy with which a message is conveyed by a speaker and recovered by listener. This definition highlights the essential role of both speaker and listener in the communication process. Speech intelligibility has traditionally been viewed as an attribute of the speaker, but intelligibility is actually a function of both speakers and listeners [9]. Generally speaking, dysarthria management has focused primarily on the individual speaker (i.e. teaching patients to improve their ability to produce speech or strategies to compensate for their deficits). However, many individuals with dysarthria are not well-suited for traditional speaker-oriented intervention approaches due to concomitant difficulties [34].

Recent studies have shown that speech intelligibility can be improved by training the listener to better understand the degraded speech signal [11–14,35]. In other words, listeners can improve their ability to recognize and understand speech that is difficult to understand without placing additional demands on the speaker with a perceptual training. This perceptual training involves familiarizing the listener with the degraded speech patterns and written targets of what is being said. Pre- and post-tests reveal intelligibility improvements as a function of training. It has recently been shown that vocal imitation during training can increase the magnitude [12] and longevity [14] of intelligibility improvement following training. Further, a significant relationship between imitation accuracy during training and subsequent intelligibility improvements has been identified — subjects who were better at imitating the degraded speech signal were better able to understand it in subsequent encounters [12].

Motivated by this observation, it is desirable to develop and test assistive technology that aids vocal imitation in perceptual training paradigms. In this work, we explore a computer-based training tool to eventually be incorporated into a program to help train listeners to learn to imitate dysarthric speech, by providing subjects with feedback about the accuracy of their imitation attempts during training. The key to the learning tool is a means of comparing the productions of a speaker with dysarthria with the imitation attempts of someone without dysarthria (a healthy subject), in a way that accounts for both spectral and temporal variations. This is achieved using a two-level dynamic warping algorithm in both temporal and spectral domains.

As an initial step in the development and evaluation of this tool, we performed a human subject test to examine the following hypotheses: Is this computer-based DTW tool able to distinguish between mere repetition from written prompts (i.e. the subjects normal speaking voice) and attempts to imitate a dysarthric verbal prompt? Additionally, does having multiple opportunities to imitate the dysarthric prompt result in improved accuracy when compared to a single opportunity?

3.1.1 Computer-Based Training Tool Using Dynamic Warping

In this Section, DW, which is a combination of DTW and DFW algorithms, was applied to dysarthric speech to provide visual cues to coach individuals to understand how to modify their speech to provide better imitation (as a coach to them on how to modify their speech), by developing a computer-based tool which provide subjects with feedback about the accuracy of their imitation attempts during training. The learning key of the tool is to compare the dysarthric speech with the imitation speech attempts of a healthy speaker, in a way that accounts for degradations in both spectral and temporal domains. Because of the degradations in spectral and temporal of the dysarthric speech, both temporal and spectral alignment is used to make a comparison. So, this comparison is done using DW.

Our description of this tool starts with a two speakers, a speaker with dysarthria and a healthy speaker. According to (2.6), let $\mathcal{S}_1$ be the sequence of dysarthric speech feature vectors (target speaker) and $\mathcal{S}_2$ be the sequence of healthy speech feature vectors (subject speaker). The minimum cost , (2.8), between sequences $\mathcal{S}_1$ and $\mathcal{S}_2$ is recursively computed. At the end of speech sequences, there is a cost value, which is the overall warped distance between the sequence $\mathcal{S}_1$ and $\mathcal{S}_2$, which may be denoted as $d_T(\mathcal{S}_1, \mathcal{S}_2)$, where the subscript $_T$ emphasizes that this is warping in time. As mentioned in Chapter 2, DTW computes a sequence of indices $\mathbf{a}_N$ and $\mathbf{b}_N$, such that the sequence $\mathbf{s}_1(\mathbf{a}_N, :)$ and $\mathbf{s}_2(\mathbf{b}_N, :)$ is minimized. If $\mathcal{S}_1$ and $\mathcal{S}_2$ are temporally aligned, then $T_1 = T_2$, and $\mathbf{a}_N = \mathbf{b}_N = (1, 2, \ldots, T_1) \overset{\triangle}{=} \mathbf{I}_0$. The amount of deviation of $(\mathbf{a}_N, \mathbf{b}_N)$ from $(\mathbf{I}_0, \mathbf{I}_0)$ is a measure of how much the path of sequences had to be "stretched" to achieve a best-matching alignment. Let $D_T(\mathcal{S}_1, \mathcal{S}_2)$ denote the path stretch distortion function. There are two ways of measuring the amount of distortion between the signals. There is the distance between the signals themselves $d_T(\mathcal{S}_1, \mathcal{S}_2)$, (2.8), and the path distortion $D_T(\mathcal{S}_1, \mathcal{S}_2)$. This is portrayed in Figure 3.1, which schematically illustrates the warping between a target speaker sequence and a subject speaker sequence.

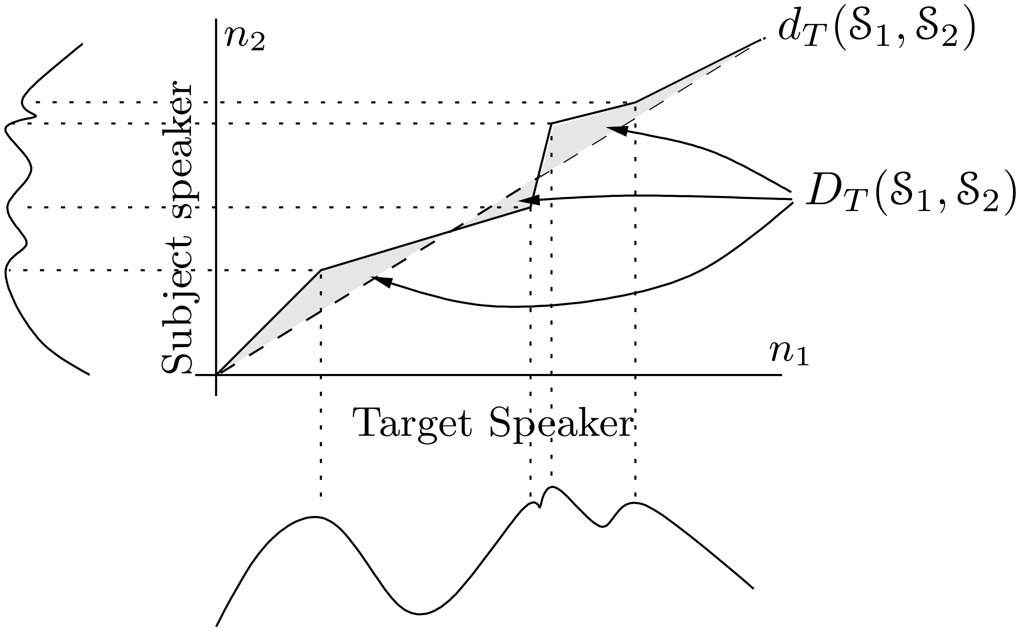

n_2
$d_T(\mathcal{S}_1, \mathcal{S}_2)$
$D_T(\mathcal{S}_1, \mathcal{S}_2)$
Subject speaker
n_1
Target Speaker

Fig. 3.1: Warping and distance measures for dynamic time (outer) warping

Figure 3.2 shows different ways (area measure, horizontal distance and vertical measure) of measuring this path distortion distance. In this work, area measure between the line determined by $(\mathbf{I}_0, \mathbf{I}_0)$ and the path $(\mathbf{a}_N, \mathbf{b}_N)$ was considered.

As mentioned in Chapter 2, the metric distance between frequency vectors $\mathbf{s}_1$ and $\mathbf{s}_2$ in (2.8), $d(\mathbf{s}_1(t_1, :), \mathbf{s}_2(t_2, :)$, can itself be computed using warping between these spectral

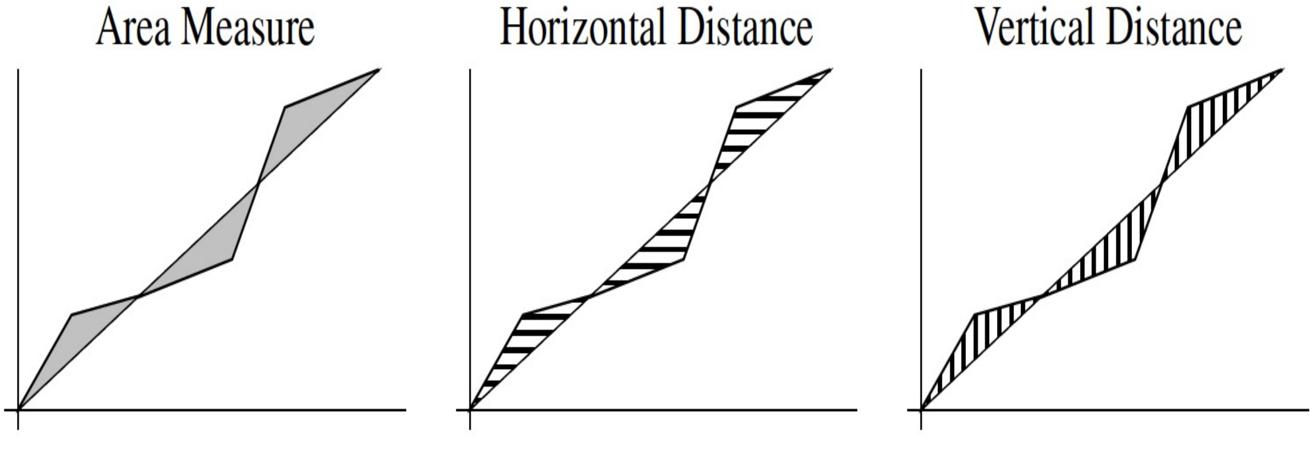


Fig. 3.2: Different measuring ways for distortion distance

At the end of frequency warping process, the distance between spectral vectors, $d_F(k_1, k_2)$, become $d_F(K, K)$, where K is the number of elements in each spectral feature vector and where the subscript $_F$ refers to a frequency (or spectral) warping process. As mentioned in section (2.2.2), DFW produces two spectral warping function paths ($\mathbf{a}_M$ and $\mathbf{b}_M$), which can be used to compute a measure of the spectral distortion $D_F(\mathbf{s}_1, \mathbf{s}_2)$ needed to obtain the best match of spectral features. The spectral distance $d_F(\mathbf{s}_1, \mathbf{s}_2)$, (2.9), and the path distortion $D_F(\mathbf{s}_1, \mathbf{s}_2)$ are combined into a scalar value using α_F as a weighting parameter. When there are low energy speech segments (such as from unvoiced speech) which have no particular spectral information to match, the distance $d_F(\mathbf{s}_1, \mathbf{s}_2)$ is generally meaningless. In order to downplay this effect, when the signals are combined, the minimum energy, $E_{\min} = \min(\|\mathbf{s}_1\|^2, \|\mathbf{s}_2\|^2)$ is used to scale the distance. The distance returned from the inner DFW to be used in 2.8 is thus

$$\mathbf{d}_{ret} = E_{\min} d_F(\mathbf{s}_1, \mathbf{s}_2) + \alpha_F D_F(\mathbf{s}_1, \mathbf{s}_2) \overset{\triangle}{=} d(\mathbf{s}_1(t_1, :), \mathbf{s}_2(t_2, :)) \tag{3.1}$$

where the experimental validation found good performance with $\alpha_F = 0.1$. The method for doing DW for this tool is shown in Algorithm 3.1. The difference between this one and Algorithm 2.1, Chapter 2, was highlighted in blue color.

3.2 Feature Extraction

In most applications of speech processing, the speech signal is processed to extract useful features (or parameters). These parameters keep most of the information needed to recognize and identify the spoken units and the speech characteristics. Spectral features information are believed to cover more speaker information, like individuality [36]. So, for the purpose of making the analysis of the speech signal more convenient and also the manipulation, speech signal should be represented with more suitable features. Speech data in this tool was sampled at 8000 sample/sec. Different kinds of spectral information were used in our experiments.

Algorithm 3.1 DW to compare Dysarthric speech and healthy subject

```
Input:
       Dysarthric spectral sequence, $\mathcal{S}_1 = \{\mathbf{s}_1(1, 1 : n_T), \mathbf{s}_1(2, 1 : n_T), \ldots, \mathbf{s}_1(n_F, 1 : n_T)\}$
       Subjected spectral sequence, $\mathcal{S}_2 = \{\mathbf{s}_2(1, 1 : m_T), \mathbf{s}_1(2, 1 : m_T), \ldots, \mathbf{s}_1(m_F, 1 : m_T)\}$
       window size $w_T$ for DTW process
       window size $w_F$ for DFW process
Output:
       Distance between $\mathcal{S}_1$ and $\mathcal{S}_2$
       $\mathbf{D}_T$
Begin
       Initialize DTW array, $DTW = array[0 \ldots n_T, 0 \ldots m_T]$
       Initialize DFW array, $DFW = array[0 \ldots n_F, 0 \ldots m_F]$
       Adapt window size, $w_T = max(w_T, abs(n_T - m_T))$
       For $i_T = 1$ to $n_T$
         For $j_T = 1$ to $m_T$
           $DTW[i_T, j_T] = \infty$
         End For $j_T$
       End For $i_T$
       Set $DTW[0, 0] = 0$
       For $i_F = 1$ to $n_F$
         For $j_F = 1$ to $m_F$
           $DTW[i_F, j_F] = \infty$
         End For $j_F$
       End For $i_F$
       Set $DFW[0, 0] = 0$
       For $i_T = 1$ to $n_T$
         For $j_T = max(1, i_T - w_T)$ to $min(m_T, i_T + w_T)$
           Adapt window size, $w_F = max(w_F, abs(n_F - m_F))$
           For $i_F = 1$ to $n_F$
             For $j_F = max(1, i_F - w_F)$ to $min(m_F, i_F + w_F)$
               $cost = norm(abs(log_{10}(\mathcal{S}_1(i_T, i_F))) - (log_{10}(\mathcal{S}_2(j_T, j_F)))$
               $DFW[i_F, j_F] := cost + minimum(DFW[i_F - 1, j_F],$
                                                $DFW[i_F, j_F - 1],$
                                                $DFW[i_F - 1, j_F - 1])$
           End For $j_F$
         End For $i_F$
         Searching minimum path through $DFW[i_F, j_F]$, save $\mathbf{a}_F$, $\mathbf{b}_F$
         $cost_{new} = DFW[n_F, m_F]$
         Area between $[(\mathbf{a}(1), \mathbf{b}(1)), (\mathbf{a}(end), \mathbf{b}(end))]_F$ and $(\mathbf{a}, \mathbf{b})_F$
         Minimum energy $[\mathcal{S}_1(i_T, :)$ and $\mathcal{S}_2(j_T, :)]$
         $\mathbf{d}_{ret} = \mathbf{E}_{min}\ \mathbf{cost_{new}} + \alpha_{\mathbf{F}}$ area
           $DTW[i_T, j_T] := \mathbf{d}_{ret} + minimum(DTW[i_T - 1, j_T],$
                                                $DTW[i_T, j_T - 1],$
                                                $DTW[i_T - 1, j_T - 1])$
       End For $j_T$
       End For $i_T$
       $\mathbf{D}_T$ = Area between $[(\mathbf{a}(1), \mathbf{b}(1)), (\mathbf{a}(end), \mathbf{b}(end))]_T$ and $(\mathbf{a}_T, \mathbf{b}_T)_T$
```

3.2.1 Linear Predictive Coding Magnitude Spectrum

Linear Predictive Coding (LPC), often call an *autoregressive* (AR) model in other domains [37], predicts the main features of speech using all-pole system function H(z) where

$$H(z) = \frac{S(z)}{E(z)} = \frac{1}{1 - \sum_{i=1}^{K} a_i z^{-i}} = \frac{1}{A(z)} \tag{3.2}$$

where $S(z)$ is the z-transform of the output speech signal, $E(z)$ is the z-transform of the input speech signal and A(z) is the inverse filter. The LPC model preserves the acoustic information in the speech [37].

In our experiments, 14 LPC coefficients are used in each frame window of 25ms with 50% overlap. The coefficients are converted to an LPC magnitude spectrum $|1/H(e^{j\omega})|$ for $K = 50$ values of ω in $[0, \pi]$, so that the spectrum can be meaningfully warped.

3.2.2 Spectrogram Features

Speech signals like many other signals change spectral content with time, like many other signals. A spectrogram displays the temporal characteristics of the signal and provides information about the distribution of frequency components of the signal as it varies with time. The spectrogram plot contains time information along the x-axis, frequency information along the y-axis, and the amount of energy amplitude in the signal at any given time and frequency is displayed as a level of grey (or color). The spectrogram is computed by calculating the Short-Time Fourier Transform ($STFT$). The STFT is computed by computing the FFTs of segments of data samples, where each sample maybe overlapped in time [38, 39].

In this work, the spectrogram is computed using 25-ms segments of speech sample, windowed using a Hamming window, with 50% overlap transformed using a 256-point FFT. The $K = 128$ positive frequency elements were used as the feature vector.

3.2.3 Pitch Information

Air from our lungs causes the vocal cords to vibrate. These vibrations acts as an excitation source, which is controlled by the mass and tension of the vocal cords [3]. The vibrations make the cords to open and close, which breaks the air-stream up into pulses. The repetition rate of the pulses is known as *pitch*. Generally speaking, pitch in speech is the “highness” or “lowness” of a tone as perceived by the ear, which depends on the number of vibrations per second produced by the air passing through the vocal cords. Pitch is the perceptual correlate of tone and intonation. There are many algorithms for pitch estimation [40]. In this work MATLAB pitch command was used to find the pitch information for each phrase.

3.3 Method

A clinical test was performed on the speech data to determine if the speech feature vectors and dynamic warping (DW) are able to distinguish between healthy subjects reading a phrase in their ”own voice” and healthy subjects imitating that same phrase produced by a speaker with dysarthria.

3.3.1 Participants

Thirty two young, healthy adults (23 women and 9 men) participated in the experiment. All participants were native speakers of American English. Per self-report, participants had no history of speech, language, or cognitive disorders and no prior experience with individuals with dysarthria. Participants were recruited from undergraduate classes at Utah State University. Institutional review board consent was obtained before the start of the experiment.

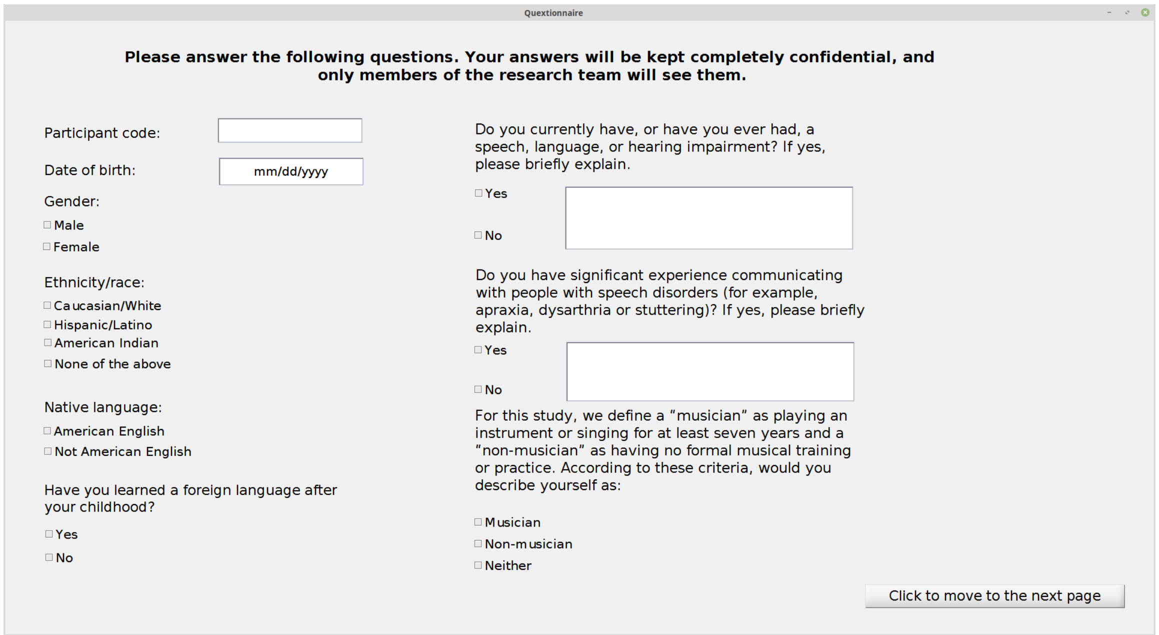
3.3.2 Speech Stimuli

The speech stimuli consisted of 40 semantically plausible phrases ranging in length from four to eight words, containing between four and 12 syllables per phrase. Phrases elicited from a 26-year-old male native speaker of American English with dysarthria secondary to traumatic brain injury. The speaker who provided the speech stimuli presented with a moderate spastic dysarthria, as diagnosed by three independent speech-language pathologists with expertise in assessment and diagnosis of motor speech disorders. Speech was characterized perceptually by monopitch, slow speaking rate, imprecise articulation, and a strained–strangled vocal quality. As reported in [12], acoustic analysis of the dysarthric phrases, relative to the same phrases produced by an age and gender-matched healthy control, confirmed that the dysarthric phrases were characterized by reduced F0 variation (manifested perceptually as reduced pitch variation or monotone) and slow speaking rate.

3.3.3 Procedure

The experiment was conducted in three phases. All participants (32 subjects) were involved in the three phases of the experiment identically. Each participant took the whole experiment session (three phases), at one sitting time, in a quiet lab at Utah State University. Upon obtaining permission, each participant was seated in front of computer preloaded with the experimental computer application, designed using MATLAB Graphical user interfaces (GUIs). Each participant was informed that the whole experiment would be delivered via the computer program. All participant were fitted with sound-attenuating headphones. The participant were informed that they would be listen a short phrases produced by a speaker with dysarthria, they were told that the phrases contained real English words (e.g., "The bread is stale"). The experiment was started first by filling a questionnaire page, Figure 3.3. In the page, each participant is asked to answer all the questions and not skip anyone otherwise the program return an error with the personal information of the participant. The program saves all the questionnaire information for all participants into one Excel file.

- **Phase 1** : The actual experiment begins with this phase for one participant at a time. During this phase, each participant is first instructed to read aloud each of the 40 phrases of the speech stimuli using his or her own voice and recording the spoken phrase. The phrases were presented one at a time. (As shown in Figure 3.4, all the instructions regarding phase 1 were written on this computer designed page.) After the participant had finished recording the 40 phrases of the speech stimuli using his or her normal voice, he or she was prompted to move on to the next phase, Figure 3.5.



Quextionnaire

Please answer the following questions. Your answers will be kept completely confidential, and only members of the research team will see them.

Participant code:

Date of birth: mm/dd/yyyy

Gender:
□ Male
□ Female

Ethnicity/race:
□ Caucasian/White
□ Hispanic/Latino
□ American Indian
□ None of the above

Native language:
□ American English
□ Not American English

Have you learned a foreign language after your childhood?
□ Yes
□ No

Do you currently have, or have you ever had, a speech, language, or hearing impairment? If yes, please briefly explain.
□ Yes
□ No

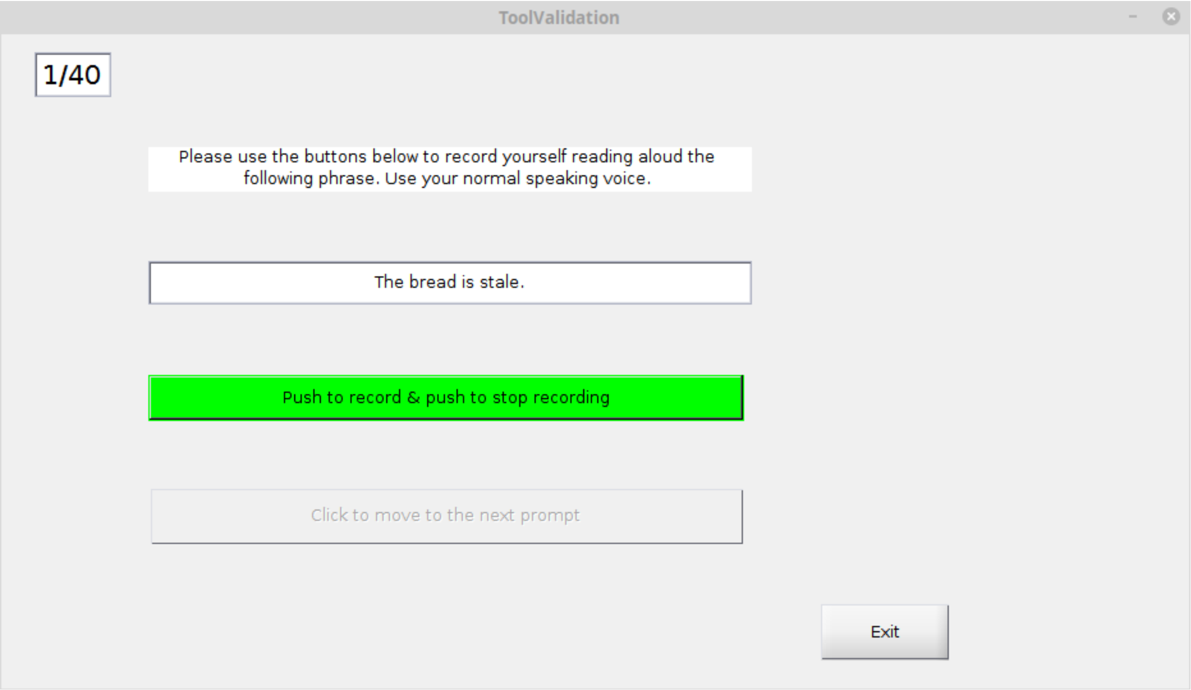
Do you have significant experience communicating with people with speech disorders (for example, apraxia, dysarthria or stuttering)? If yes, please briefly explain.
□ Yes
□ No

For this study, we define a "musician" as playing an instrument or singing for at least seven years and a "non-musician" as having no formal musical training or practice. According to these criteria, would you describe yourself as:
□ Musician
□ Non-musician
□ Neither

Fig. 3.3: Questionnaire page of the clinical test

- **Phase 2** : Each participant hears a single phrase once, one phrase at a time, spoken by a speaker with dysarthria. The participant is instructed to imitate the phrase as best as possible and recording his or her imitation attempt. As shown in Figure 3.5, all the instructions for this phase were written down in the designed page. The participant repeats the recording procedure for phase 2 for the 40 phrases of the speech stimuli. When the participant records his or her 40 imitation attempts, he or she can press the button "*Continue to next prompt*" to move to the next phrase, Figure 3.6. In the below discussion, this phase referred to as the "one time " attempt.

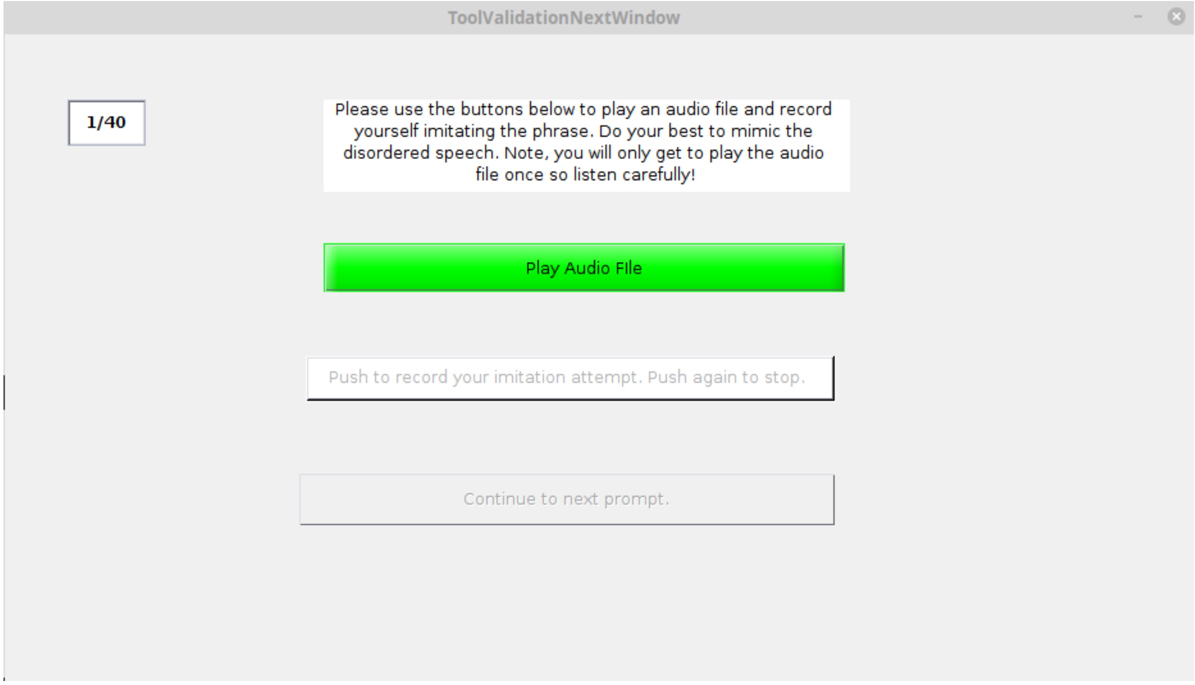
- **Phase 3** : This phase consider the last technical part of this experiment. In phase 3, instead of hearing the spoken prompt once, the participant is instructed to play the prompt as often as desired, providing opportunity to practice imitating before recording his or her imitation attempt. As shown in Figure 3.6, all the instructions were written down in the designed page for this specific phase. The participant can play the dysarthric speech as often as they choose and practice as much as can but the participant can record once his or her imitation attempt. Again, each participant


ToolValidation
1/40
Push to record & push to stop recording
Click to move to the next prompt
Exit

Please use the buttons below to record yourself reading aloud the following phrase. Use your normal speaking voice.

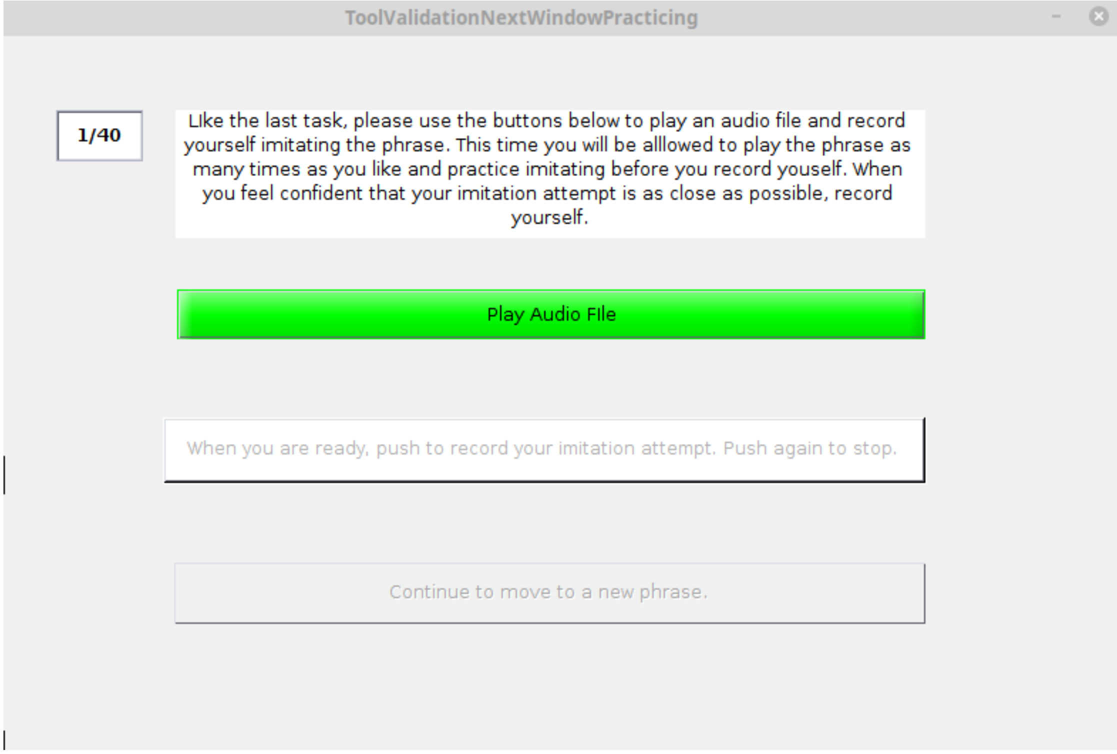
The bread is stale.

Fig. 3.4: Own voice recording phase of the clinical test



Please use the buttons below to play an audio file and record yourself imitating the phrase. Do your best to mimic the disordered speech. Note, you will only get to play the audio file once so listen carefully!

Fig. 3.5: First attempt imitation phase of the clinical test


ToolValidationNextWindowPracticing
1/40
Play Audio FIle
When you are ready, push to record your imitation attempt. Push again to stop.
Continue to move to a new phrase.

LIke the last task, please use the buttons below to play an audio file and record yourself imitating the phrase. This time you will be alllowed to play the phrase as many times as you like and practice imitating before you record youself. When you feel confident that your imitation attempt is as close as possible, record yourself.

Fig. 3.6: Multi attempts imitation phase of the clinical test

should go with the recording. This is referred to as the “multi” attempt .

3.4 Analysis

The total data set consisted of 1280 recorded phrases (.wav) for each phase. After identifying the starting and ending points of each speech signal for the three phases, spectral feature vectors are extracted for each 25 ms segment of speech, with 50% overlap, using either LPC magnitude spectrum or FFT-based features. Distance and path distortion measures $d_T(\mathcal{S}_1, \mathcal{S}_2)$ and $D_T(\mathcal{S}_1, \mathcal{S}_2)$ are computed for each segment, comparing the subject recording against the target (dysarthric speech) recording. The participant recordings of “one-time attempt” and “multi attempt” were compared against the “own voice” attempt. Here $\mathcal{S}_1$ is the target speech, and $\mathcal{S}_2$ is one of the participant speech attempts.

It is known that speakers with dysarthria tend to have lower variance in their pitch production [12]. For an alternative comparison, the pitch variance was also computed to see if the participants were matching pitch variance in their attempt to imitate. The MATLAB `pitch` command was applied to 52 ms speech segments with 41 ms overlap. The resulting pitch estimates were smoothed using bidirectional median filtering to remove meaningless pitch estimates for unvoiced speech, then further smoothed using the MATLAB `smooth command`. The variance of the resulting pitch estimates was computed as a scalar measure, denotes as PV($\mathcal{S}$), where $\mathcal{S}$ denotes either target, own voice, one-time, or multi speech signals. The method for doing this whole clinical test is given in Algorithm 3.2

3.4.1 Results

Figure 3.7 shows the histogram of distortion measure D_T comparing the target voice with the own voice, one-time, and multi attempts. Also indicated is the mean value of the information. It is apparent that D_T(target,one-time) is typically less than D_T(target,own-voice), indicating that the participants are better matched to the target when they attempt to imitate than when they read in their own voice. The feature vector used in all of these is the LPC feature because the LPC-based feature vector performed significantly better in the classifier than the FFT-based feature vector. Figure 3.8 shows the histogram of d_T comparing the target with the own voice, one-time, and multi attempts These measures are not as effective at distinguishing between own voice and imitation attempts.

Figure 3.9 shows the histogram of the pitch variance (PV) for the target, across all phrases. Figure 3.9(b,c,d) shows the histogram of the PV for own-voice, one-time, and multi, across all phrases and all participants. From the mean values, it is apparent that the pitch variation does decrease with imitation, with the multi attempt being close to the target.

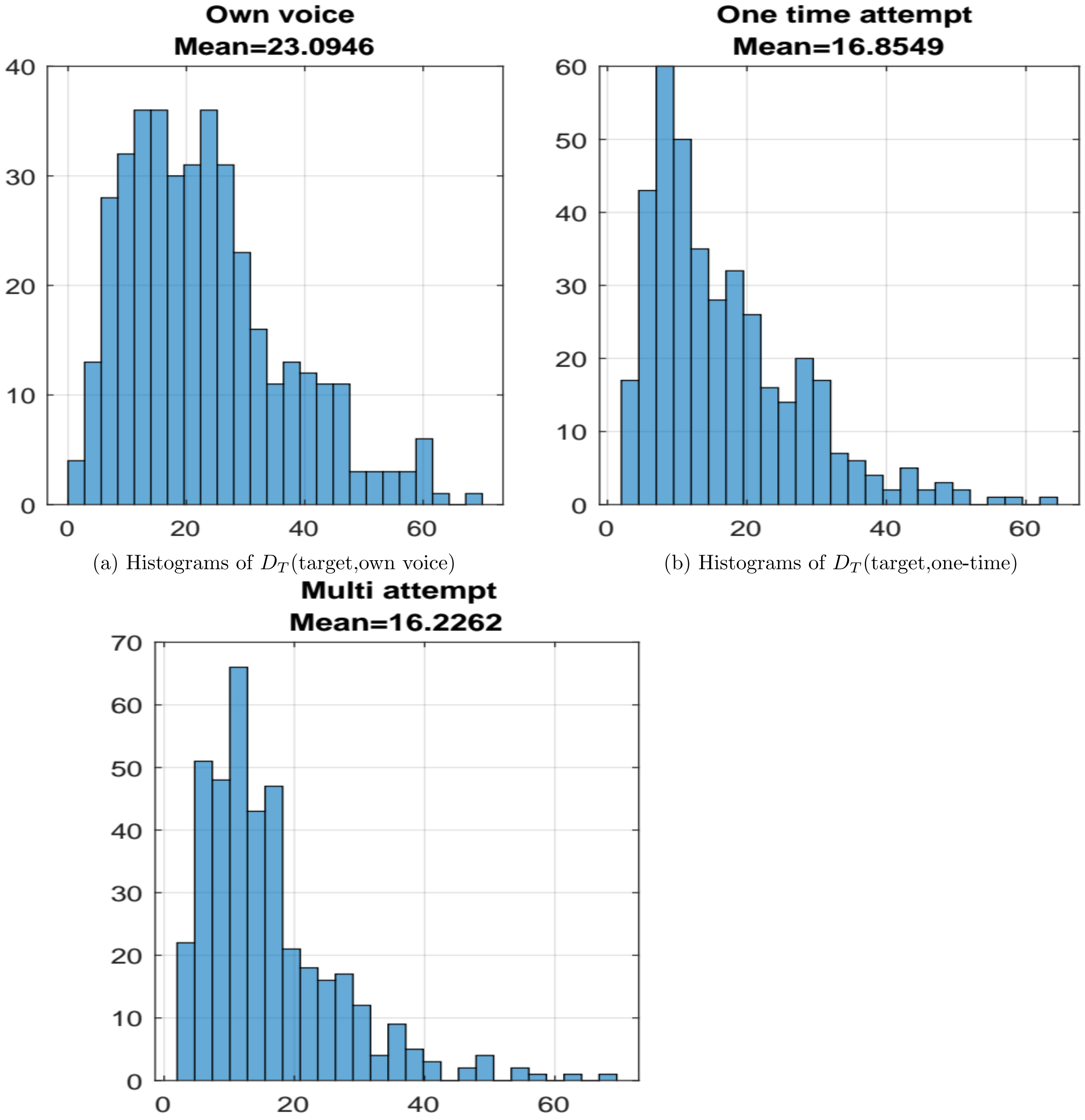
These features were used in simple classifiers to determine the overall performance. For D_T scores, the classifier is defined follows:

$$\text{if } D_T(\text{target,one-time}) \leq D_T(\text{target,own-voice}) \text{ then success}$$

Success for the D_T score classifier means that the participants are better matched to the target when they attempt to imitate than when they read in their own voice, similarly for the d_T scores classifier and for the multi data. The probability of success is tabulated in

Algorithm 3.2 Whole clinical test using two-level DW

```
Input:
        Dysarthric speech information, $target\_speech(n)$, $n$: number of phrases
        Own voice recording information, $own\_speech(m)$, $m$: number of phrases
        One attempt imitation recording information, $one\_speech(m)$, $m$: number of phrases
        Multi attempt imitation recording information, $multi\_speech(m)$, $m$: number of phrases
        $p$: number of participant
Output:
        Histogram
        Classification probability
Begin
        For $i = 1$ to $n$
          For $j = 1$ to $p$
            Function: Identifying Starting and ending points
                $target\_clip(i) = identify\_start\_end\_function(target\_speech(i))$
                $own\_clip(i,j) = identify\_start\_end\_function(own\_speech(i,j))$
                $one\_clip(i,j) = identify\_start\_end\_function(one\_speech(i,j))$
                $multi\_clip(i,j) = identify\_start\_end\_function(multi\_speech(i,j))$
            Function: LPC magnitude spectrum features
                $target\_lpc(i) = lpc\_mag\_function(target\_clip(i))$
                $own\_lpc(i,j) = lpc\_mag\_function(own\_clip(i,j))$
                $one\_lpc(i,j) = lpc\_mag\_function(one\_clip(i,j))$
                $multi\_lpc(i,j) = lpc\_mag\_function(multi\_clip(i,j))$
            Function: Spectrogram features
                $target\_spect(i) = spect\_function(target\_clip(i))$
                $own\_spect(i,j) = spect\_function(own\_clip(i,j))$
                $one\_spect(i,j) = spect\_function(one\_clip(i,j))$
                $multi\_spect(i,j) = spect\_function(multi\_clip(i,j))$
            Function: Algorithm 3.1
                $[d_{T,lpc}, D_{T,lpc}]_{(target,own)} = dtw\_algorithm(\mathcal{S}_1 = target\_lpc(i), \mathcal{S}_2 = own\_lpc(i,j))$
                $[d_{T,lpc}, D_{T,lpc}]_{(target,one)} = dtw\_algorithm(\mathcal{S}_1 = target\_lpc(i), \mathcal{S}_2 = one\_lpc(i,j))$
                $[d_{T,lpc}, D_{T,lpc}]_{(target,multi)} = dtw\_algorithm(\mathcal{S}_1 = target\_lpc(i), \mathcal{S}_2 = multi\_lpc(i,j))$
            Function: Algorithm 3.1
                $[d_{T,spec}, D_{T,spec}]_{(target,own)} = dtw\_algorithm(\mathcal{S}_1 = target\_spect(i), \mathcal{S}_2 = own\_spect(i,j))$
                $[d_{T,spec}, D_{T,spec}]_{(target,one)} = dtw\_algorithm(\mathcal{S}_1 = target\_spect(i), \mathcal{S}_2 = one\_spect(i,j))$
                $[d_{T,spec}, D_{T,spec}]_{(target,multi)} = dtw\_algorithm(\mathcal{S}_1 = target\_spect(i), \mathcal{S}_2 = multi\_spect(i,j$
            Function: Pitch variation
                $PV(target) = pitch\_function(target\_clip(i,j))$
                $PV(own) = pitch\_function(own\_clip(i,j))$
                $PV(one) = pitch\_function(one\_clip(i,j))$
                $PV(multi) = pitch\_function(multi\_clip(i,j))$
          End For $j_T$
        End For $i_T$
        Plot Histogram
        Find classification portability
```

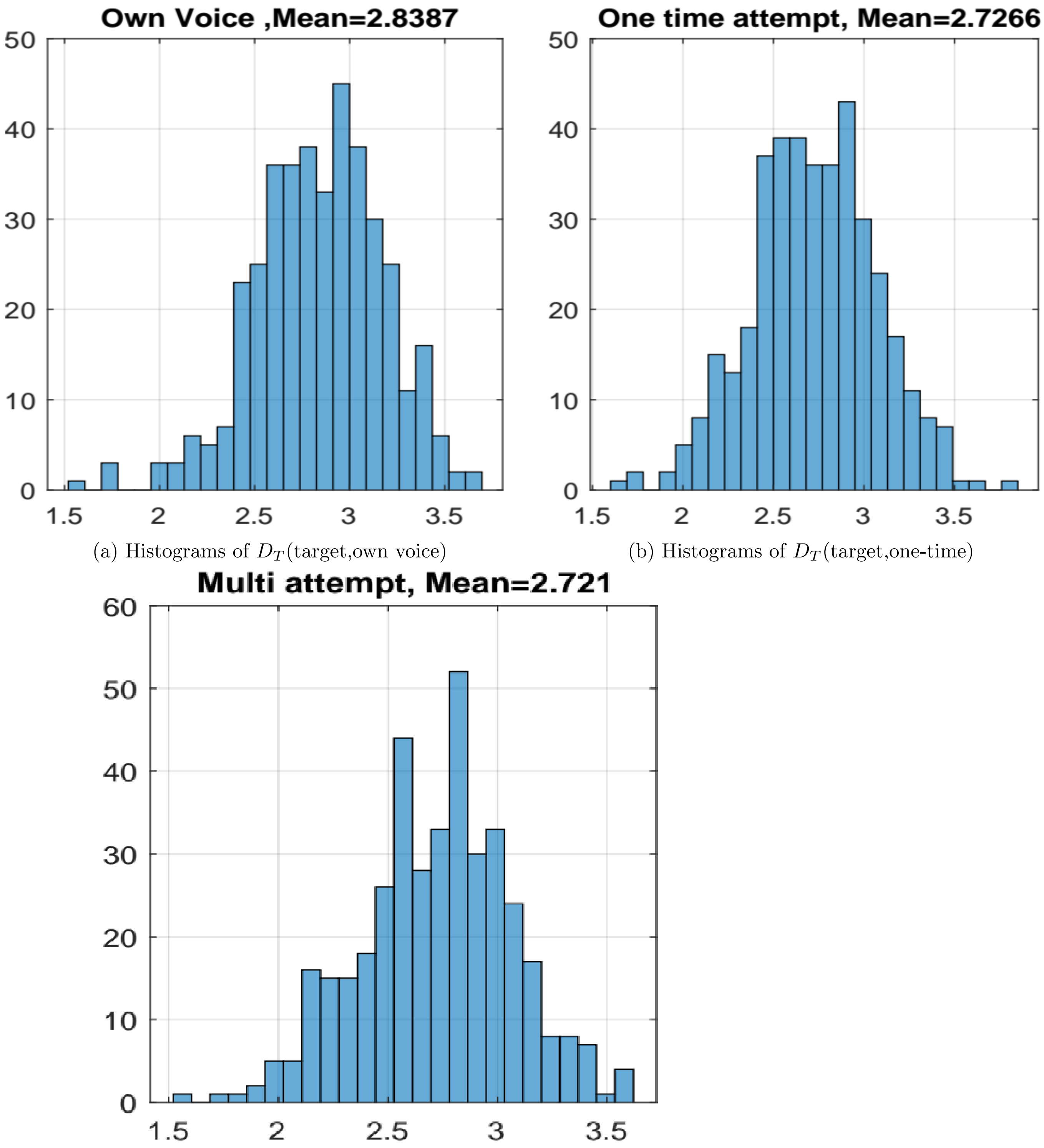


(a) Histograms of D_T(target,own voice)

(b) Histograms of D_T(target,one-time)

(c) Histograms of D_T(target,multi)

Fig. 3.7: Histograms of D_T across all participants (males and females) and all phrases

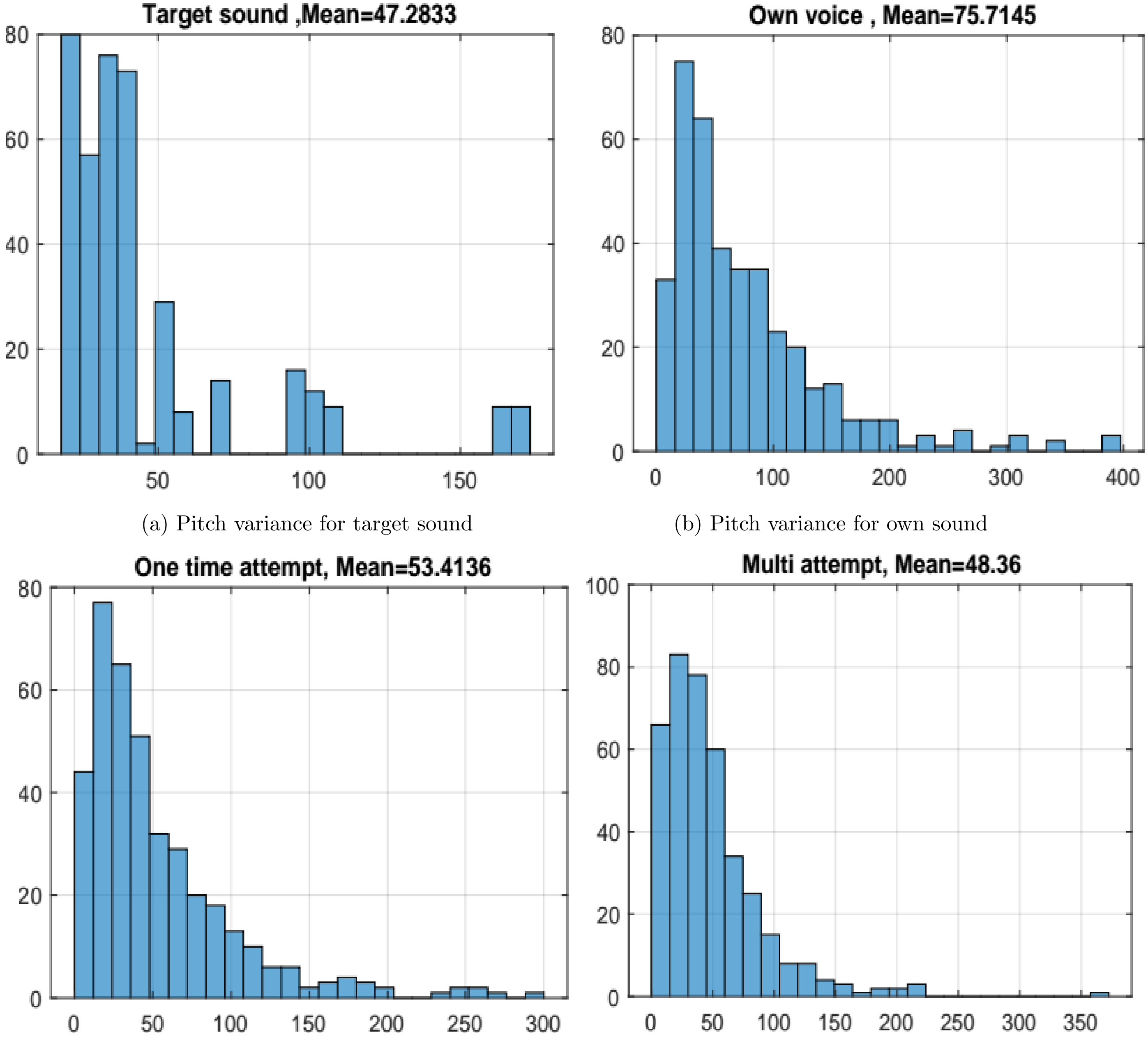


(a) Histograms of D_T(target,own voice)

(b) Histograms of D_T(target,one-time)

(c) Histograms of D_T(target,multi)

Fig. 3.8: Histograms of d_T across all participants (males and females) and all phrases



(a) Pitch variance for target sound

(b) Pitch variance for own sound

(c) Pitch variance for one time attempt sound

(d) Pitch variance for multi attempt sound

Fig. 3.9: Histograms of pitch variance

table 3.1, where the testing is done over all phrases for all participants.

The classifier for the PV is

$$\text{if PV(one-time) } < \text{PV(own voice) then success}$$

Also, success means that the pitch variance for the participants are better matched to the pitch variance of the target when they attempt to imitate than when they read in their own voice (and similarly for the multi data). The probability of success for this classifier is also shown in table 3.1.

Test	Classification Probability
D_T own *vs.* D_T one-time	87.5%
D_T own *vs.* D_T multi	90.1%
d_T own *vs.* d_T one-time	87.8%
d_T own *vs.* d_T multi	90.6%
$\mathbf{d}_{ret}$ own *vs.* $\mathbf{d}_{ret}$ one-time	91.8%
$\mathbf{d}_{ret}$ own *vs.* $\mathbf{d}_{ret}$ multi	89%
PV(own) *vs.* PV(one-time)	67.0%
PV(own) *vs.* PV(multi)	64.5%

Table 3.1: Comparison Results

Classification performance is summarized in Table 3.1. As shown, the distortion score alone D_T achieved 87% correct classifying between “own voice” and “one-time”. This raises slightly for classifying with the “multi” speech. This provides some indication that practice may produce an improvement, but that the attempts to imitate produces such a dramatic difference that the difference is barely distinguishable (using this tool). Table 3.1 also shows the performance when the distance and distortion are combined according to $\mathbf{d}_{ret} = d_T + \beta D_T$ (where $\beta = 0.1$). Slight improvements are observed.

Table 3.1 also shows the results of using the pitch variance to distinguish between “own voice” and imitation attempts. Pitch variation on its own does not perform as well as the DW methods.

3.5 Summary

The method of two-level dynamic warping, described in Chapter 2, was applied to dysarthric speech to help train listeners to learn to imitate dysarthric speech. This chapter presents a computer-based training tool that could eventually be used to provide care givers with feedback about the accuracy of their imitation attempts during training. The key to the learning tool is a means of comparing the productions of a speaker with dysarthria with the imitation attempts of someone without dysarthria (healthy subject), in a way that accounts for both spectral and temporal variations. This is achieved using a two-level dynamic warping algorithm. Clinical test was performed on the speech data to determine if the speech feature vectors and the two-level DW are able to distinguish between healthy subjects reading a phrase in their "own voice" and healthy subjects imitating that same phrase produced by a speaker with dysarthria. The results presented in Table 1 based on this clinical study indicate that the analysis performed on the speech signals is able to distinguish between own voice and imitation attempts with high probability.

CHAPTER 4

TRAINING SPEECH IMITATION ACCURACY USING DYNAMIC WARPING

Voice transformation, for example, from a male speaker to a female speaker, is achieved here using a two-level dynamic warping algorithm. An outer warping process, which temporally aligns blocks of speech (dynamic time warp, DTW), invokes an inner warping process, which spectrally aligns based on magnitude spectra (dynamic frequency warp, DFW). The mapping function produced by the dynamic frequency warp is used to move spectral information from a source speaker to a target speaker. This warping mapping process involves only spectral magnitudes, and has been found to introduce significant deleterious signal processing artifacts. It has been found that reconstruction of phase information significantly improves the quality of transformed speech. Information obtained by this process is used to train an artificial neural network to produce spectral warping output information based on spectral input data. Objective evaluation measure of spectral features and warping paths was applied.

4.1 Introduction

Speech is an important and essential oral human communication tool. Speech processing and synthesis is an important and interesting subject today. One example of speech processing is the modification of speech of a person as if it is spoken by another person. This is called voice transformation.

Voice transformation (VT) refers to the process of changing the parameters of the speech or changing voice personality, to convert the speech uttered by one speaker (source speaker) to sound as if other speaker (target speaker) had spoken it, for example, from a male speaker to a female speaker [15]. Voice transformation has applications such as text-to-speech synthesis (TTS), international dubbing, health-care, multi-media, language education, music, security-related usage, vocal restoration, speech-to-speech translation, and preprocessing for speech recognition, etc. [41–43]. A perfect voice transformation system should convert the following characteristics from a speaker:

- **Vocal Tract Characteristics**
- **Prosody Characteristics**
- **Glottal Excitation**

These characteristics are very important and related to speaker identification and that eventually affect the VT process.

4.2 Voice Transformation Background

The most common case of VT is done when the source speaker and target speaker are speaking the same language, that means no need to change the language content from the original speech signal. Cross lingual VT (which means when the source speaker and the target speaker are spoke different languages) has also been achieved [44]. Cross lingual VT is not within the scope of this work.

Figure 4.1 shows typical framework for VT system. VT, usually, can be achieved in two stages.

- **Training Stage**: which is offline voice transformation stage.
- **Online Voice Transformation – Synthsis Stage**: a real-time process.

The inputs for the training stage are the source speech signal and target speech signal. Both signals pass through a speech analysis model to extract the information of acoustic feature parameters, like pitch and formant information. After this analysis, an appropriate mapping function is devised to perform voice transformation by mapping the acoustic features of the source speaker into the acoustic features of the target speaker.

In the transformation stage, the input to this stage is the source speech signal only. This signal passes through speech analysis module to extract the suitable features. Once these features information are computed, they pass through the transformation function

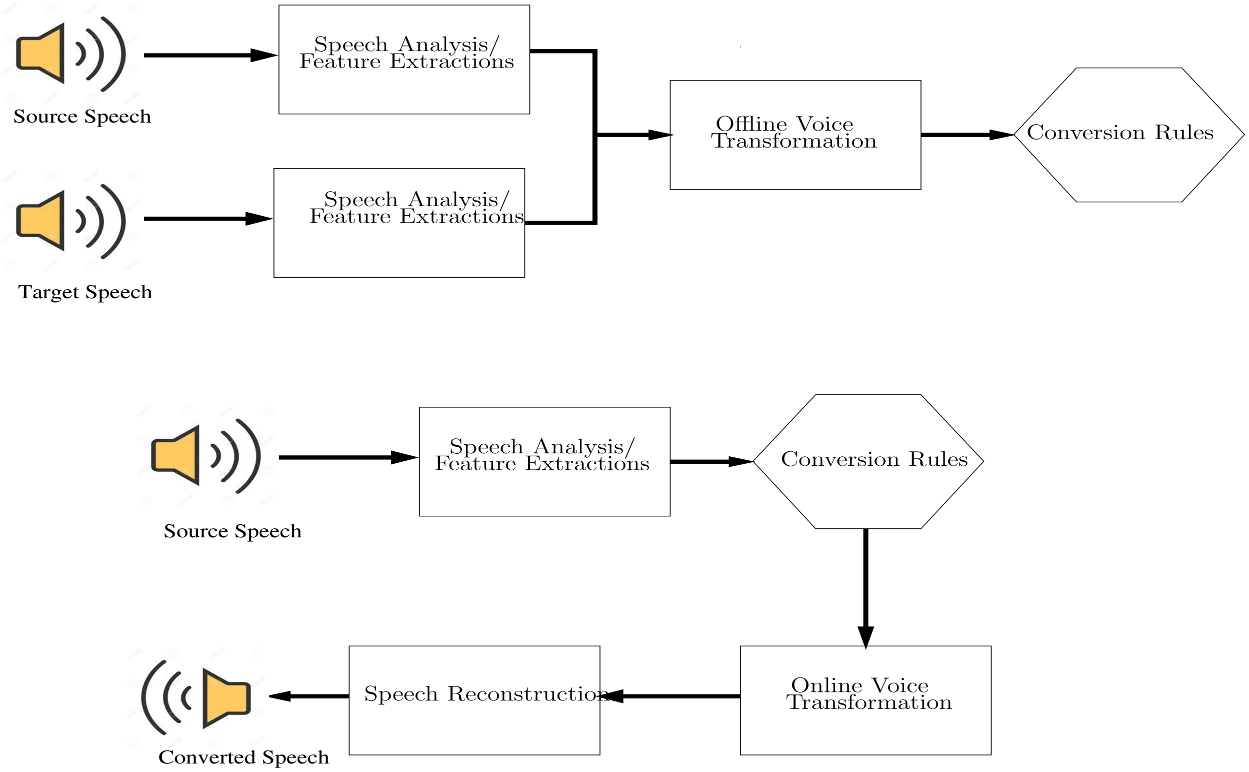

Source Speech
Speech Analysis/ Feature Extractions
Target Speech
Speech Analysis/ Feature Extractions
Offline Voice Transformation
Conversion Rules
Source Speech
Speech Analysis/ Feature Extractions
Conversion Rules
Online Voice Transformation
Speech Reconstruction
Converted Speech

Fig. 4.1: Typical Speech Transformation System

model, obtained from training stage, to perform the mapping and produce a new vectors of feature information. The transformed features are passed through a speech synthesis step (speech reconstruction module) to produce the transformed speech signal.

The speech analysis model represents the speech signal as vectors of feature information that have enough represent the whole speech signal, while the reconstruction model is responsible to reconstruct or recreate the speech signal from the transformed acoustic feature vectors. Generally speaking, the speech signal does not enter the transformation model directly. As shown in Figure 4.1, transformation is done on the feature information.

4.3 Voice Transformation's Related Works

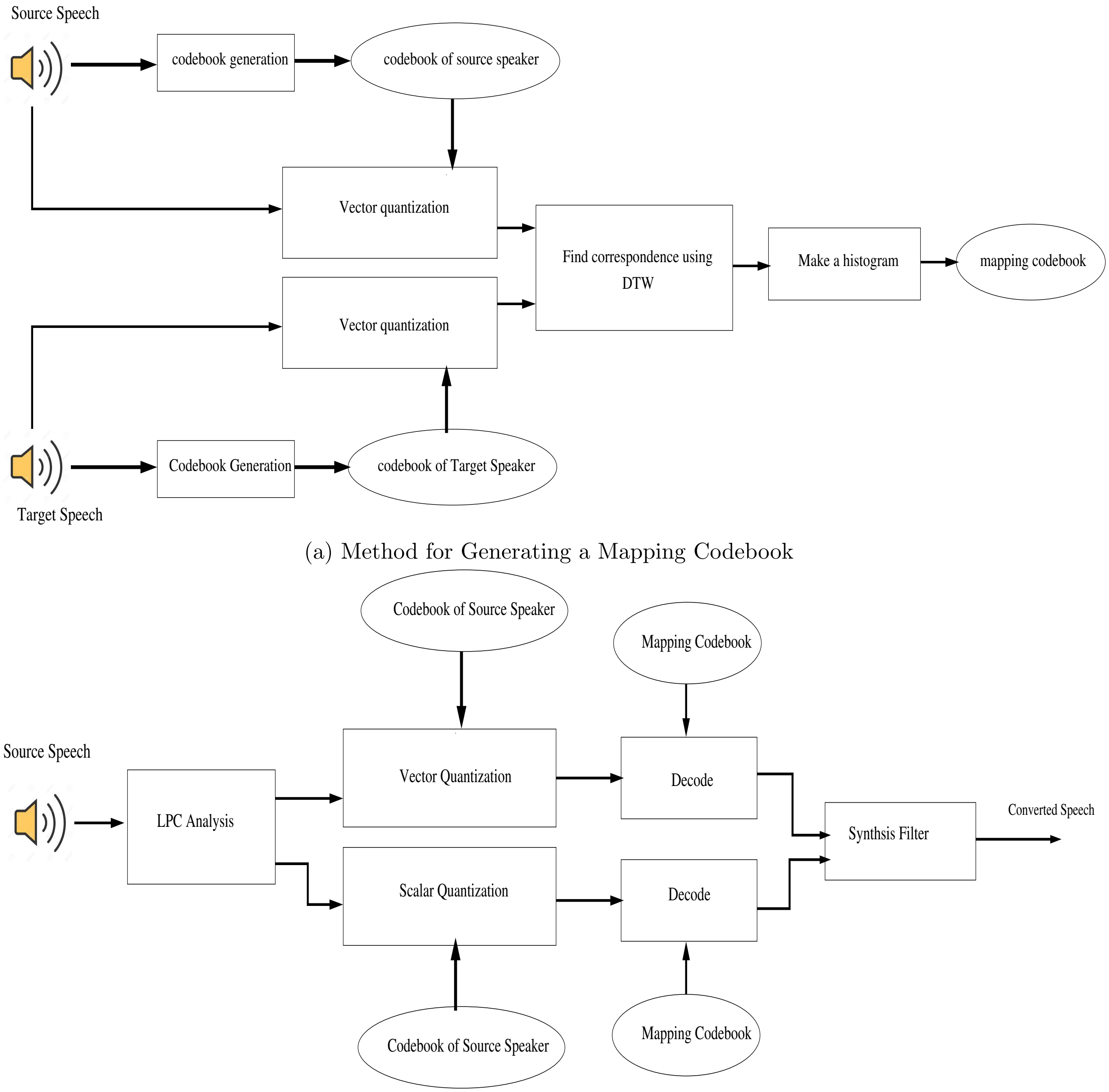
Many techniques have been proposed by researchers for the voice transformation especially for the mapping function. In this section, we will discuss some of these important techniques.

4.3.1 Codebook Mapping Method

Codebook method, [45], is considered as an early work on voice transformation. The authors in [45] proposed a way of using vector quantization (VQ) with codebook and spectrum mapping to do the transformation. Other work, [46, 47], show that the acoustic features data, like pitch information, formant frequencies and bandwidth, spectral tilt, etc, are crucial features that relate to speech individuality. While it is difficult to control the speech individuality by modifying all acoustic feature parameters independently, generating codebooks can be used with vector quantization to represent all these features. This work include two steps: first, learning step; and second, conversion-synthesis step. These steps are similar to the model portrayed by Figure 4.1. In the first step, codebooks were generated for both source speakers and target speaker. After that, the speech generated by speakers were vectors quantized using speaker's codebook. Dynamic time warping was used to find the correspondence between these vectors. These correspondences are then accumulated to find a histogram, which is used as a weighting function to compute the mapping codebook from source speaker to a target speaker. In the real time step, the speech was quantized using speaker's codebook, then all feature parameters are transformed with the mapping codebook form training step, and finally the speech was synthesized with LPC vocoder. Figure 4.2 illustrates the block diagram of the procedure for generating a mapping codebook and the conversion-synthesis step.

4.3.2 Dynamic Frequency Warping

The converted speech using mapping codebooks method has a voice quality turned out to be poor. [48] propose a method of doing voice transformation based on PSOLA technique (PSOLA stands for Pitch Synchronous Overlap and Add) to improve the quality of the converted speech signal. In this work, the authors introduced dynamic frequency warping to the voice transformation. Dynamic frequency warping (DFW) attempts to compensate for the differences between acoustic features spectral information of the source speaker and the acoustic features of the target speaker by finding an optimal non-linear warping function [49]. While the changing in the vocal tract length produces a non-linear transformation of acoustic



(a) Method for Generating a Mapping Codebook

(b) Voice Conversion from a Source Speaker to a Target Speaker

Fig. 4.2: Block Diagram for Mapping Codebooks Method

parameters, DFW is closely related to the acoustic theory of speech production. They extracted cepstral information form both source and target speakers to do frequency warping. In [50, 51] and [52, 53], the authors proposed using DFW and bilinear frequency warping ($BLFW$), respectively. Because of the frequency warping method does not modify the relative amplitude of the meaningful parts of the spectrum information, the accuracy and quality of the converted speech signal is moderate.

The work in [51–53] achieved voice transformation in which frequency warping is complemented with some type of amplitude scaling (AS) to compensate for the spectral conversion inaccuracy and improve the spectral modification. AS modifies the vertical axis of the frequency-warped spectra by means of corrective filters

4.3.3 Gaussian Mixture Model

Gaussian mixture model (GMM) is a model used in many pattern recognition techniques [54], whose efficiency for text-independent speaker recognition and has been illustrated by many studies [55, 56]. GMM may be the most popular method proposed for speech conversion [57–61]. GMM is used in speech transformation because of its ability to model the acoustic parameters of a speaker as a combination of several components [55].

Speech conversion is performed by computing a linear transformation function. While using a general linear transformation limits the performance of the conversion process [62], GMM is considered a practical solution by modeling the source data and the target data with a Gaussian mixture model to produce transformation functions for each Gaussian.

Two approaches are mainly used for GMM-based speech transformation techniques: first modeling the source information with a GMM [57, 58] and second modeling the joint density GMM ($JDGMM$) between the feature of the source speaker and the feature of the target speaker [59–61]. The distribution of the source speaker's spectral space is modeled with a GMM to estimate the parameters (mean vectors and the covariance matrices). The conversion function is typically assumed to be probabilistic piece-wise linear mapping function for each Gaussian. The unknown parameters are calculated by solving the normal linear conversion equations for a least squares solution [57, 58].

JDGMM is the most popular approach. The transformation function can be optimized by using different objective function. The most popular functions are the Mean Square Error (MSE) and the Maximum Likelihood Estimation (MLE). The combination of the spectral source aligned vectors and the corresponding target aligned spectral vectors is used to estimate GMM parameters for the joint density model by using Expectation Maximization (EM) algorithm [62]. Both approaches need two separate stages (training and online transformation stages) to do the transformation process. The computation of the Gaussian distribution parameters is part of the training stage.

4.3.4 Neural Network Methods

Many researchers tried to apply neural network techniques to achieve speech transformation problem because of the general success of neural network (NN) technique in the field of machine learning analyses and applications. While vocal tract shape between two speakers is nonlinear and the neural networks are exceptionally good at learning nonlinear models, NN was employed in mapping the source speech feature vectors into the feature vectors of the target speaker [63].

An artificial neural network (ANN) form family of models inspired by biological neural networks [64]. An ANN model organized as a set of layers that contains interconnected nodes, where each node represents an artificial neuron, with weight associated with each interconnection between two nodes [63]. The network is initialized with initial weights, then the network adjusts its weights to establish a relationship between the input data and the output data that comprise the training data. The network learn to estimate, classify, and make predictions from new data based by training the network to find relationships between input data and output data.

Work by M. Narendranath, [65], was considered as one of the first attempts of using ANN to transform the source speaker formants to target speaker formants. In [66], the authors proposed a method for voice conversion using NN with three layers. This method based on LPC spectral features using radial basis function neural network (RBF). However, both techniques in [65, 66], used carefully prepared trained data which prepared carefully.

Also, for both the source and target speakers, they manually select the regions of vowels or syllable. This is an inconvenient way to align correctly the source and target features for real-world application scenarios.

In [67], the authors compare between using an ANN to achieve the voice conversion and the state-of-the-art GMM. GMMs capture the joint distribution of the source features and the target feature, while the work in [67] directly maps the spectral source feature information onto the spectral target feature information. Also, with GMM they use Maximum likelihood parameters generation ($MLPG$) to obtain a smooth trajectory of spectral features, while the mapping with ANN provide best transformation results without using MLPG. Two stages, training and online transformation, are used for the ANN to achieve the speech transformation. During the training stage of the work in [67], 25 Mel-cepstral coefficients are extracted from the recording source and target speakers, the back-propagation is used to adjust the weights of the NN. At the online transformation, the features to be transformed are propagated from input, first layer, until last layer. The new transformed features are used to recreate the new converted speech.

In contrast to the traditional way of using GMM method, using restricted Boltzmann machines (RBMs) was proposed in [68] as a probability density model to model the joint distributions of the source spectral feature information and target spectral feature information, where they replaced the RBM. The aim behind this step of replacing is to achieve better capturing for the correlations between the joint spectral features of the source and target speakers. After this work of using RBM, Nakashima *et al.* proposed using of two Deep Belief Network (DBF) networks connected by a simple feed-forward NN to build high-order eigen spaces of the source/target speakers to achieve a high-order feature space that can be converted [69].

4.4 Voice Transformation Process

4.4.1 Two-Level Dynamic Warping for the Transformation

Voice transformation is a topic that still has a lot to be explored. The approach taken here avoids the need to find acoustic parameters, such us pitch or formant model, or to model the joint density between the feature of the source speaker and the feature of the target speaker, instead deals directly with spectral information. The transformation is accomplished using a two-level dynamic warp (DW). Based on the two-level DW it is straightforward to map the source speech to target speech when both are available.

It is obvious from the name "dynamic time warping" that DTW, or outer dynamic warping, temporally aligns block of features to compensate for different speech rates. For example, temporal alignment between male speech segments and female speech segments by comparing certain features occurring as a function of time. Dynamic frequency warping (DFW) or "inner dynamic warping" can also be used to perform spectral alignment based on spectral magnitudes of blocks of speech data, such as aligning spectral features of male speaker to spectral features of female speaker. In this work, the combination of inner and outer warping, simply referred to "dynamic warping" or DW, is used to achieve the transformation.

Our speech transformation process starts with a two speakers, source speaker and target speaker. Following (2.6), let $\mathcal{S}_2$ be the sequence of target speech feature vectors and $\mathcal{S}_1$ be the sequence of source speech feature vectors. The minimum cost, (2.8), between sequences $\mathcal{S}_1$ and $\mathcal{S}_2$ is recursively computed, allowing some alignment motion between blocks of the two different users in both the time and frequency domains. At the end of that specific portion of speech sequences, time warping produces an overall warped metric distance $d_T(T_1, T_2)$, (2.6), between the sequences $\mathcal{S}_1$ and $\mathcal{S}_2$, which may be denoted as $d_T(\mathcal{S}_1, \mathcal{S}_2)$, where the subscript $_T$ emphasizes that this is warping in time. This is suggested in Figure 4.3. This DTW accounts for temporal shifts due to the rate of speaking differences between the two speakers. The output of this DTW is the temporal alignment of the source signal with respect to the target signal.

The $\text{dist}(\mathbf{s}_1(k_1), \mathbf{s}_2(k_2))$ in (2.9) represents the metric distance between elements of the frequency feature vectors. In this work, we use spectral feature vectors computed as the positive frequency components of FFT of windowed data. The metric distance used looks

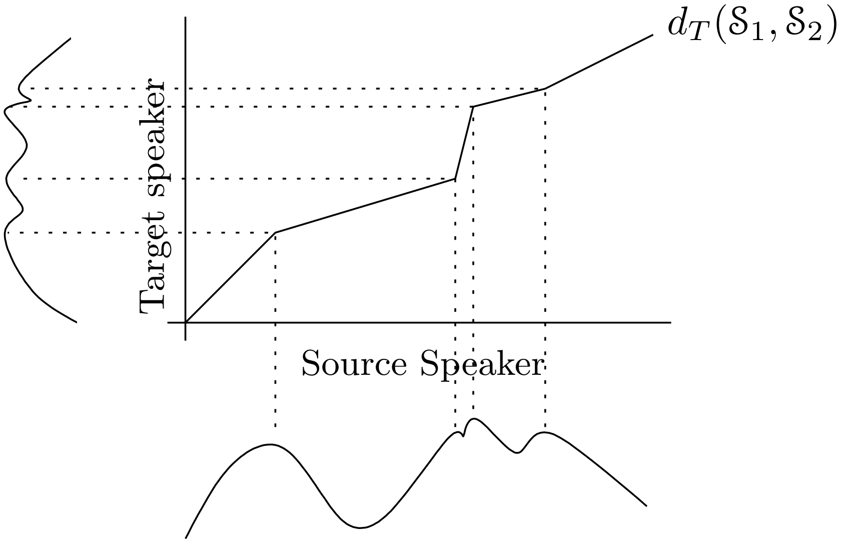


Fig. 4.3: Distnce Measure for Outer Dynamic Warping for VT

only at the magnitude of the spectral information, so

$$\text{dist}(\mathbf{s}_1(k_1), \mathbf{s}_2(k_2)) = |\,|\mathbf{s}_1(k_1)| - |\mathbf{s}_2(k_2)|\,|. \tag{4.1}$$

At the end of frequency warping process, the returned metric distance between spectral vectors to be used in (2.8) is computed as

$$d(\mathbf{s}_1(k_1, :), \mathbf{s}_2(k_2, :)) \overset{\triangle}{=} d_F(K, K), \tag{4.2}$$

where K is the number of elements in each spectral feature vector.

At the end of the temporal alignment process, DTW produces a sequence of indices $\mathbf{a}_T = (a_T(1), a_T(2), \ldots, a_T(N))$ and $\mathbf{b}_T = (b_T(1), b_T(2), \ldots, b_T(N))$, which are called the temporal warping function paths. Also, DFW produces another sequence of indices $\mathbf{a}_F = (a_F(1), a_F(2), \ldots, a_F(M))$ and $\mathbf{b}_F = (b_F(1), b_F(2), \ldots, b_F(M))$, which are called the spectral warping function paths. The method for doing two-level dynamic warping to achieve voice transformation is given in Algorithm 4.1. The difference between this one and Algorithm 2.1, Chapter 2, was highlighted in blue color.

Algorithm 4.1 Dynamic Wrapping (DW) for Voice Transformation

```
Input:
    First spectral sequence, $\mathcal{S}_1 = \{\mathbf{s}_1(1, 1 : n_T), \mathbf{s}_1(2, 1 : n_T), \ldots, \mathbf{s}_1(n_F, 1 : n_T)\}$
    Second spectral sequence, $\mathcal{S}_2 = \{\mathbf{s}_2(1, 1 : m_T), \mathbf{s}_1(2, 1 : m_T), \ldots, \mathbf{s}_1(m_F, 1 : m_T)\}$
    window size $w_T$ for DTW process
    window size $w_F$ for DFW process
Output:
    Distance between $\mathcal{S}_1$ and $\mathcal{S}_2$
    Indices $\mathbf{a}(N)$ and $\mathbf{b}(N)$ for Temporal alignment
    Indices $\mathbf{b}(M)$ and $\mathbf{b}(M)$ for spectral alignment
Begin
    Initialize DTW array, $DTW = array[0 \ldots n_T, 0 \ldots m_T]$
    Initialize DFW array, $DFW = array[0 \ldots n_S, 0 \ldots m_F]$
    Adapt window size, $w_T = max(w_T, abs(n_T - m_T))$
    For $i_T = 1$ to $n_T$
      For $j_T = 1$ to $m_T$
          $DTW[i_T, j_T] = \infty$
      End For $j_T$
    End For $i_T$
    Set $DTW[0, 0] = 0$
    For $i_F = 1$ to $n_F$
      For $j_F = 1$ to $m_F$
          $DTW[i_F, j_F] = \infty$
      End For $j_F$
    End For $i_F$
    Set $DFW[0, 0] = 0$
    For $i_T = 1$ to $n_T$
      For $j_T = max(1, i_T - w_T)$ to $min(m_T, i_T + w_T)$
          Adapt window size, $w_F = max(w_F, abs(n_S - m_F))$
          For $i_F = 1$ to $n_F$
            For $j_F = max(1, i_F - w_F)$ to $min(m_F, i_F + w_F)$
              $cost = abs(abs(\mathcal{S}_1(i_T, i_F))) - (abs(\mathcal{S}_2(j_T, j_F)))$
              $DFW[i_F, j_F] := cost + minimum(DTW[i_F - 1, j_F],$
                                                $DFW[i_F, j_F - 1],$
                                                $DFW[i_F - 1, j_F - 1])$
            End For $j_F$
          End For $i_F$
        Searching minimum path through $DFW[i_F, j_F]$, save $\mathbf{a}_F$, $\mathbf{b}_F$
        $cost_{new} = DFW[n_F, m_F]$
          $DTW[i_T, j_T] := cost_{new} + minimum(DTW[i_T - 1, j_T],$
                                                $DTW[i_T, j_T - 1],$
                                                $DTW[i_T - 1, j_T - 1])$
    End For $j_T$
    End For $i_T$
    Searching minimum path through $DTW[i_T, j_T]$, save $\mathbf{a}_T$, $\mathbf{b}_T$
```

4.4.2 Speech Database

Current voice conversion techniques need a parallel database [59,70] in which the source and target speakers record the same set of utterances. In this work, the speech data was carried out on CMU ARCTIC database. The CMU ARCTIC databases were constructed at the Language Technologies Institute at Carnegie Mellon University as phonetically balanced, designed for the purpose of speech synthesis research [71]. The ARCTIC database consists of seven primary sets of recordings, recorded by two US males, one Canadian male with English accent, one Scottish male with English accent, one Indian male, and two US females. Each speaker recorded a set of 1132 English utterances, most being between one and four seconds long. In experiments presented here one US male and one US female are chosen to do the voice transformation from male to female and to apply phase reconstruction on the output transformed voice.

4.4.3 Spectral Feature Extraction

In order to perform spectral mapping, the network must be fed with spectra from source and target speakers utterances or some representation of it. To extract features for this experiment, speech data was sampled at 16000 samples/sec. These feature vectors that were used in the two-level DW were considered to be the positive-frequency spectral information in the frequency domain, calculated using the FFT. Each sentence was temporally segmented into 32-ms segments, using a Hamming window with 16-ms overlapping, zero-padded, then transformed using a 512-point FFT. The K = 256 positive frequency spectral elements in the frequency domain were used as a feature vector.

4.4.4 Analysis

Two-level DW was applied on the spectral magnitudes feature vectors that are extracted from the FFT for each 32-ms segment of a phrase with 50% overlap windowed using a Hamming window. When the temporal alignment reaches the end of the specific speech segment, DTW produces a sequence of indices ($\mathbf{a}_T$ and $\mathbf{b}_T$). These indices describe the temporal alignment for source speech signal as

$$\mathbf{s}_{1_{TA}}(:, a_T(i)) = \mathbf{s}_1(:, b_T(i))), i = 1, 2, \ldots, N. \tag{4.3}$$

such that the new temporal aligned spectral information $\mathbf{s}_{1_{TA}}(t_1, :)$ and $\mathbf{s}_2(t_2, :)$ are as similar as possible (e.g., peaks and valleys of $\mathbf{s}_1$ align with peaks and valleys, respectively, of $\mathbf{s}_2$.)

At every stage of the temporal alignment, DFW was applied to do the spectral alignment and to find the sequence of path indices ($\mathbf{a}_F$ and $\mathbf{b}_F$). The temporally aligned source spectrum vector information of speaker $\mathbf{s}_{1_{TA}}$ is transformed to spectrally match the target spectrum information of speaker $\mathbf{s}_2$ by creating a modified source spectrum information of speaker $\hat{\mathbf{s}}_1$ according to

$$\hat{s}_{1_{TF}}(a_F(i)) = s_{1_{TA}}(b_F(i)), i = 1, 2, \ldots, M. \tag{4.4}$$

This spectral alignment map drags spectral components of source blocks to produce transformed speech in the frequency domain. This data is inverse Fourier transformed and added in sequence to produce the transformed signal ($\hat{s}_{1_{warped}}$). After warping, filtering is performed to mitigate signal processing artifacts.

4.4.5 Results

Figure 4.4(a) shows a typical spectrogram for male speaker using the phrase “Author of the danger trail, Philip Steels, etc.” from the CMU ARCTIC database (Press ”Play” box to play the male sound Play). Figure 4.4(b) shows the spectrogram for female speaker for the same phrase mentioned above (Press ”Play” box to play the female sound Play). Figure 4.4(c) shows the voice transformation from male speaker to female speaker using same phrase (Press ”Play” box to play the transformed voice form male to female sounds Play). From Figure 4.4(a) and (b), we can see on a large scale we have some big motion, and on a fine scale we have small motion lines. These small lines are due to pitch and they are well defined lines but the warped signal in Figure 4.4(c) generally shows the distribution of the energy about right but the pitch lines are not that sharp and it is fazeier than the signals in Figure 4.4(a) and (b).

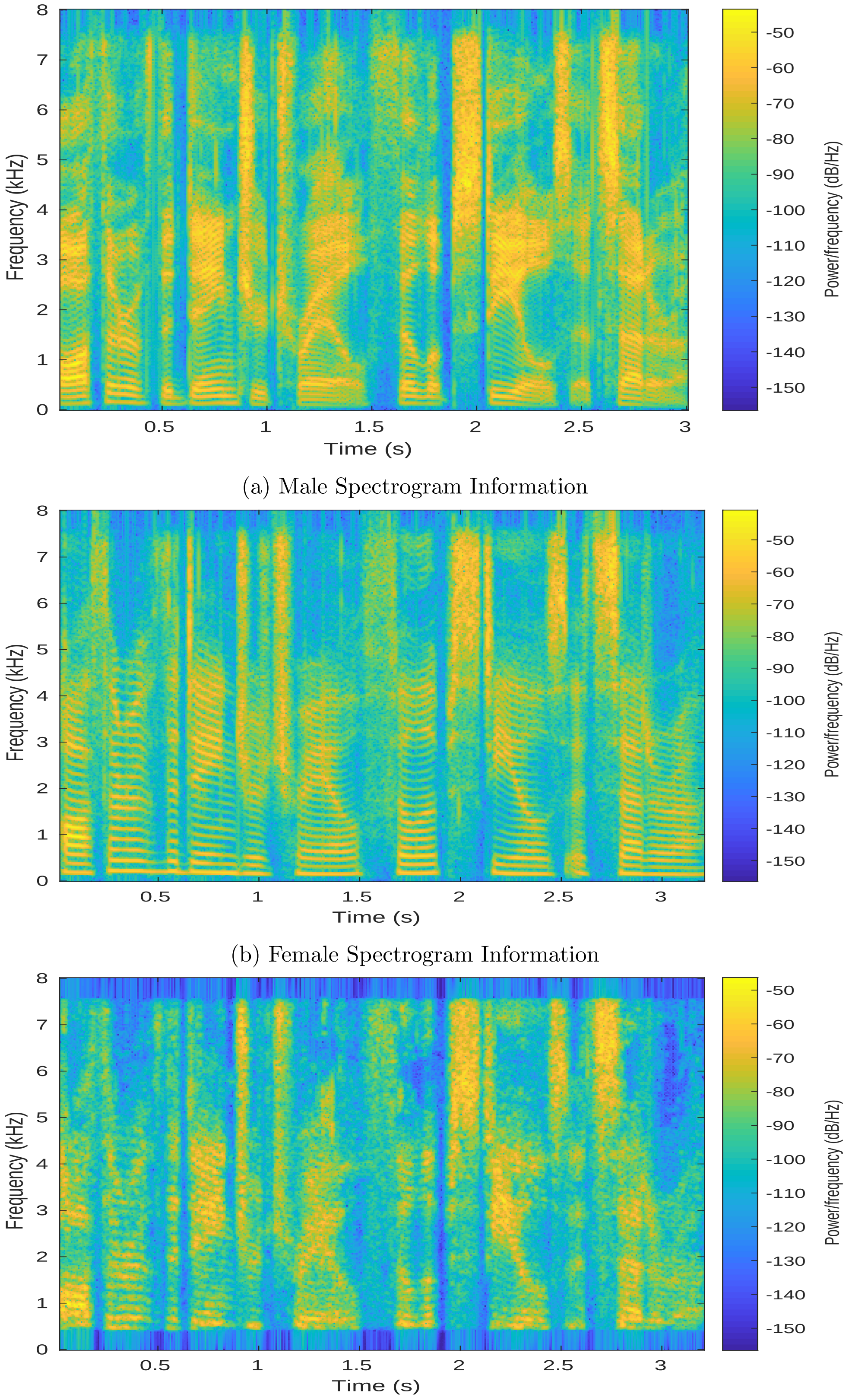
Figure 4.5(a) shows the spectral feature information for one segment of speech for female speaker. Figure 4.5(b) shows the spectral feature information for one time aligned segment (aligned with the segment shown in Figure 4.5(a)) of speech for male speaker. By applying the DFW, the peaks and valleys of the male spectrum are aligned to the locations of the peaks and valleys of the female spectrum, this process portrayed by Figure 4.5(c), which shows clearly the time aligned spectral segment for male speaker be spectrally aligned with the corresponding spectral segment shown in Figure 4.5(a). Figure 4.5(c) shows the way that the bins were connected within one segment (small green circles on the graph), and this may contribute to the distortion in the warped signal. Acoustically, the transformed signal, Figure 4.4(c), looks like female signal but the final sound has significant signal processing artifacts. Press ”Play” box to play the transformed sound Play.

4.5 Effectiveness of Phase Reconstruction on Warped Speech

The method described above for achieving speech transformation from one speaker's voice to another, which operates by moving speech magnitude information from a source speaker to a target speaker using a process involving dynamic warping in both the time domain and the frequency domain, involves only spectral magnitudes. This has been found to introduce significant deleterious signal processing artifacts. These are greatly ameliorated when phase information is reconstructed from the magnitude-only signals and improves the quality of the transformed speech. This process demonstrates the importance of phase in some speech processing tasks. This improvment flies in the face of convention, since conventional knowledge was considered phase information in speech is not significant to speech intelligibility.

Phase information in speech signal has usually been neglected in speech synthesis [72]. But with the increasing requirements of speech signal quality [73], phase effects should be considered [74, 75], as we have done here.

Many phase reconstruction algorithms have been proposed [76]. Those algorithms

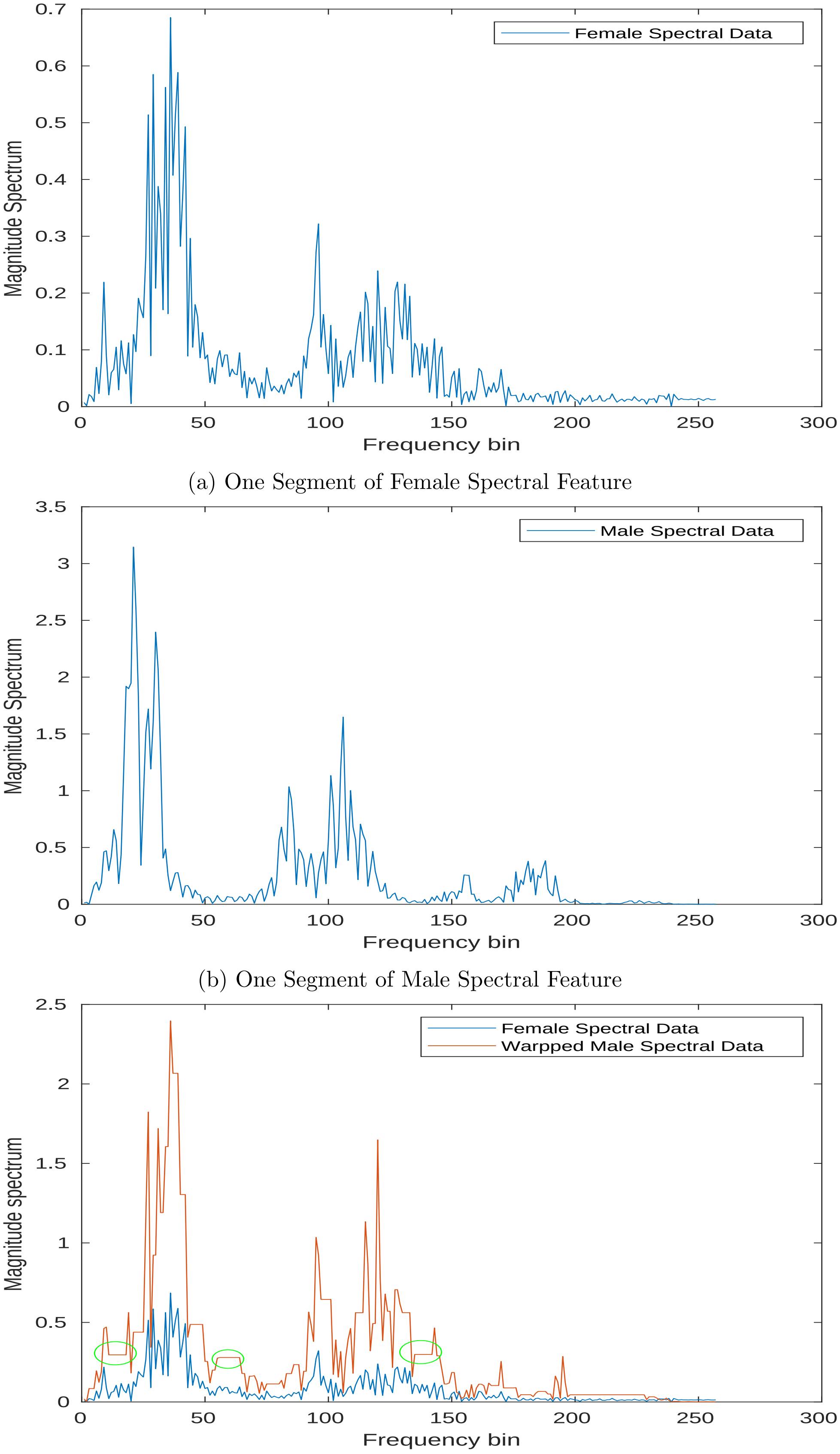


(a) Male Spectrogram Information

(b) Female Spectrogram Information

(c) Warped Male to Female

Fig. 4.4: Spectrogram for Male, Female and Warped Male to Female



(a) One Segment of Female Spectral Feature

(b) One Segment of Male Spectral Feature

(c) Warped Spectral Male with Female Spectral Fea- ture

Fig. 4.5: One Spectral Segment Feature for Male, Female and Warped Male to Female

basically can be divided into two groups. The first group of algorithms processes the entire signal offline. The second group works in (near) real-time [77]. In this work, the Griffin-Lim algorithm (GLA), [76], was applied to do phase reconstruction on an entire signal because it is based only on the consistency and does not take any prior knowledge about the target signal into account. This is described below.

4.5.1 Phase Reconstruction Process

GLA is arguably the most generic algorithm for phase reconstruction [77]. GLA phase reconstruction is computationally simple and is based on minimizing the squared error between STFT magnitude of $|\hat{S}1_{warped}(w)|$ and $|\hat{S}1_{TF}(w)|$ in each iteration.

Let $\hat{s}^i_{1_{warped}}$ refers to the estimated warped transformed signal $\hat{s}1_{warped}$ after ith iteration. The new estimate of $\hat{s}^{i+1}_{1_{warped}}$ at iteration $(i+1)$ is computed as follows:

- Find the STFT of the $\hat{s}^i_{1_{warped}}$ at iteration i, $(\hat{S}1_{warped}(w)^i)$.
- Find the magnitudes of $\hat{S}1_{warped}(w)^i$ and $\hat{S}1_{TF}(w)$.
- Replace the magnitude of the $\hat{S}1_{warped}(w)^i$ by the magnitude of the modified source spectrum $\hat{S}1_{TF}(w)$.
- Compute the new value of $\hat{S}1_{warped}(w)^i$ as

$$\hat{S}^i_{1_{warped}}(w)_{new} = |\hat{S}1_{TF}(w)| \times \exp(j\measuredangle \hat{S}^i_{1_{warped}}(w)). \tag{4.5}$$

- The updated signal is

$$\hat{s}^{i+1}_{1_{warped}} = \frac{H_w \times \mathscr{F}^{-1}(\hat{S}^i_{1_{warped}}(w)_{new})}{H_w^2} \tag{4.6}$$

  where H_w is the length of Hamming analysis window used in the STFT.

The algorithm is portrayed in Figure 4.6. It can be shown [76] that the algorithm in Figure 4.6 decreases the following distance measure between $|\hat{S}1_{warped}(w)|$ and $|\hat{S}1_{TF}(w)|$.

$$dist[\hat{s}_{1_{warped}}, |\hat{S}_{1_{TF}}(w)|] = \frac{1}{2\pi}\int_{w=-\pi}^{\pi}[|\hat{S}_{1_{warped}}(w)| - |\hat{S}_{1_{TF}}(w)|]^2 dw. \tag{4.7}$$

4.5.2 Experiment

The phase reconstruction experiment was carried out on CMU ARCTIC database. From that database, for test purposes we chose the following phrase: “Author of the danger trail, Philip Steels, etc.”, which is spoken by a US male and a US female.

Figure 4.7 shows a typical spectrogram for a male speaker (part (a)), also a typical spectrogram for a female speaker (part (b)). Figure 4.7(c) shows a typical spectrogram for a warped male speaker using magnitude only information without applying phase reconstruction, also Figure 4.7(d) shows the spectrogram for the warped speech with the GLA phase reconstruction (Press ”Play” box to play the transformed voice form male to female sounds using phase reconstruction algorithm $\boxed{\text{Play}}$). The pitch lines are somewhat stronger in the GLA phase reconstruction.

Figure 4.8(a) shows the spectral feature information for one time aligned segment of speech for a male speaker, and Figure 4.8(b) also shows the spectral feature information for one segment of speech for a female speaker. Figure 4.8(c), shows that the locations of peaks and valleys of male spectrum for that specific segment are aligned to the locations of the peaks and valleys of the female spectrum due to using DFW process of the two-level dynamic warping (inner process) without and with using phase reconstruction algorithm, respectively. The effects of using the phase reconstruction algorithm or not. Figure 4.8(c) shows the way that the bins were connected within one segment (small green circles on the graph). It is very clear that there is phase mismatch, and that phase mismatch contribute to the distortion in the warped signal. Figure 4.8(d) shows how the phase reconstruction algorithm takes care of the phase between bins within that specific segment (small green circles on the graph). The final sound of this transformation using GLA phase reconstruction is much better than if not using phase reconstruction algorithm. The whole process for the

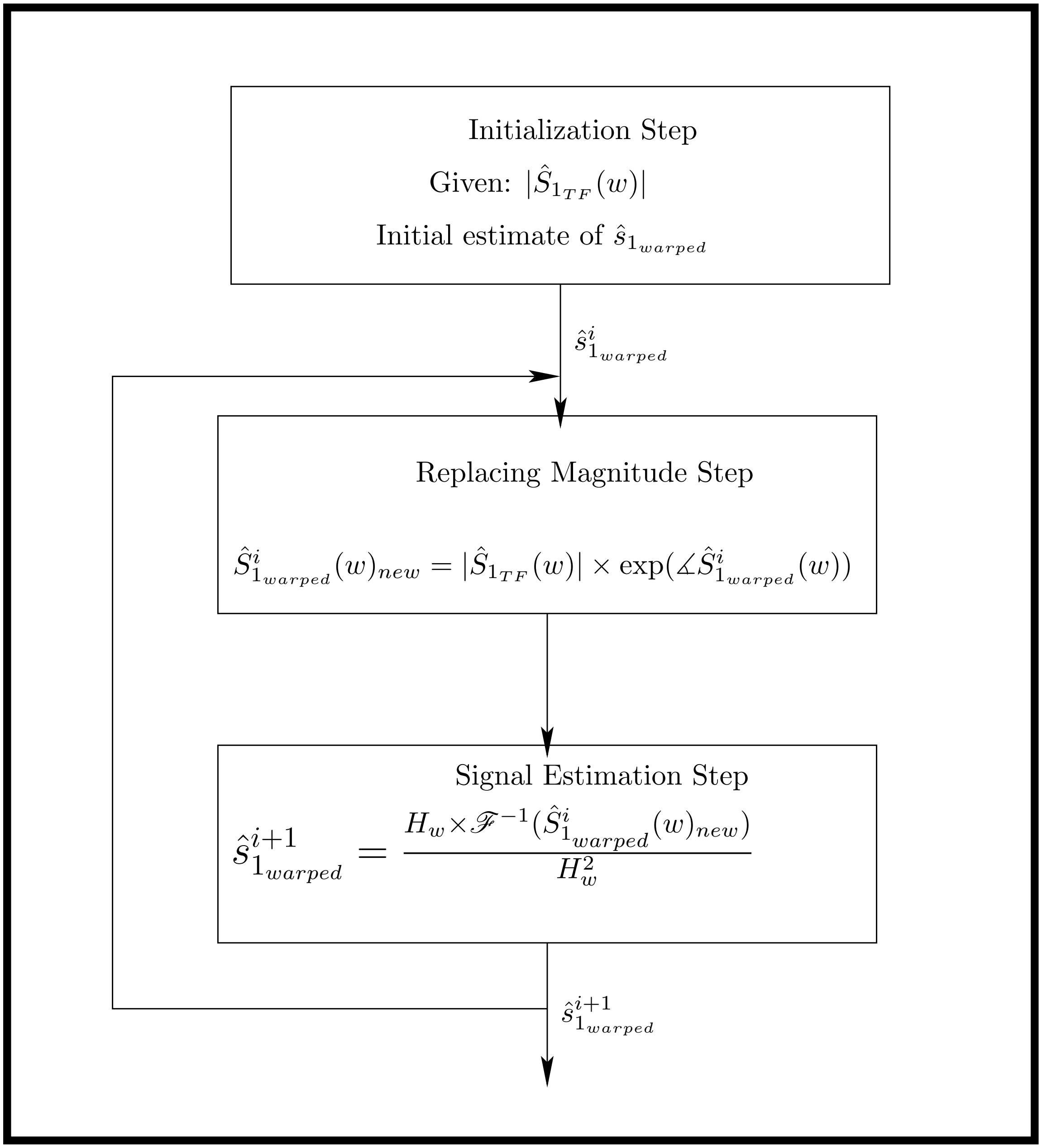
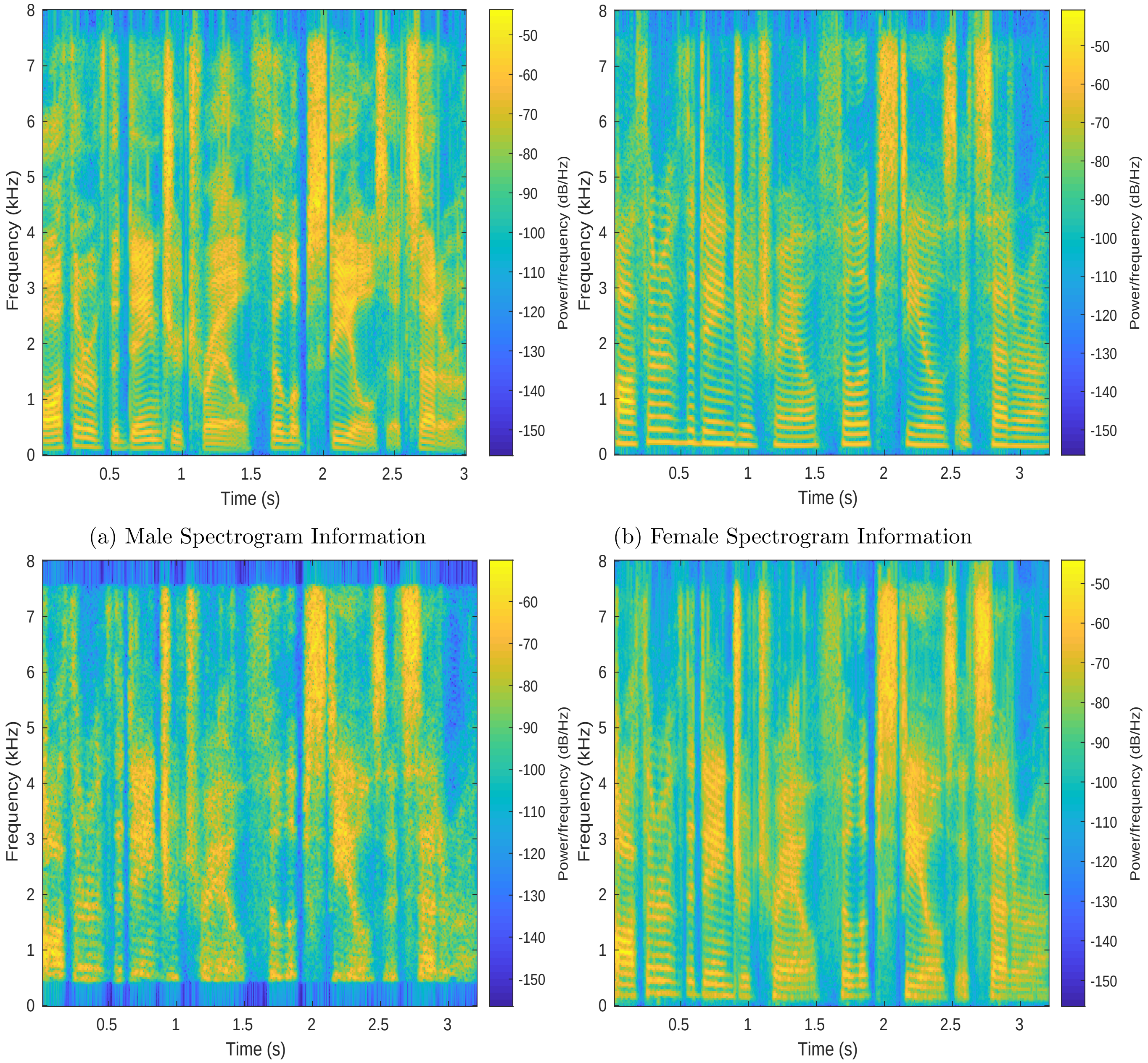

Initialization Step
Given: $|\hat{S}_{1_{TF}}(w)|$
Initial estimate of $\hat{s}_{1_{warped}}$
$\hat{s}^{i}_{1_{warped}}$
Replacing Magnitude Step
$\hat{S}^{i}_{1_{warped}}(w)_{new} = |\hat{S}_{1_{TF}}(w)| \times \exp(\measuredangle \hat{S}^{i}_{1_{warped}}(w))$
Signal Estimation Step
$\hat{s}^{i+1}_{1_{warped}} = \frac{H_w \times \mathscr{F}^{-1}(\hat{S}^{i}_{1_{warped}}(w)_{new})}{H_w^2}$
$\hat{s}^{i+1}_{1_{warped}}$

Fig. 4.6: Griffin-Lim (GLA) Algorithm



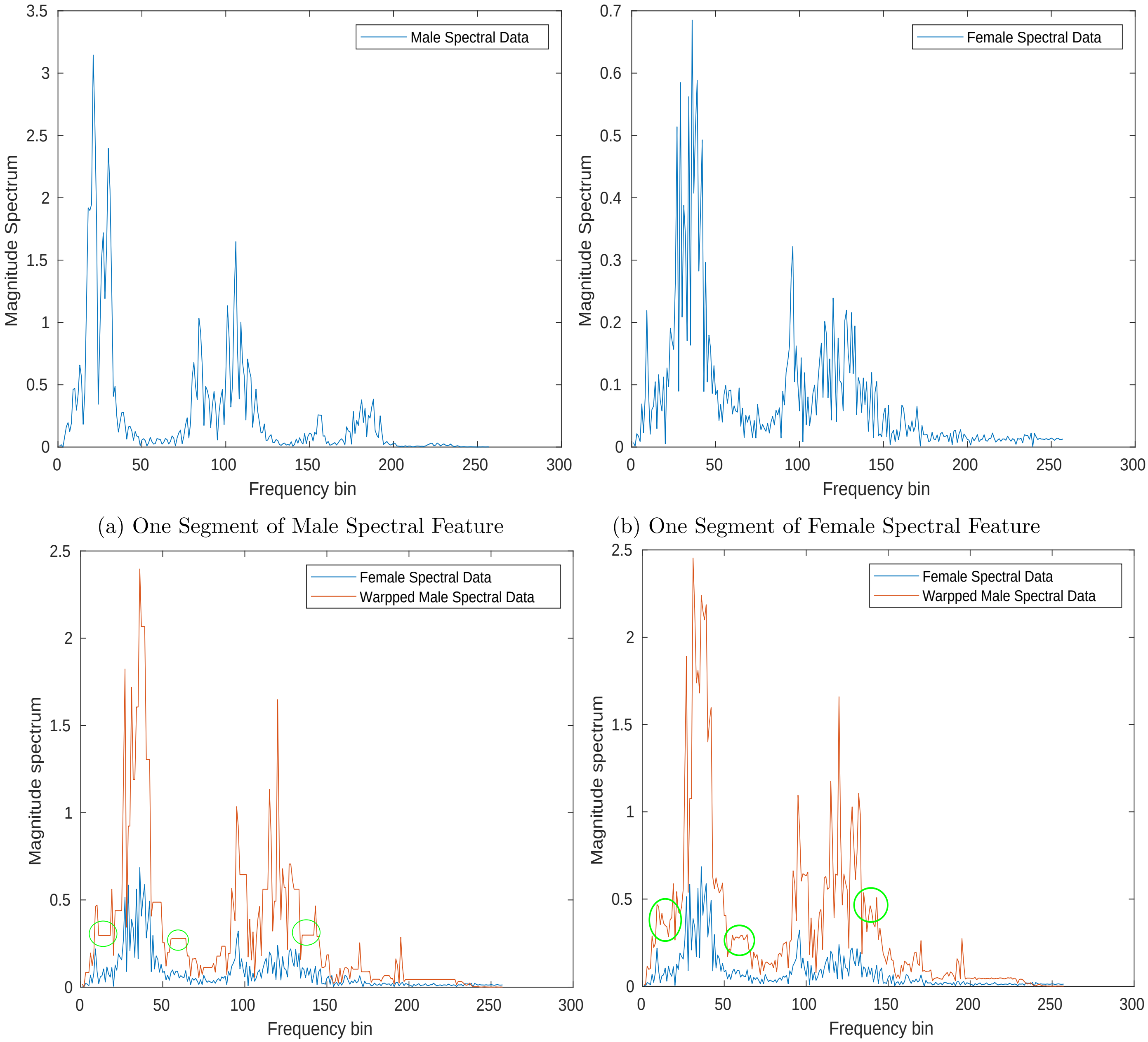
(a) Male Spectrogram Information

(b) Female Spectrogram Information

(c) Warped Male to Female no GLA

(d) Warped Male to Female with GLA

Fig. 4.7: Spectrogram Information for Male, Female and Warped Male to Female with and without GLA



(a) One Segment of Male Spectral Feature

(b) One Segment of Female Spectral Feature

(c) Warped Spectral Male with Female Spectral, no GLA

(d) Warped Spectral Male with Female Spectral, with GLA

Fig. 4.8: One Spectral Segment Feature for Male, Female and Warped Male to Female with and without GLA

speech transformation using two-level DW with the phase reconstruction is shown in Figure 4.9 and Algorithm 4.2.

4.6 Spectral Warping using ANN

Speech conversion has become an area of high interest in speech signal processing. Most of the traditional voice conversion techniques assume availability of parallel training data. In other words, the mapping function is computed on paired utterances of the same linguistic content spoken by source and target speaker [78]. The speech transformation approach described above is accomplished using a two-level dynamic warp (DW) when source speaker and target speaker are available which avoids the need to find acoustic parameters. But if the target speech is already available saying the desired target sentence, why bother with the transformation? A more challenging, but realistic, setting is when the target is not available saying the desired statement. Thus, a second phase of this transformation approach is to train an Artificial Neural Network (ANN) to produce the spectral warping function from only the source speaker information, based upon which the source speech may be warped to the target speech.

Machine learning (ML) is the technique of using statistical models for computers to learn certain tasks without explicitly giving the computer instructions on how to perform the task. Because of the success of the neural techniques in the field of machine learning and it is a fast growing research area, many researchers tried to apply and used some of the techniques to the speech transformation application, these networks are exceptionally good at learning of non-linear models [67, 79].

4.7 Artificial Neural Network

An Artificial neural network (ANN) is a model that tries to mimic the behavior of the human brain. The term neural network originates from as far back as the 1940's and was a first attempt to describe the human brain in a mathematical way [80, 81]. An ANN consists of many interconnected processing nodes that computes responses based on inputs. Each node represents the model of an artificial neuron, and there is an associated weight for each

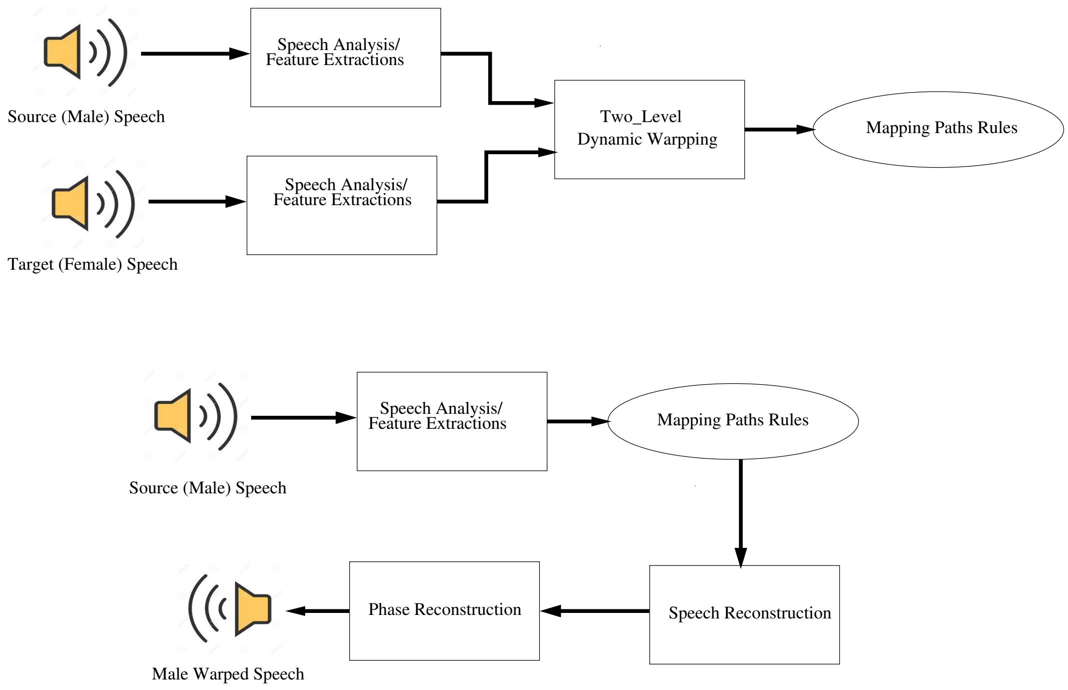


Fig. 4.9: Block Diagram of Voice Transformation using two_level DW

Algorithm 4.2 Whole Voice Transformation Process using two-level DW

```
Input:
        Male speech information, $source\_speech$
        Female speech information, $target\_speech$
Output:
        Warped speech
Begin
        Function: Identifying Starting and ending points
                  $source\_speech\_clip = identify\_start\_end\_function(source\_speech)$
                  $target\_speech\_clip = identify\_start\_end\_function(target\_speech)$
        Function: Spectral feature extraction
                  $\mathcal{S}_1 = spect\_function(source\_speech\_clip)$
                  $\mathcal{S}_2 = spect\_function(target\_speech\_clip)$
        Function: Algorithm 4.1
                  $[(\mathbf{a}_T, \mathbf{b}_T), (\mathbf{a}_F, \mathbf{b}_F)] = dtw\_algorithm(\mathcal{S}_2 = target\_spect, \mathcal{S}_= source\_spect)$
        Function: Time Alignment step
                  $\mathcal{S}_{1_{TA}}(:, \mathbf{a}_T) = \mathcal{S}_1(:, \mathbf{b}_T)$
        Function: Spectral Alignment step
                  For $i_F = 1$ to $size((\mathcal{S}_{1_{TA}}), 2)$
                     $\mathcal{S}_{1_{TF}}(i, \mathbf{a}_F) = \mathcal{S}_{1_{TA}}(i, \mathbf{b}_F)$
                  End $i_F$
        Function: Phase Reconstruction Step
                  $warped\_speech = phase\_algorithm(\mathcal{S}_{1_{TF}})$
        Play $warped\_speech$
```

interconnection between the neuron.

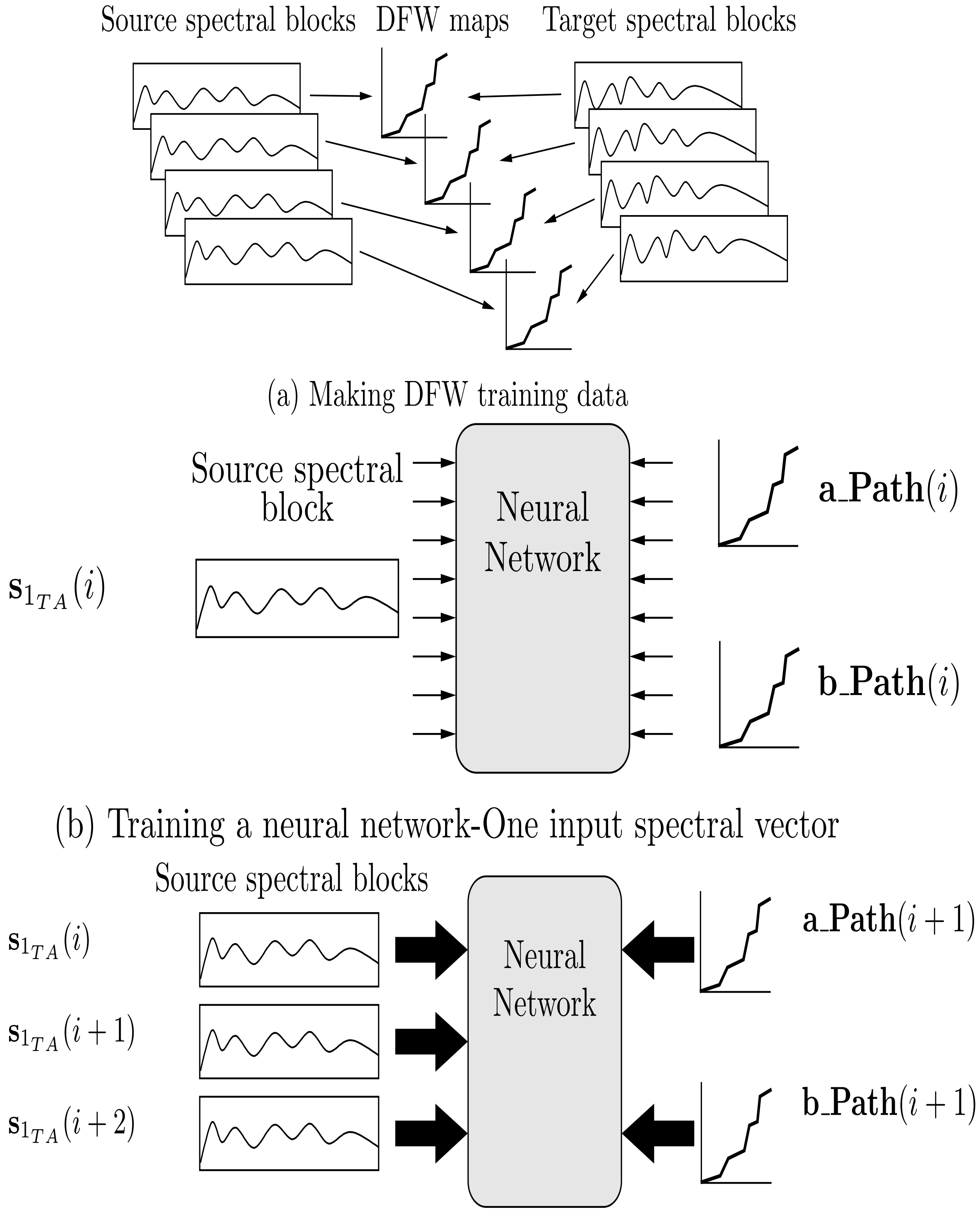
In the last few years, ANN models with different topologies and architectures have been used to solve a variety of tasks, like in language modeling, text-to-speech synthesis, also have perform different pattern recognition tasks.

4.8 Proposed Method of Spectral Transformation using ANN

In this work, a multi-layer feedforward neural network is used to obtain the mapping function between the input and the output vectors. The two-level dynamic warping procedure described above was used to obtain training data to train an artificial neural network. The ANN input data was spectral feature vectors from the source speaker. The ANN output data was spectral warping path information $\mathbf{a}_F$ and $\mathbf{b}_F$, which can be used to do the spectral warping function.

This experiment was carried out on CMU-ARCTIC database, the first $(600 - 1000)$ phrases recorded by a US male and the corresponding $(600 - 1000)$ phrases recorded by a US female are chosen to achieve the voice transformation form male (source) speaker to female (target) speaker and to train the ANN network. Speech data was sampled at 16000 samples/sec. These feature vectors that were used in the two-level DW were considered to be the positive-frequency spectral information in the frequency domain that are extracted using the FFT. Each sentence was temporally segmented into 32-ms segments, using a Hamming window with 16-ms overlapping, zero-padded, then transformed using a 512-point FFT. The K = 256 positive frequency spectral elements in the frequency domain were used as a feature vector.

Training data is obtained as portrayed in Figure 4.10, as follows: Source and target spectral information are temporally aligned using DTW. The source spectral features are temporally aligned with respect to the target spectral features, these time aligned spectral features are used as an input to the ANN. The warping paths $\mathbf{a}_F$ and $\mathbf{b}_F$ computed by the two-level DW when it calls the inner spectral warping are saved as training data, this procedure produces a pool of training data as shown in 4.10. The ANN is trained to map a sequence of time aligned source speaker's spectral feature information to the spectral warping



(a) Making DFW training data

(b) Training a neural network-One input spectral vector

(c) Training a neural network-Three input spectral vectors

Fig. 4.10: Making DFW training data, and using this to train a neural network

path information (Once the training is complete, we get a weight matrix that represents the mapping function between the source and the target speaker spectral). A generalized back propagation learning is applied to adjust the weights matrix of the desired network such as to minimize the mean square error between the calculated and actual spectral warping paths ($\mathbf{a}_F$ and $\mathbf{b}_F$).

4.8.1 Interpolating the data to achieve constant length

Given a magnitude spectrum as the input (as discussed above), the length of the warping paths $\mathbf{a}_F$ and $\mathbf{b}_F$ may be different from one segment of speech to another. Since the length of the warping paths $\mathbf{a}_F$ and $\mathbf{b}_F$ may vary form phrase to phrase, this poses a problem for the neural network. In order to have the neural network be able to have a constant output length (without zero-padding the output data, which would introduce neural network artifacts) the $\mathbf{a}_F$ and $\mathbf{b}_F$ information are interpolated to produce a modified $\mathbf{a}_F$ and $\mathbf{b}_F$ which is the same (maximum) length for all segments of the speech vector. In our experiments the maximum length of the warping paths was 482. This interpolation is computed as described here using an example, which employs Matlab-liked notation and functions.

- Let $\mathbf{a} = (1, 2, 3, 4, 4, 5)$ (length of $\mathbf{a}$ is 6), which has the indices of $[1, 2, 3, 4, 5, 6]$. (That is, 1-based indexing.) Let the new length is $\max_{\texttt{index}} = 10$. ($\max_{\texttt{index}}$ is the maximum length of vector $\mathbf{a}$ in the training data set.)
- Create a new set of interpolated indices (generally with non-integer values) as

$$\tilde{\mathbf{n}} = [1 : (\max_{\texttt{index}})] \times \frac{\texttt{length}(\mathbf{a})}{\max_{\texttt{index}}}$$

  In the example, the new indices are $\tilde{\mathbf{n}} = [1, 1.2, 1.8, 2.4, 3, 3.6, 4.2, 4.8, 5.4, 6]$. The new indices still fall within the range of 1 to 6, but there are now $\max_{\texttt{index}}$ index values.
- Interpolate the $\mathbf{a}$ values using the interpolated indices as

$$\tilde{\mathbf{a}} = \texttt{interpolate}(\mathbf{a}, [1 : \texttt{length}(\mathbf{a})], \tilde{\mathbf{n}})$$

In this example the new $\tilde{\mathbf{a}}$ values are $\tilde{\mathbf{a}} = (1, 1.5556, 2.1111, 2.6667, 3.2222, 3.7778, 4.000, 4.000, 4.4444, 5)$.

The above process can be repeated for function path $\mathbf{b}$ to find $\tilde{\mathbf{b}}$. Figure 4.11 illustrates normal function paths for $\mathbf{a}$ and $\mathbf{b}$ with interpolated paths. ($\tilde{\mathbf{a}}$ and $\tilde{\mathbf{b}}$ respectively.)

In order to get the original length of the warping paths $\mathbf{a}$ and $\mathbf{b}$, reverse interpolation was applied to the estimated $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$. This reversing was applied as described here using an example.

- Let $\tilde{\mathbf{a}} = (1, 1.5556, 2.1111, 2.6667, 3.2222, 3.7778, 4.000, 4.000, 4.4444, 5)$ (length of $\tilde{\mathbf{a}}$ is 10). Suppose the original length of $\mathbf{a}$ is $\texttt{length}(\mathbf{a}) = 6$ (as in the previous example).
- Create a new set of reverse interpolated indices as

$$\tilde{\mathbf{n}}_{revs} = [1 : \texttt{length}(\mathbf{a})] \times \frac{\texttt{length}(\tilde{\mathbf{a}})}{\texttt{length}(\mathbf{a})}$$

  In the example, the new reversing indices are $\tilde{\mathbf{n}}_{revs} = [1, 3.3333, 5, 6.6667, 8.3333, 10]$.
- Reverse interpolation of the $\tilde{\mathbf{a}}$ values using reverse interpolated indices are

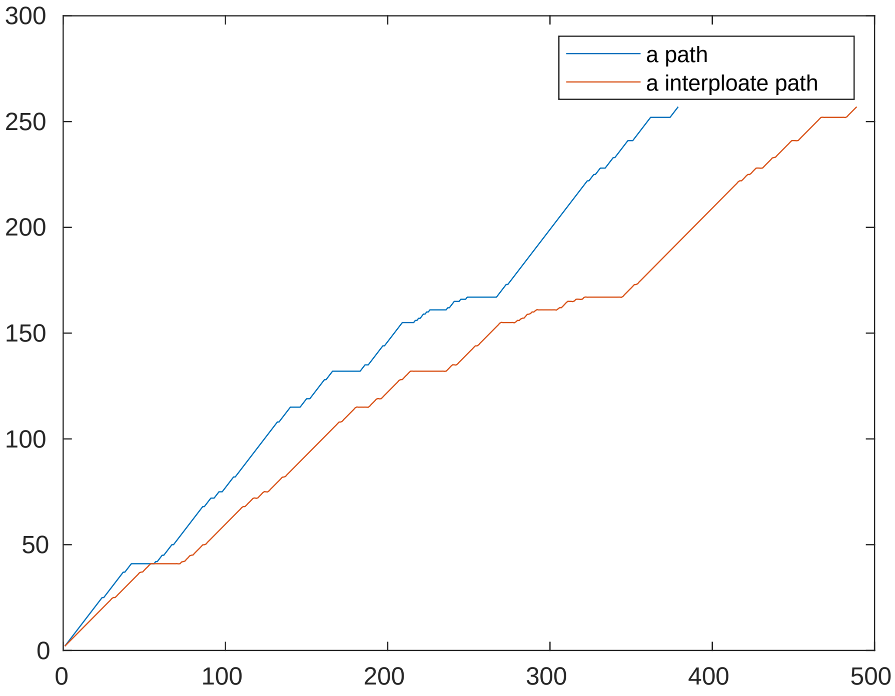
$$\tilde{\mathbf{a}}_{revs} = \texttt{interpolate}(\tilde{\mathbf{a}}, [1 : \texttt{length}(\tilde{\mathbf{a}})], \tilde{\mathbf{n}}_{revs})$$

  In this example the new $\tilde{\mathbf{a}}_{revs}$ values are $\tilde{\mathbf{a}}_{revs} = (1, 2, 3, 3.8667, 4.1333, 5)$.

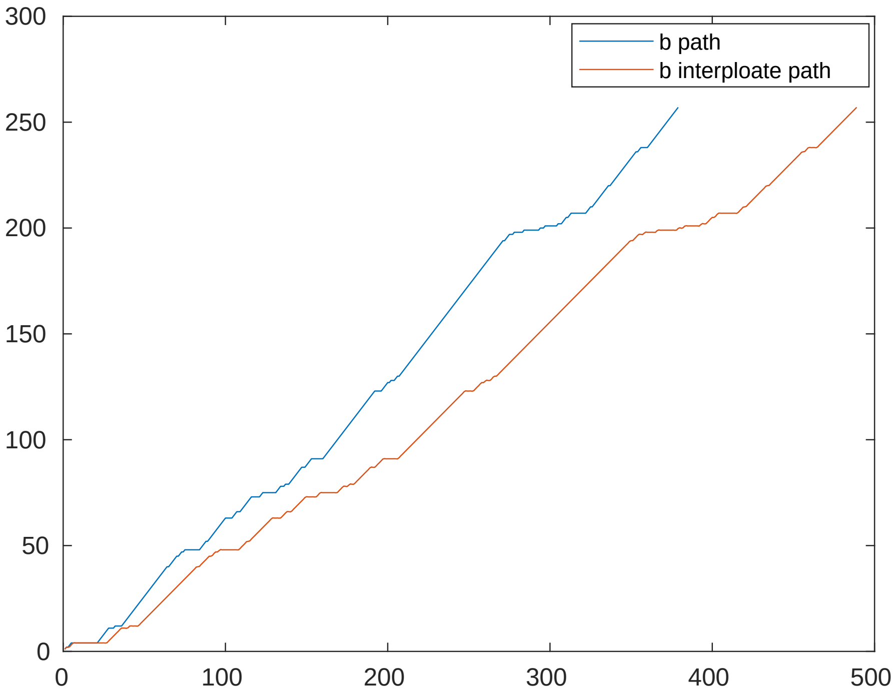
The above process can be repeated for function path $\tilde{\mathbf{b}}$ to find $\tilde{\mathbf{b}}_{revs}$. The non integer index values of $\tilde{\mathbf{a}}_{revs}$ and $\tilde{\mathbf{b}}_{revs}$ are rounding to get integer values.

4.9 Mel-Cepstral Distortion as an Objective Measure

In this work, we use MCD to evaluate the quality of the transformed speech, and to be able to compare our work to other work in the literature. Mel-Cepstral Distortion (MCD) has been used as an objective error measure for evaluating the quality of synthetic voice [59]. It is a measure of the difference between two sequences of mel-cepstra. MCD is related to filter characteristics and hence is an important measure to check the performance of



(a) **a** and $\tilde{\mathbf{a}}$



(b) **b** and $\tilde{\mathbf{b}}$

Fig. 4.11: Normal and interpolated paths

mapping obtained by ANN network. MCD is essentially a weighted Euclidean distance, that is defined by:

$$MCD = (10/ln10)\sqrt{2\sum_{i=1}^{25}(mc_i^{(t)} - mc_i^{(w)})^2} \tag{4.8}$$

where $mc_i^{(t)}$ is the ith mel-cepstral of the a frame of target speech,and $mc_i^{(w)}$ is the ith mel-cepstral of the corresponding frame in the warped speech signal.

4.10 ANN Architecture and Experiment

The experiment, of training the ANN to estimate the warping paths $\mathbf{a}_F$ and $\mathbf{b}_F$, was conducted in four phases. Phase one was conducted to validate the general question, can a neural network learn the transformation from the speaker warping paths. It is thus a general proof of concept. Phase two was conducted on a big data to achieve voice transformation. Phase three was conducted using clustering method to improve the quality of the transformation. Phase four was conducted to see who the convolutional neural network works with this application. The important task in building the ANN based voice conversion system is to find an optimal architecture for ANN. To experiment with different ANN architectures we considered, as mentioned in section 4.6.2, the source male time aligned spectral features are used as an input to the ANN and the warping paths $\mathbf{a}_F$ and $\mathbf{b}_F$ are used as an output to the ANN in two main architectures:

- **First Architecture:** One-Input/One-Output ($1/IP - 1/OP$). As shown in Figure 4.12(a), we considered one source time aligned spectral features vector ($\mathbf{s}_{1_{TA}}(:, i)$, $i = 1, \ldots,$ number of all source time aligned vectors) of length 256 as an input and the corresponding warping paths $(\mathbf{a_F}(i), \mathbf{b_F}(i))_{actual}$ as a training data to the ANN at each iteration of training. Each warping path has length equal to 482, where the total length of the output vector used in this experiment is 964. The length 482 was calculated using the interpolation way as shown in section 6.4.2. This ANN (1/IP-1/OP) architecture shown in Figure 4.12(a) was used to capture the transformation

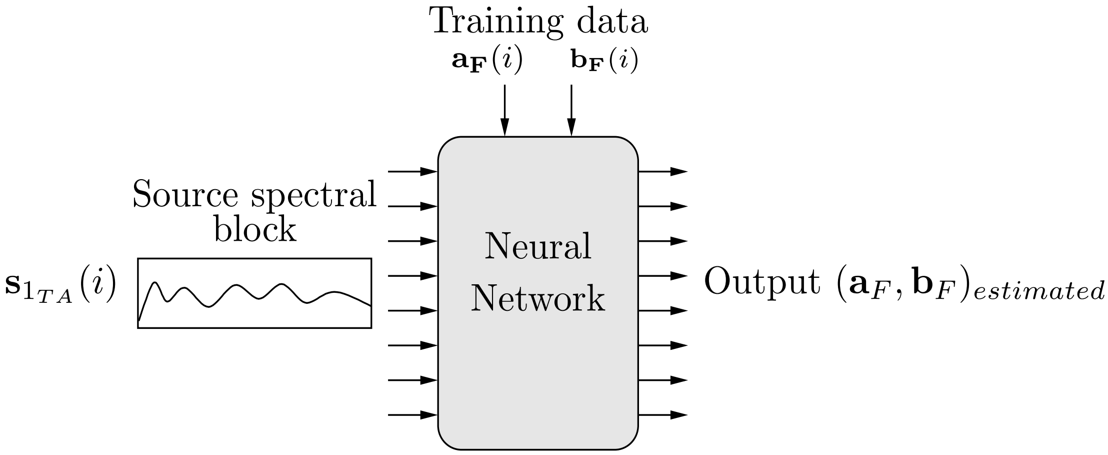


Fig. 4.12: Block diagram of ANN of $1/IP - 1/OP$ architecture

function for mapping the source time aligned vectors onto the corresponding warping paths. Once the training is complete, we get a weight matrix which can be used to estimate the corresponding wrapping paths $(\mathbf{a_F}(i), \mathbf{b_F}(i))_{estimated}$.

- **Second Architecture:** Three-Input/One-Output $(3/IP - 1/OP)$. As shown in Figure 4.13(b), instead of using one one source time aligned spectral features vector of length 256 as an input to the ANN, we considered three consecutive time aligned spectral vectors ($\mathbf{s}_{1_{TA}}(:, i)$, $\mathbf{s}_{1_{TA}}(:, i+1)$, and $\mathbf{s}_{1_{TA}}(:, i+2)$, where $i = 1, \ldots,$ number of all source time aligned vectors) of the source information of length 768. Also, the warping paths corresponding with $(i+1)^{th}$ source time aligned spectral features vector. This ANN 3/IP-1/OP architecture shown in Figure 4.13(b) used to capture the transformation function for mapping the source time aligned vectors onto corresponding warping paths. Once the training is complete, we get a weight matrix which can be used to estimate the corresponding wrapping paths $(\mathbf{a_F}(i+1), \mathbf{b_F}(i+1))_{estimated}$. The reason behind choosing this kind of architecture is that having more contexts helps understand how it can warp, and then explore the evidence that we have got to support our hypothesis.

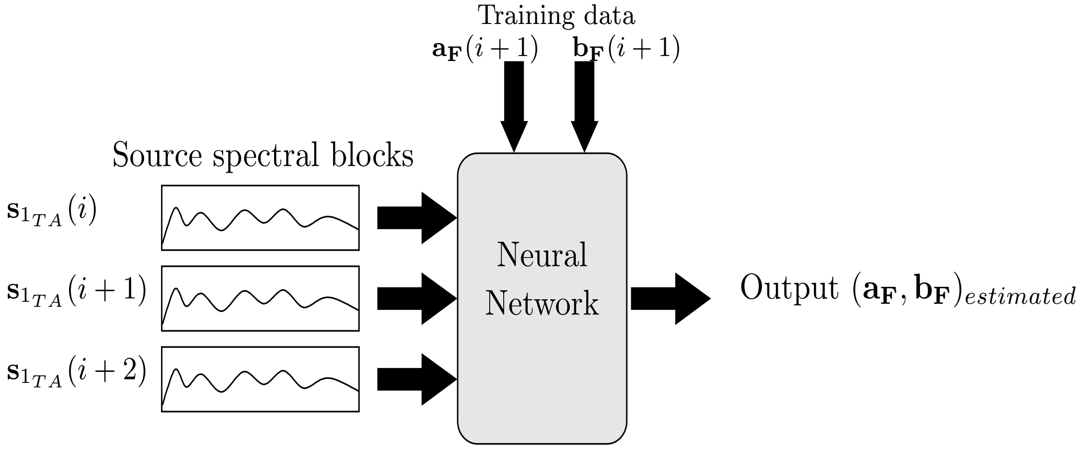


Fig. 4.13: Block diagram of ANN of $3/IP - 1/OP$ architecture

4.10.1 Phase One

In this phase, only one phrase was selected to be trained using the two architectures mentioned above to achieve voice transformation. This phase was contacted to decide which architecture for ANN that achieve good quality for the process of voice transformation to continue with it to the next phases. It also validates the general question, can a neural network learn the transformation from the speaker warping paths. It is thus a general proof of concept.

The selected phrase, "Author of the danger trail, Philip Steels, etc." was selected for the CMU-ARCTIC database. After identifying the starting point and ending point of the selected phrase, feature information are extracted first from the FFT for each 32-ms segment of speech with 50% overlap windowed using a Hamming window. Temporal alignment (DTW) was done on the source spectral feature vectors with respect to the target female signals. DFW was applied to do the spectral alignment and to find the sequence of path indices (a_F and b_F), the process of doing two-level DW was done using Matlab program. After that the spectral data was arranged in a suitable shape using Matlab, training was performed by Tensorflow and Keras using the Python language to estimate the warping paths $\mathbf{a}_F$ and $\mathbf{b}_F$.

To get an optimal architecture, we have experimented both architectures on 5-layer, 6-layer, 7-layer, 8-layer, 9-layer, 10-layer, 20-layer and 21-layer networks. The architectures are provided with number of nodes in each layer and the output function used for that layer in Table 4.1. For instance, 256L 500N 900N 1000N 964L means that it's a 5-layer network with 256 inputs, 964 output nodes with 500, 900, and 1000 nodes in the hidden layers. L represents "linear" output function and N represents "ReLU" output function (activation function). From the fields of phase one in table 4.1, we see that the nineteen layers 3/IP-1/OP architecture provides better MCD results when compared with others. Hence, for all the remaining voice transformation experiments reported in the next phases, the nineteen layers 3/IP-1/OP architecture.

Figure 4.14 shows a typical spectrogram information for a male speaker (part (a)), a typical spectrogram information for a female speaker (part (b)), a typical spectrogram information for a warped male speaker using ANN of nineteen layers and 1/Ip-1/OP architecture with a phase reconstruction algorithm (part (c)), and a typical spectrogram information for a warped male speaker using ANN of the selected number of layers and 3/IP-1/OP architecture with a phase reconstruction algorithm (part (d)). Acoustically, the warped signal in part (d) is clearly looks like the female signal more than the result in part (c) and looks like the result in Figure 4.7(d). Also, the pitch lines in part (d) is stronger than the pitch lines in part (c). The final sound of that transformation in part (d) (Play) is much better than the sound in part (c) (Play).

Figures 4.15 and 4.16 show how the ANN has learned a warping path (the **a** warping data) at various segments of the selected speech signal for different training iterations. Figure 4.15(b), which produced using the nineteen layers of the ANN 3/IP-1/OP architecture, shows more matching between the DW path and the NN-learned path than Figure 4.15(a) which produced using the nineteen layers of the ANN 1/IP-1/OP architecture. The DW path is shown in blue and the NN-learned path is in orange. As Figure 4.15(b) shows the network has learned the data so well that the training data (blue) is indistinguishable form the NN-learned data. Figures 4.16 shows how the ANN has learned a **b** warping path at various segments of the selected speech signal after 5000 training iterations. Again excellent match is achieved with NN-3/IP-1/Op architecture (Figure 4.15(b)). As before, the DW path is shown in blue and the NN-learned path is in orange. Figure 4.17 shows the mean square error between the true values of the warping paths ($\mathbf{a}_F$ and $\mathbf{b}_F$), obtained form inner DFW, and the estimated warping paths ($\tilde{\mathbf{a}}_{F,estimated,revs}$ and $\tilde{\mathbf{b}}_{F,estimated,revs}$) obtained form training the ANN for the selected architecture (3/IP-1/OP). The training takes 5000 training iterations to achieve this performance.

According to the result shown in Figures 4.14(d), 4.15(d) and 4.16(d). we decide that the voice transformation reported in the next phases of this experiment should be based on the nineteen layers 3/IP-1/OP architecture.

Phase No.	**No. of Layers**	**Architecture Type**	**ANN Architecture**	**MCD[dB]**
Phase One	5	1/IP-1/OP	256L 500N 900N 1000N 964L	20.2
	5	1/IP-1/OP	256L 600N 1200N 1000N 964L	20.5
	6	1/IP-1/OP	256L 600N 1000N 1500N 1000N 964L	18.01
	7	1/IP-1/OP	256L 600N 1000N 1500N 2000N 1500N 964L	18.002
	19	1/IP-1/OP	500L 1500N 2500N 3500N 4500N 5040N 6459N 7459N 8459N 9459N 10459N 8459N 7939N 6939N 5500N 4500N 3500N 2500N 964L	3.3
	20	1/IP-1/OP	500L 1500N 2500N 3500N 4500N 5040N 6459N 7459N 8459N 9459N 10459N 9459N 8459N 7939N 6939N 5500N 4500N	5.5

Continued on next page

Phase No.	No. of Layers	Architecture Type	ANN Architecture	MCD[dB]
Phase One			3500N 2500N 964L	
	5	3/IP-1/OP	768L 4500N 6500N 6939N 964L	16.85
	6	3/IP-1/OP	768L 4500N 6500N 7040N 6939N 964L	15.3
	7	3/IP-1/OP	768L 4500N 6500N 8040N 8939N 6939N 964L	12.92
	8	3/IP-1/OP	768L 3500N 4500N 5040N 6459N 5939N 4939N 964L	8.8
	8	3/IP-1/OP	768L 4500N 5500N 6040N 7459N 6939N 4939N 964L	9.2
	8	3/IP-1/OP	768L 4500N 6500N 7040N 9459N 7939N 6939N 964L	7
	8	3/IP-1/OP	768L 3000N 5500N 6040N 7459N 6939N 5939N 964L	6.8
	9	3/IP-1/OP	768L 4500N 6500N 7040N 9459N 8939N 7939N 6939N 964L	7.5
	10	3/IP-1/OP	768L 4500N 6500N 7040N 9459N 10459N 8939N 7939N 6939N 964L	4.7
	19	3/IP-1/OP	1500L 2500N 3500N 4500N 6500N 7040N 8459N 9459N 10459N 10459N 9459N 8459N 7939N 6939N 5500N 4500N 3500N 2500N 964L	2.55
	20	3/IP-1/OP	1500L 2500N 3500N 4500N 6500N 7040N 8459N 9459N	3.5

Continued on next page

Phase No.	No. of Layers	Architecture Type	ANN Architecture	MCD[dB]
			10459N 11459N 10459N 9459N 8459N 7939N 6939N 5500N 4500N 3500N 2500N 964L	
Phase Two	19	1/IP-1/OP	500L 1500N 2500N 3500N 4500N 5040N 6459N 7459N 8459N 9459N 10459N 8459N 7939N 6939N 5500N 4500N 3500N 2500N 964L	5.4
	19	3/IP-1/OP	1500L 2500N 3500N 4500N 6500N 7040N 8459N 9459N 10459N 10459N 9459N 8459N 7939N 6939N 5500N 4500N 3500N 2500N 964L	4.8
Phase Three	Six-Clusters.Each Cluster trained for 19 layers	1/IP-1/OP	500L 1500N 2500N 3500N 4500N 5040N 6459N 7459N 8459N 9459N 10459N 8459N 7939N 6939N 5500N 4500N 3500N 2500N 964L	3.3
		3/IP-1/OP	1500L 2500N 3500N 4500N 6500N 7040N 8459N 9459N 10459N 10459N 9459N 8459N 7939N 6939N 5500N 4500N 3500N 2500N 964L	2.8
Phase Four	7	1/IP-1/OP	256L 64N(k=3) 100N(k=3) 150N(k=3) 100N(k=3) FL D100N	13.33

Continued on next page

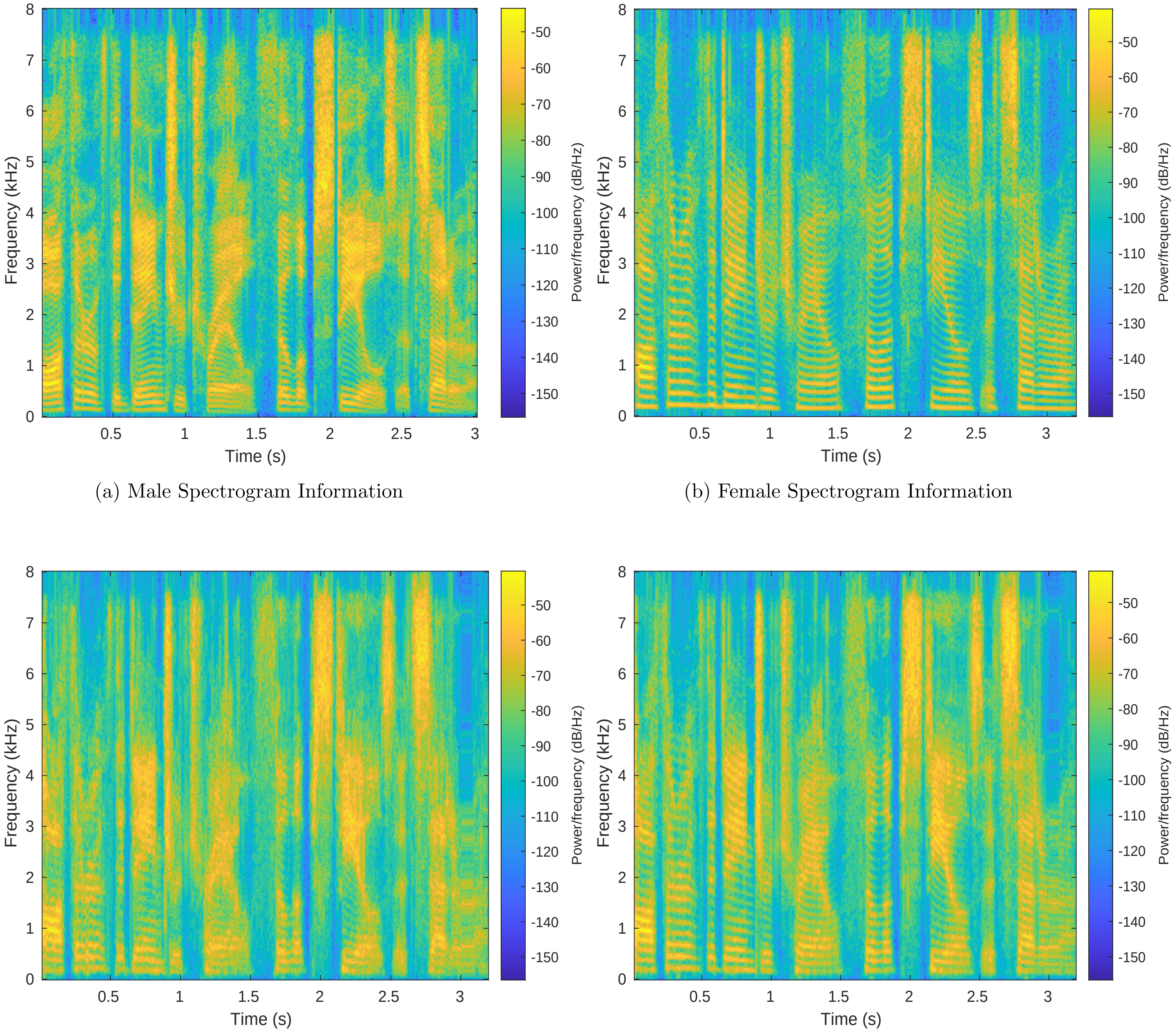
Phase No.	No. of Layers	Architecture Type	ANN Architecture	MCD[dB]
			964L	
	5	3/IP-1/OP	768L 64N(k=3) 100N(k=3) FL D100N 964L	13.5

Table 4.1: MCD's Obtained for Different Architectures using on

4.10.2 Phase Two:

In this phase, we considered 600 phrases for training and a separate set of 120 for testing to train the selected NN architecture to achieve general voice transformation, that is not learning and achieving voice conversion on one selected phrase as we did in phase one. As described in Section 4.6.2, the selected 600 phrases for each speaker were recorded by a US male and US Female and they are selected form the CMU-ARCTIC database. After identifying the starting point and ending point of the selected phrases, feature information are extracted first from the FFT for each 32-ms segment of speech with 50% overlap windowed using a Hamming window. Temporal alignment (DTW) was done on the source spectral feature vectors with respect to the target female signals. DFW was applied to do the spectral alignment and to find the sequence of path indices ($\mathbf{a}_F$ and $\mathbf{b}_F$) for each segment, the process of doing two-level DW was done using Matlab program. After that the data was arranged in a suitable shape, using Matlab, and trained with the selected NN architecture for 10000 iterations with Tensorflow and Keras using Python language and tested for the phrase chosen in phase one (this phrase was excluded form the training set and testing set).

Figure 4.18 shows the spectrogram information for a male speaker (part (a)), a typical spectrogram information for a female speaker (part (b)), a typical spectrogram information for a warped male speaker using ANN of the selected architecture (nineteen layers and 3/IP-1/OP architecture) for the same phrase used in phase one after the network reached the 5000 iterations with a phase reconstruction algorithm (part (c)), and the spectrogram



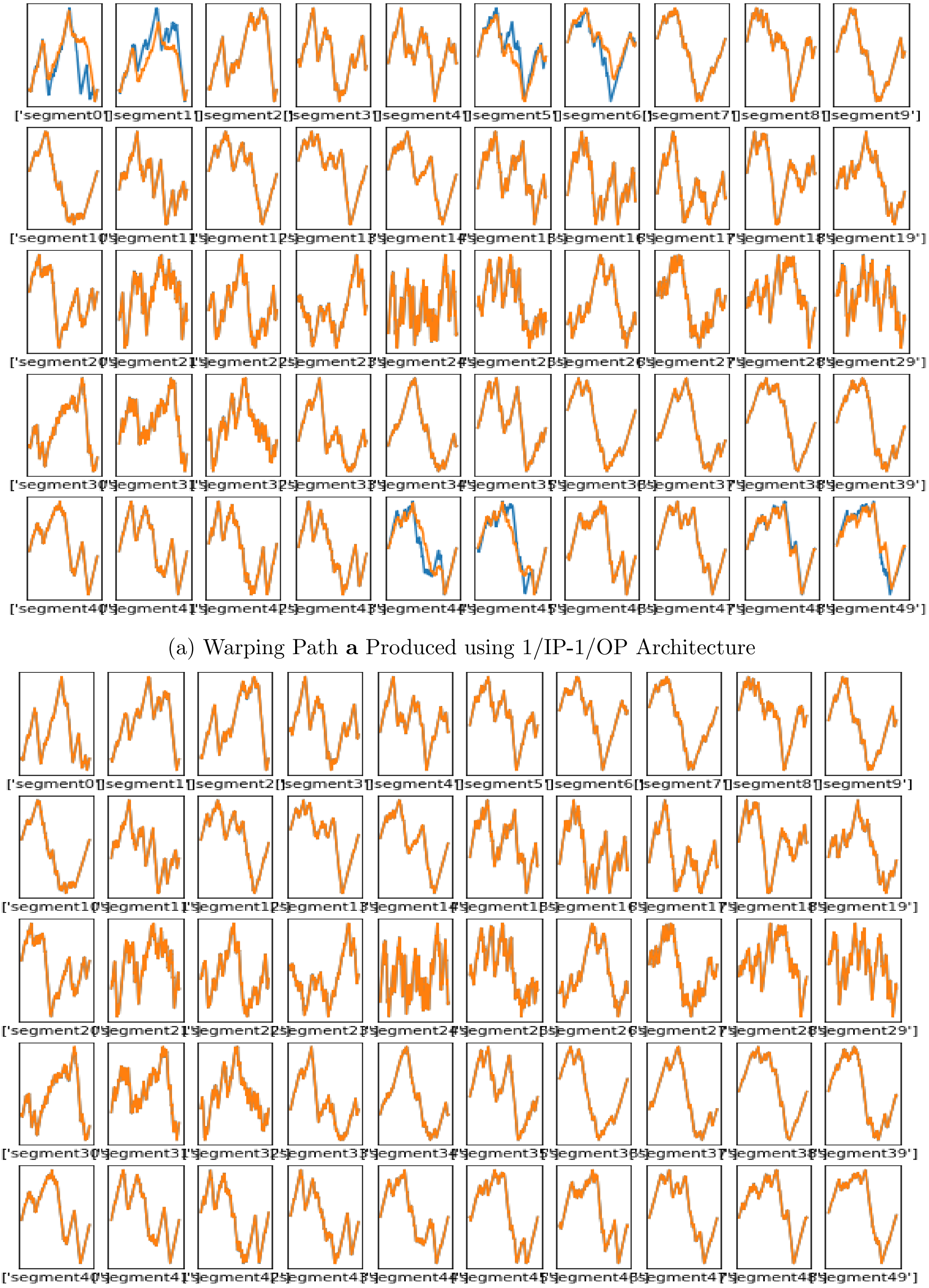
(a) Male Spectrogram Information

(b) Female Spectrogram Information

(c) Warped Male Spectrogram Information, Architecture 1

(d) Warped Male Spectrogram Information, Architecture 2

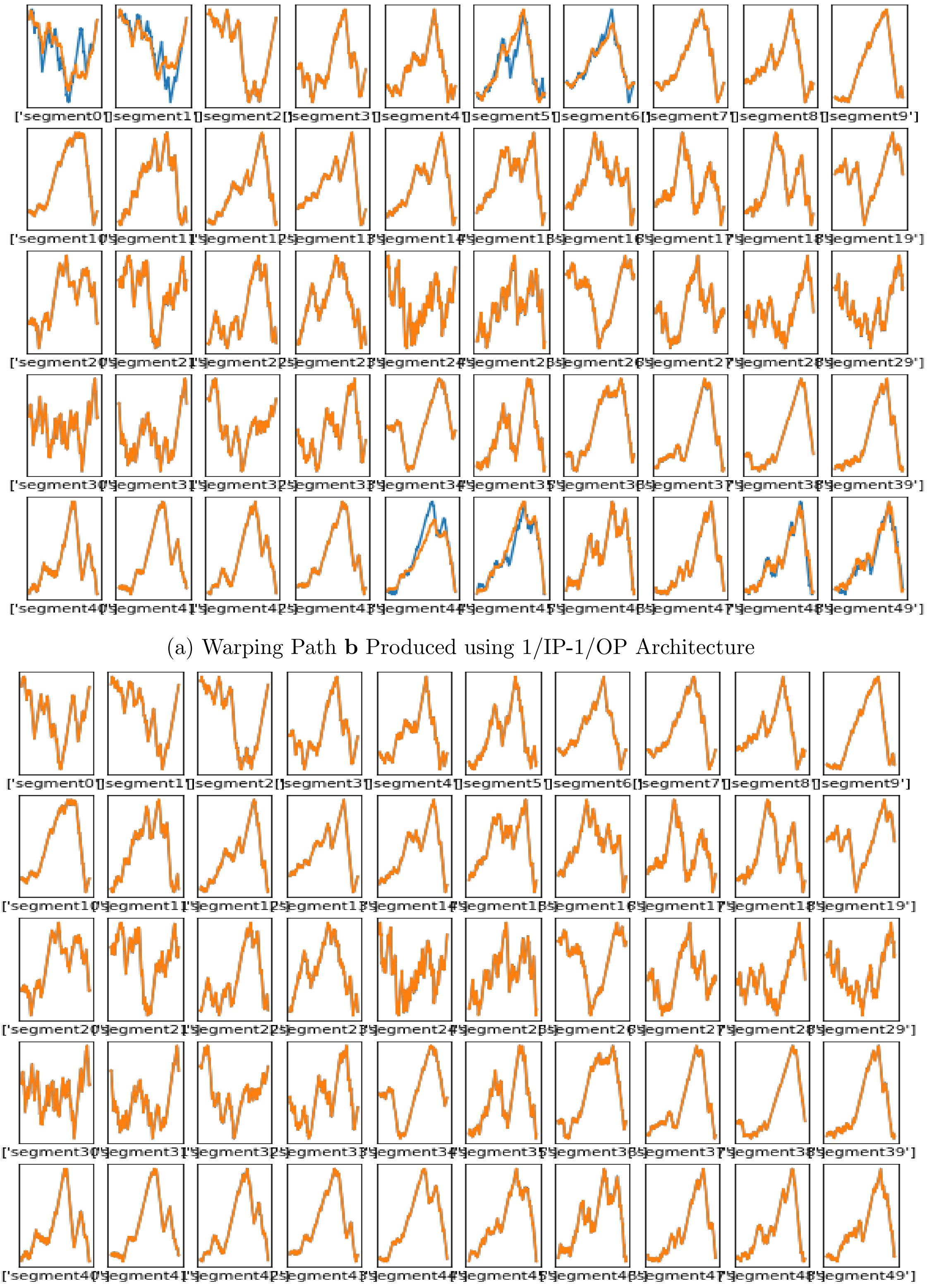
Fig. 4.14: Spectrogram Information for Warped Male Speaker, Phase One



(a) Warping Path **a** Produced using 1/IP-1/OP Architecture

(b) Warping Path **a** Produced using 3/IP-1/OP Architecture

Fig. 4.15: Learned Warping Path **a**, Phase One



(a) Warping Path **b** Produced using 1/IP-1/OP Architecture

(b) Warping Path **b** Produced using 3/IP-1/OP Architecture

Fig. 4.16: Learned Warping Path **b**, Phase One

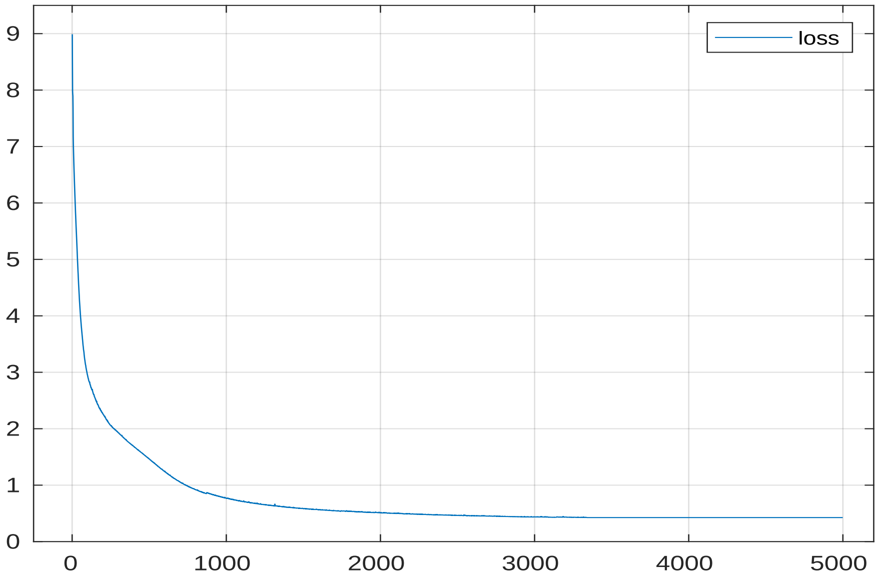
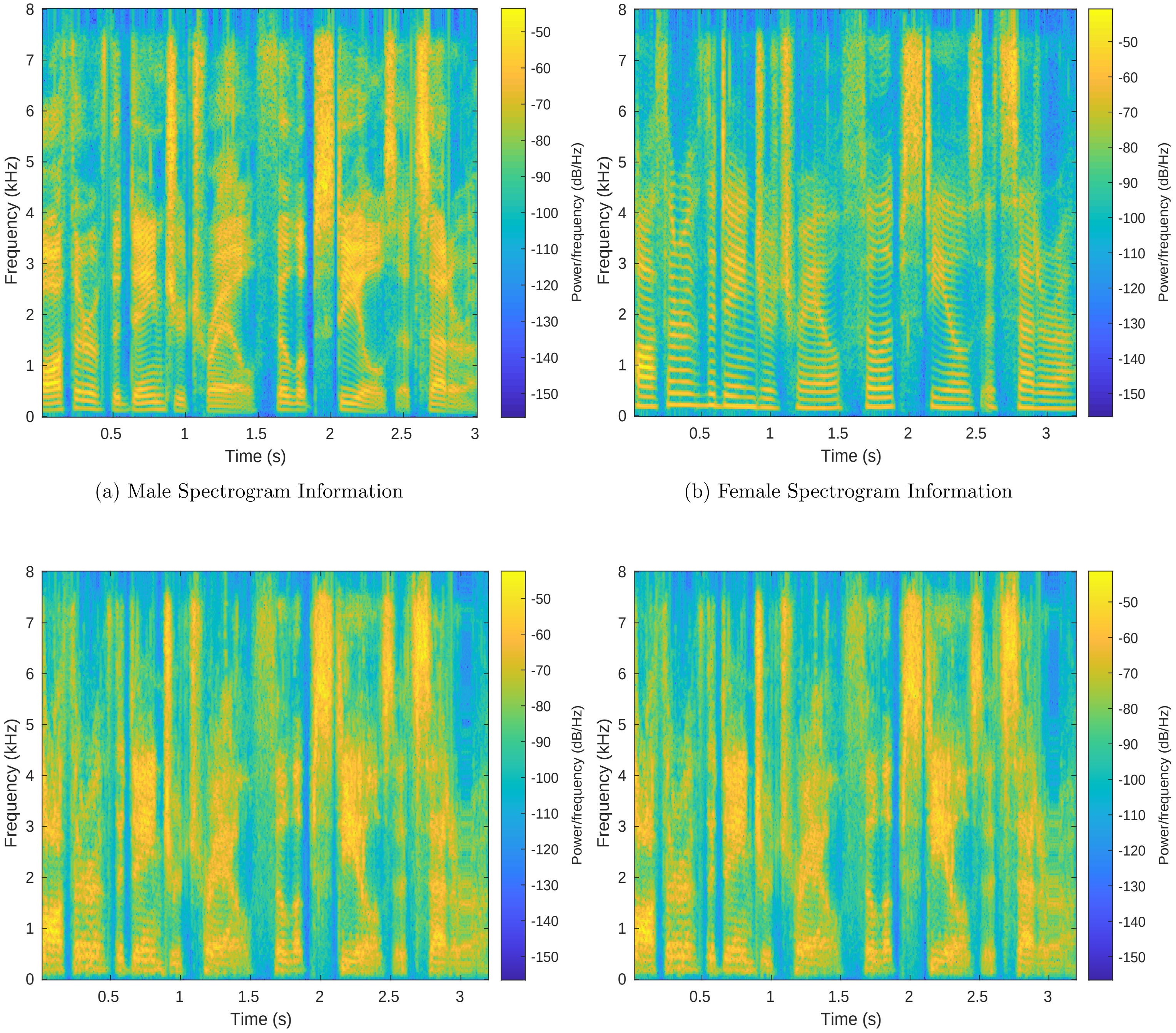


Fig. 4.17: Mean Squared Error (Learning Curve), Phase One

information for a warped male speaker using ANN of the selected number of layers and architecture with a phase reconstruction algorithm for the same phrase used in phase one after the network reached 10000 iterations (part (d)). Acoustically, the warped signal in part (d) is clearly looks like the female signal more than the result in part (c). The final sound of that transformation in part (d) (Play) is much better than the sound in part (c) (Play) but with some signal processing artifacts. Also the pitch lines in part (d) is not that sharp.

Figures 4.19 shows how the ANN has learned a warping path (the **a** warping data) at various segments of the selected speech signal at various points along the 10000 training iterations. Figure 4.19(a) was produced using the selected layers and architecture at 500 training iterations. Figure 4.19(b) produced using the selected network at 2500 training iterations. Figure 4.19(c) produced using the same selected layers and same architecture at 5000 training iterations. Finally, figure 4.19(d) produced by trained the selected network at 10000 training iterations. Figure 4.19(d) shows more matching between the DW path and the learned path than Figure 4.19(a),(b) and (c). The DW path is shown in blue and the learned path is in orange. By this point, many of the warping functions have been learned well, but there are still a significant number that have not been learned well.



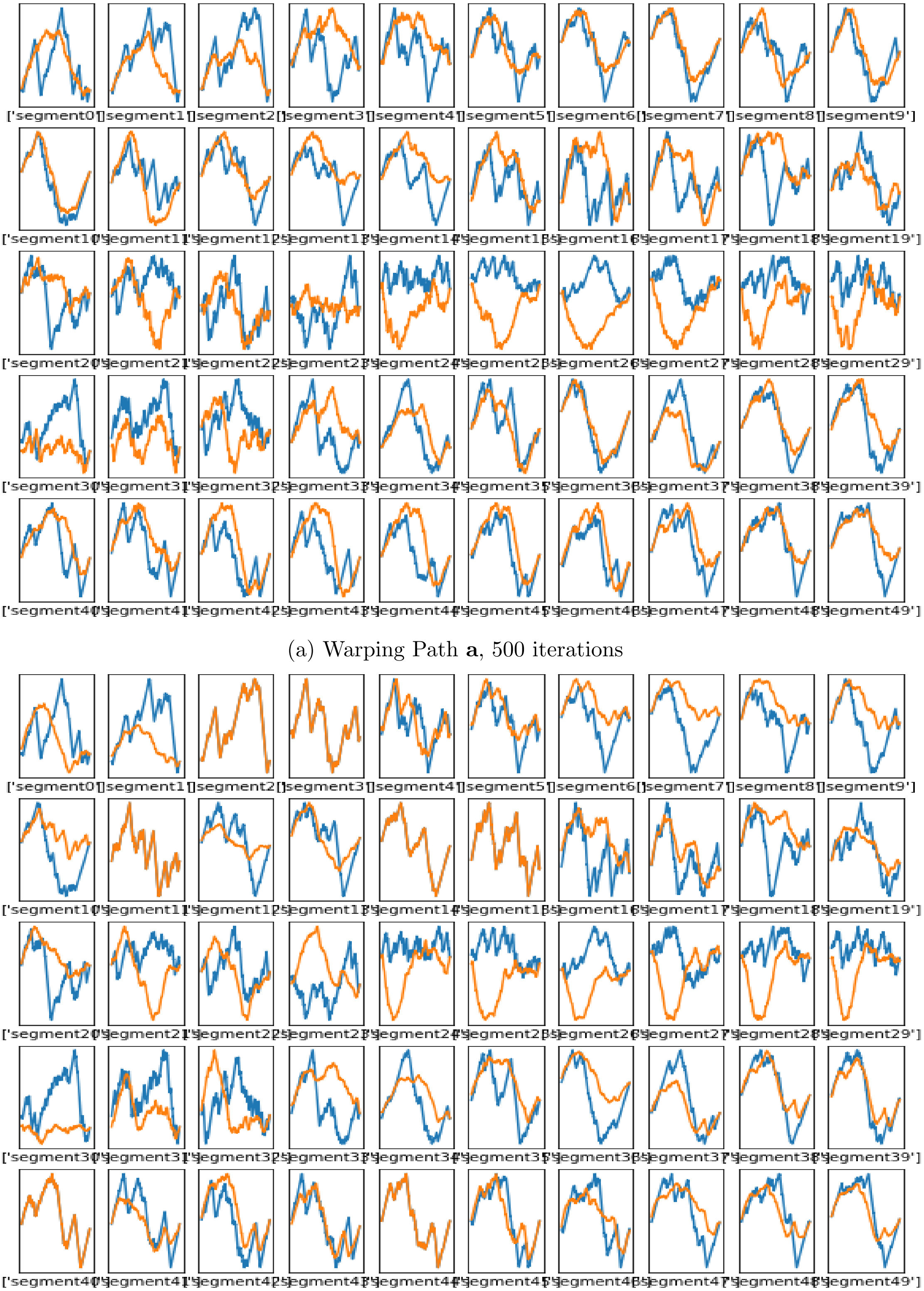
(a) Male Spectrogram Information

(b) Female Spectrogram Information

(c) Warped Male Spectrogram Information, 5000 iterations

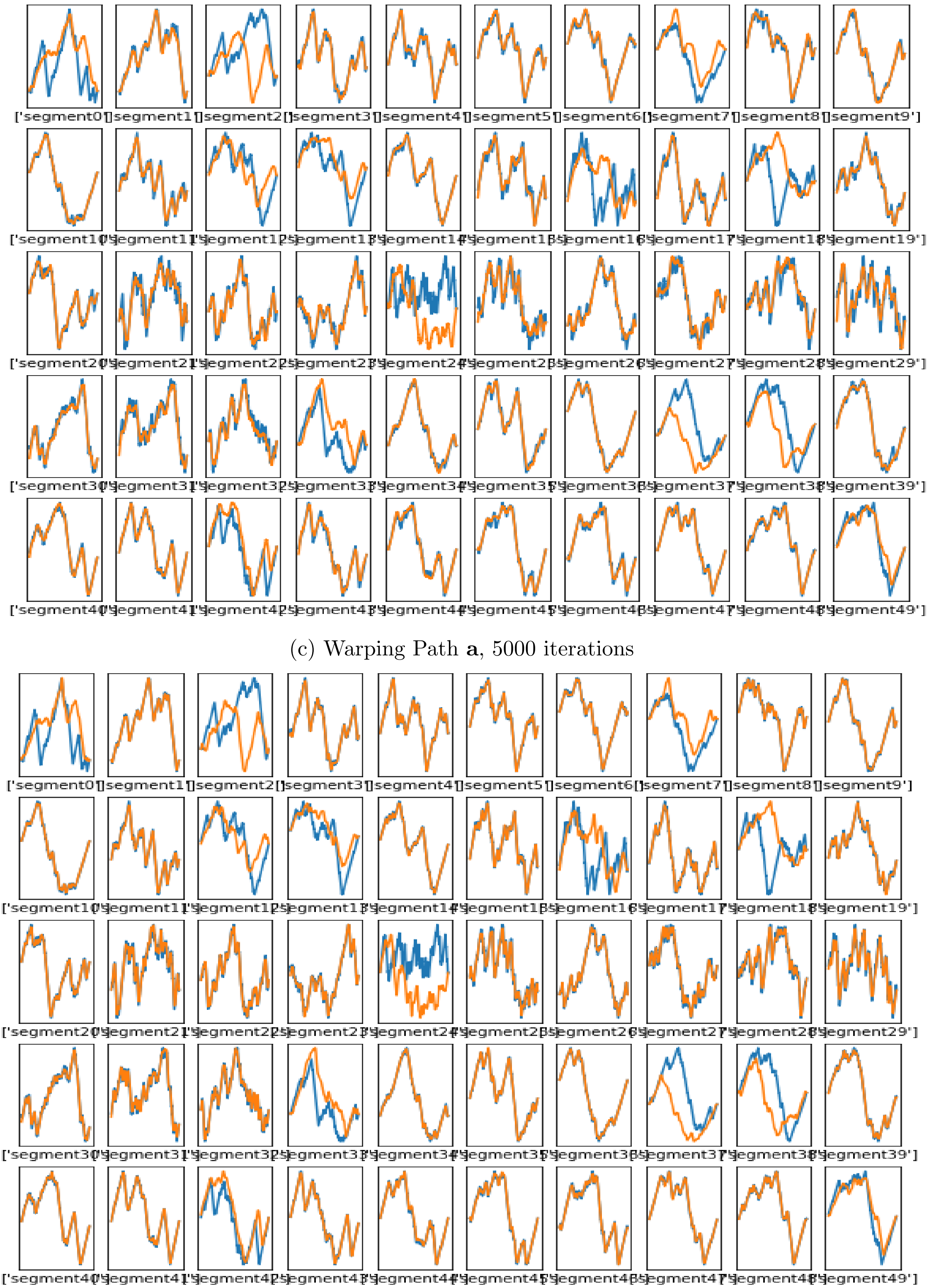
(d) Warped Male Spectrogram Information, 10000 iterations

Fig. 4.18: Spectrogram Information for Warped Male Speaker, Phase Two



(a) Warping Path **a**, 500 iterations

(b) Warping Path **a**, 2500 iterations



(c) Warping Path **a**, 5000 iterations

(d) Warping Path **a**, 10000 iterations

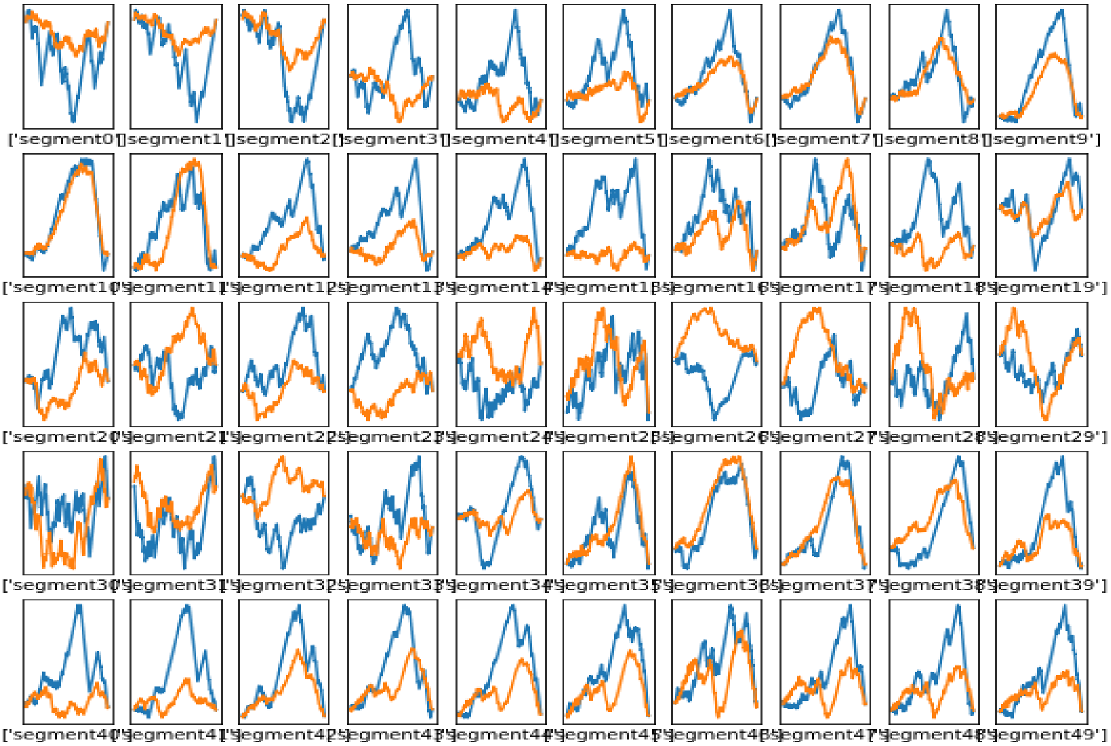
Fig. 4.19: Learned Warping Path **a**, Phase Two

Figures 4.20 shows how the ANN has learned a **b** warping path at various segments of the selected speech signal at different training iterations. Figure 4.20(a) produced at 500 training iterations, Figure 4.20(b) generated at 2500 training iterations, Figure 4.20(c) produced at 5000 training iterations and Figure 4.20(d) produced by trained the selected network for 10000 training iterations. Again, more learning matching with NN-3/IP-1/Op architecture (Figure 4.20(d)) for 10000 iterations than other NN-learned paths generated for different training iterations. The DW path is shown in blue and the NN-learned path is in orange. Figure 4.21 shows the mean square error between the true values of the warping paths (**a** and **b**) and the estimated one. The MCD values for this phase were reported in the fields of phase two in Table 4.1.

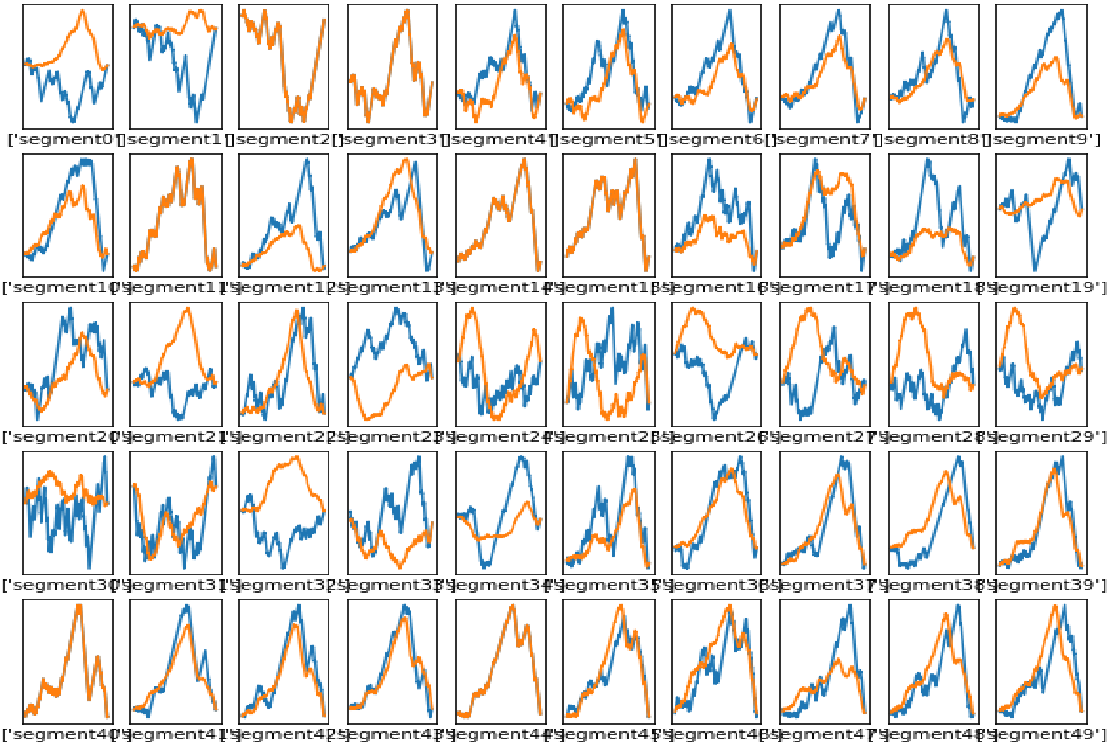
4.10.3 Phase Three

While the results of phase two were promising, there was still enough error on enough warping path estimates that it was determined to direct the neural network more by presenting data that has been clustered. Several neural networks were trained each to be responsive to data pertaining to a particular cluster.

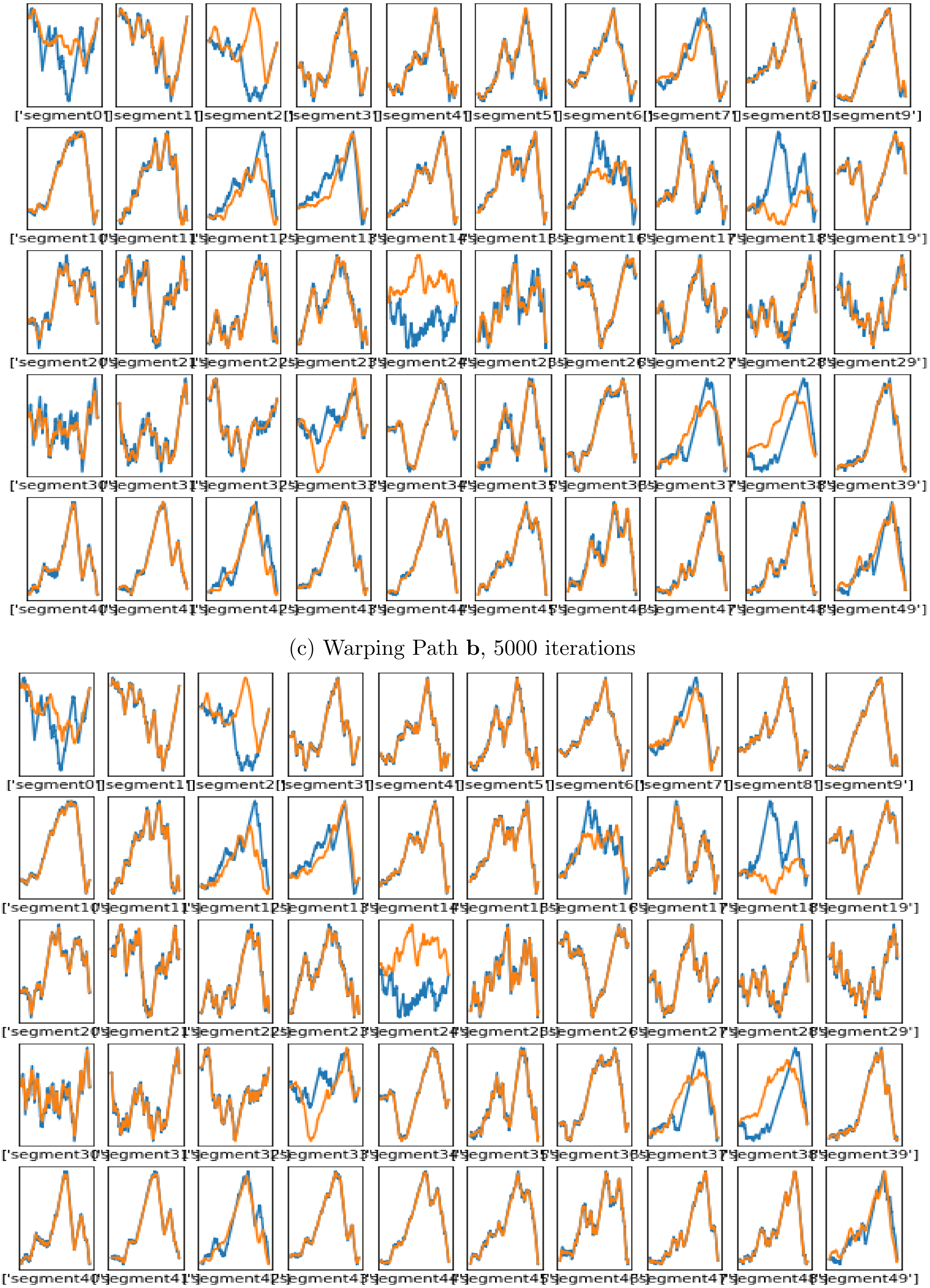
Cluster analysis, or clustering, is an unsupervised machine learning task that involves the grouping of data points. Given a set of data points, clustering algorithm can be used to classify each data point into a specific group. Many clustering algorithms follow a different set of rules for defining the similarity or distance among data points in an effort to discover the dense regions of the data. K-Means, which is probably the most well-known clustering algorithm, was chosen in this work to cluster the spectral information. Figure 4.22 shows the clusters for the selected 600 phrases spectral data with $k = 6$ clusters. Data from each of the k clusters was used to train an ANN from phase two for 1000 training iterations to predict the warping paths. The concept is shown in Figure 4.23, the six trained NN models were used to test the selected phrase, “Author of the danger trail, Philip Steels, etc.”, Figure 4.24(a). As shown in Figure 4.24(b), the partitioned spectral features for the testing phrase, $\mathbf{s}_{1:6}$, were combined together ($\mathbf{s}_i$ means the spectral features of the i^{th} cluster), and the partitioned estimated warping paths, $(\mathbf{a}_F, \mathbf{b}_F)_{(1:6),estimated}$, were combined



(a) Warping Path **b**, 500 iterations



(b) Warping Path **b**, 2500 iterations



(c) Warping Path **b**, 5000 iterations

(d) Warping Path **b**, 10000 iterations

Fig. 4.20: Learned Warping Path **b**, Phase Two

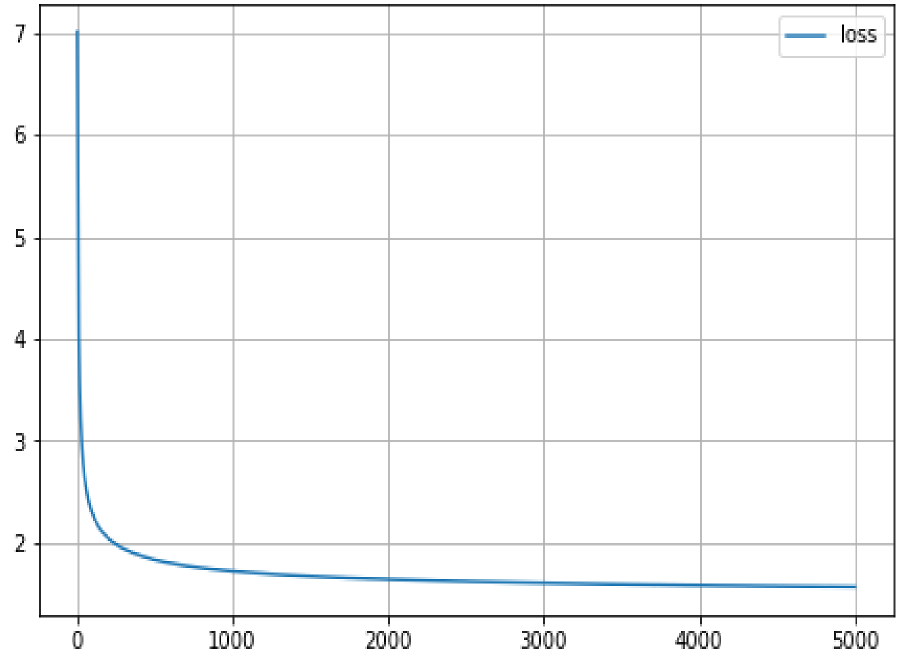


Fig. 4.21: Mean Squared Error (Learning Curve), Phase Two

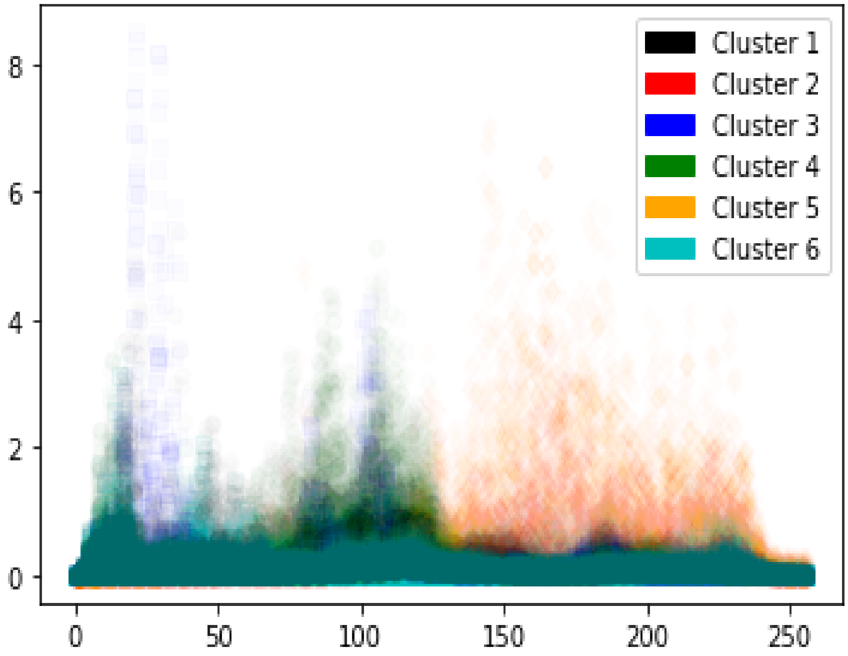
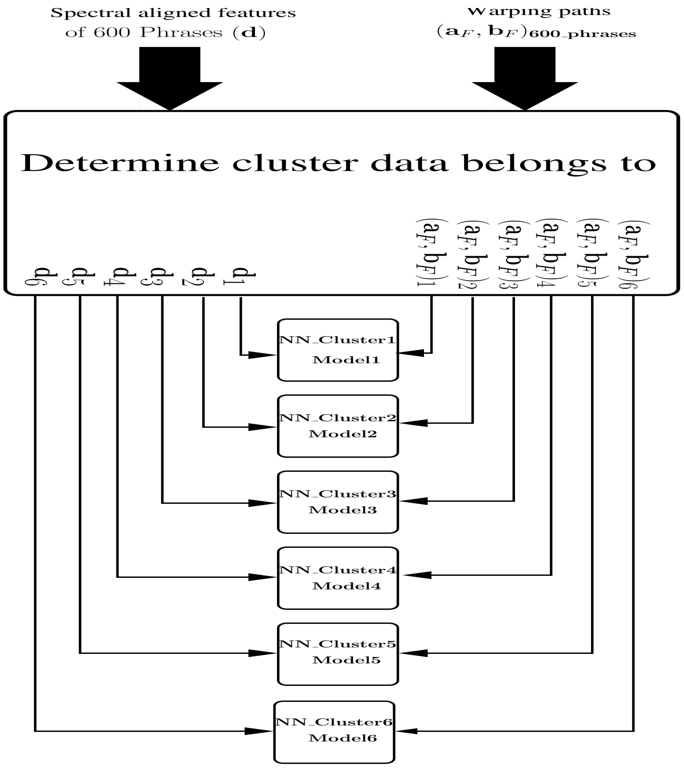


Fig. 4.22: Clustering Analysis for the Spectral Information ($k = 6$)



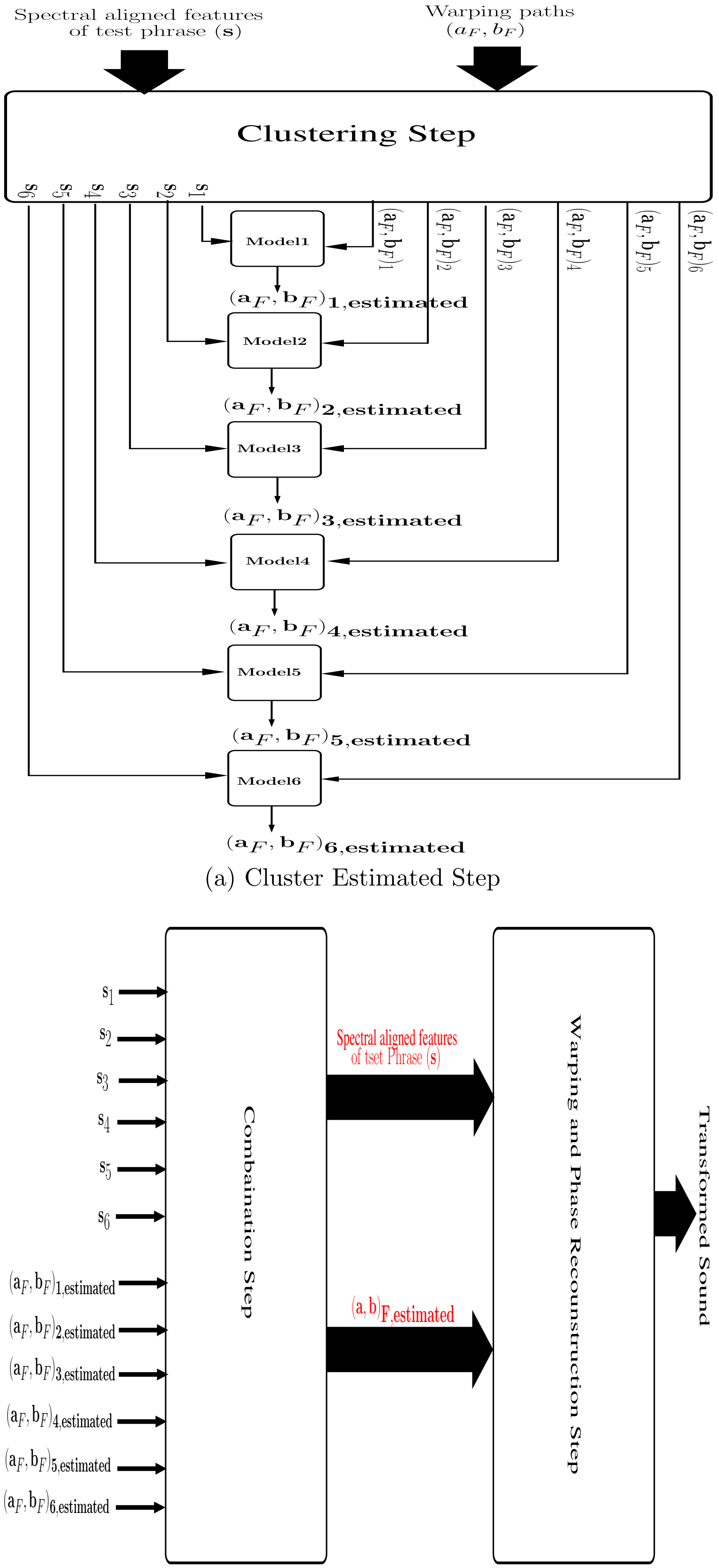
(a) Cluster training Step

Fig. 4.23: Cluster Training Method

together ($(\mathbf{a}_F, \mathbf{b}_F)_{i,,estimated}$ mean the estimated warping paths of the i^{th} cluster associated with $\mathbf{s}_i$). The combined features and estimated paths were passed through warping and phase reconstruction step to produce the transformed signal.

Figure 4.25 shows an example spectrogram information for a male speaker (part (a)), a spectrogram information for a female speaker (part (b)), a spectrogram for a warped male speaker using the selected clustering architecture ANN for the same phrase used in phase one after the network reached the 10000 iterations with a phase reconstruction algorithm (part (c)), and a spectrogram for a warped male speaker using the same clustering ANN with a phase reconstruction algorithm at 10000 training iterations (part (d)) (Press ”Play” box to play the transformed voice form male to female sounds using clustering method (Play)). The pitch lines in part (d) are stronger than the pitch lines in Figure 4.18 (part d).

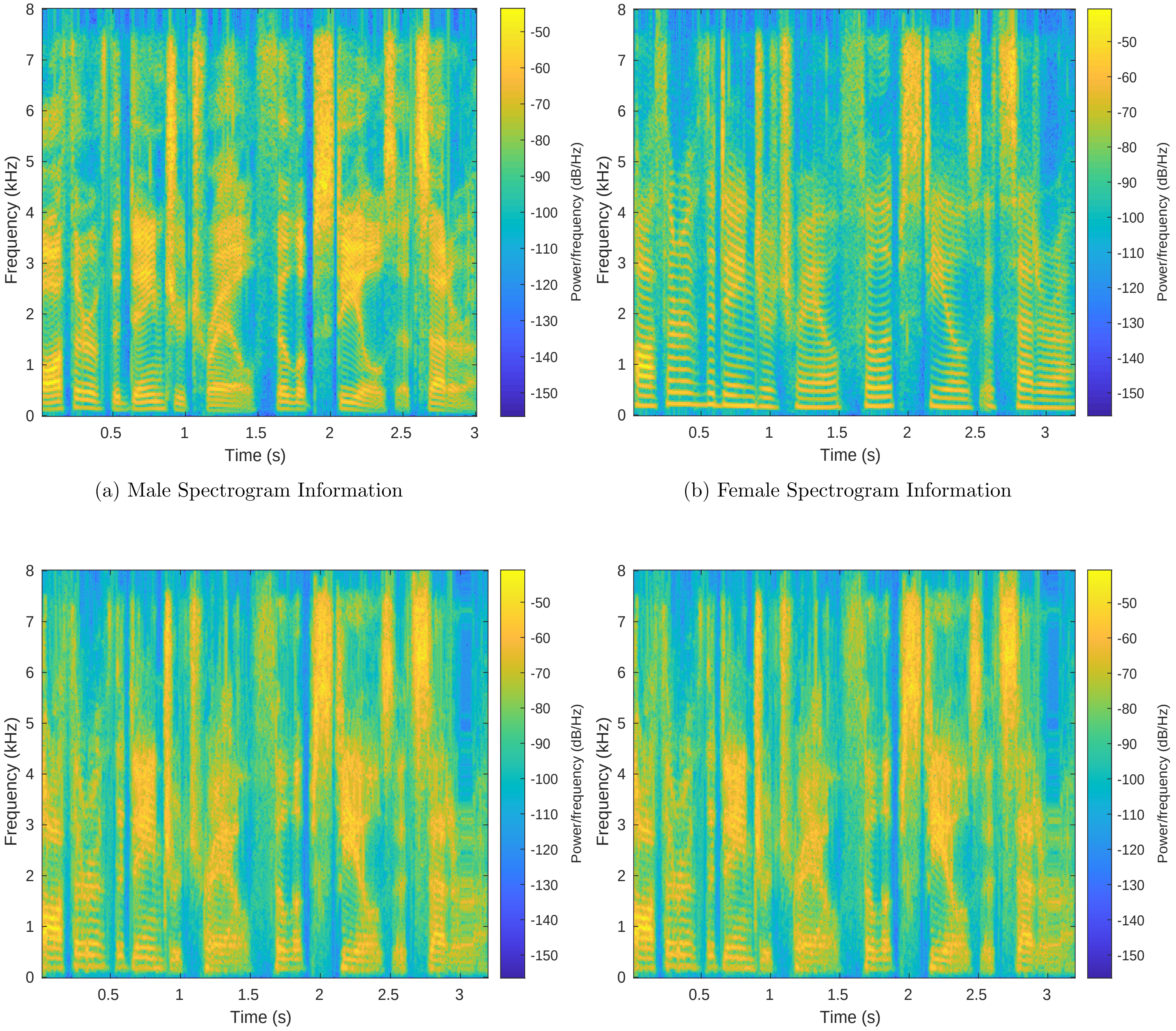
Figure 4.26 shows how the clustering ANN has learned the warping paths **a** and **b** at various segments of the selected speech signal after different number of training iterations. Figure 4.26(a),(b) and (c) show how the clustering ANN learned the warping path **a** at


Spectral aligned features of test phrase (s)
Warping paths (a_F, b_F)
Clustering Step
s1
s2
s3
s4
s5
s6
$(a_F, b_F)_1$
$(a_F, b_F)_2$
$(a_F, b_F)_3$
$(a_F, b_F)_4$
$(a_F, b_F)_5$
$(a_F, b_F)_6$
Model1
$(\mathbf{a}_F, \mathbf{b}_F)_{\mathbf{1,estimated}}$
Model2
$(\mathbf{a}_F, \mathbf{b}_F)_{\mathbf{2,estimated}}$
Model3
$(\mathbf{a}_F, \mathbf{b}_F)_{\mathbf{3,estimated}}$
Model4
$(\mathbf{a}_F, \mathbf{b}_F)_{\mathbf{4,estimated}}$
Model5
$(\mathbf{a}_F, \mathbf{b}_F)_{\mathbf{5,estimated}}$
Model6
$(\mathbf{a}_F, \mathbf{b}_F)_{\mathbf{6,estimated}}$
s_1
s_2
s_3
s_4
s_5
s_6
$(a_F, b_F)_{1,estimated}$
$(a_F, b_F)_{2,estimated}$
$(a_F, b_F)_{3,estimated}$
$(a_F, b_F)_{4,estimated}$
$(a_F, b_F)_{5,estimated}$
$(a_F, b_F)_{6,estimated}$
Combaination Step
Spectral aligned features of tset Phrase (s)
$(a, b)_{F,estimated}$
Warping and Phase Reconstruction Step
Transformed Sound

(a) Cluster Estimated Step

(b) Combination and warping Step

Fig. 4.24: Cluster Warping Method



(a) Male Spectrogram Information

(b) Female Spectrogram Information

(c) Warped Male Spectrogram Information, 5000 iterations

(d) Warped Male Spectrogram Information, 10000 iterations

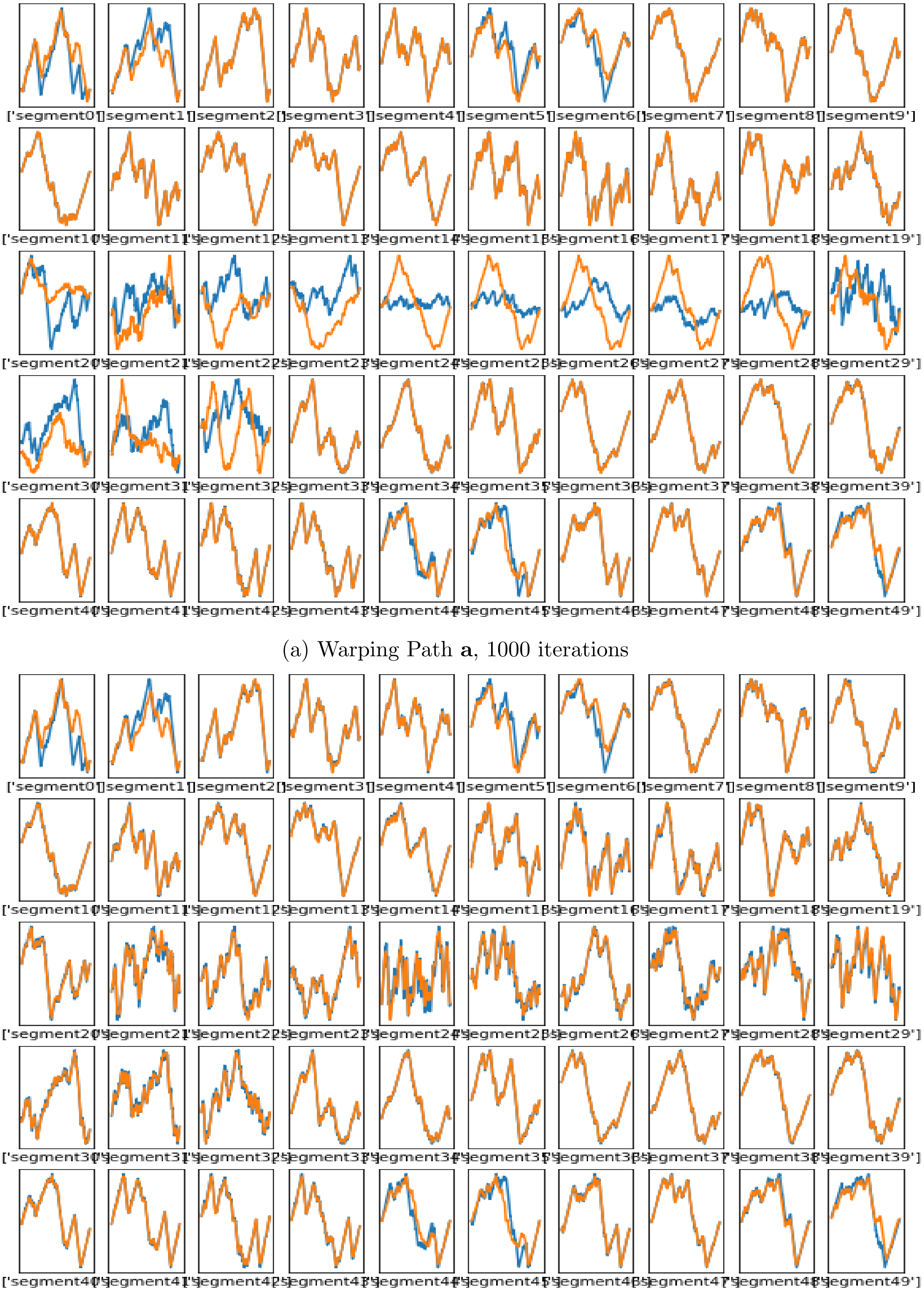
Fig. 4.25: Spectrogram Information for Warped Male Speaker, Phase Three

1000, 2500 and 5000 training iterations, respectively. Figure 4.26(d), (e) and (f) show the learned warping path **b** at 1000, 2500 and 5000 training iterations, respectively. Also, the DW path is shown in blue and the NN-learned path is in orange. The final sound at the 10000 training iterations is much better than the final sound produced from phase two and with little signal processing artifacts. Figure 4.27 shows the mean square error between the true values of the warping paths (**a** and **b**) and the estimated one for the six clusters. The MCD values for this phase were reported in the fields of phase three in Table 4.1.

4.10.4 Phase Four

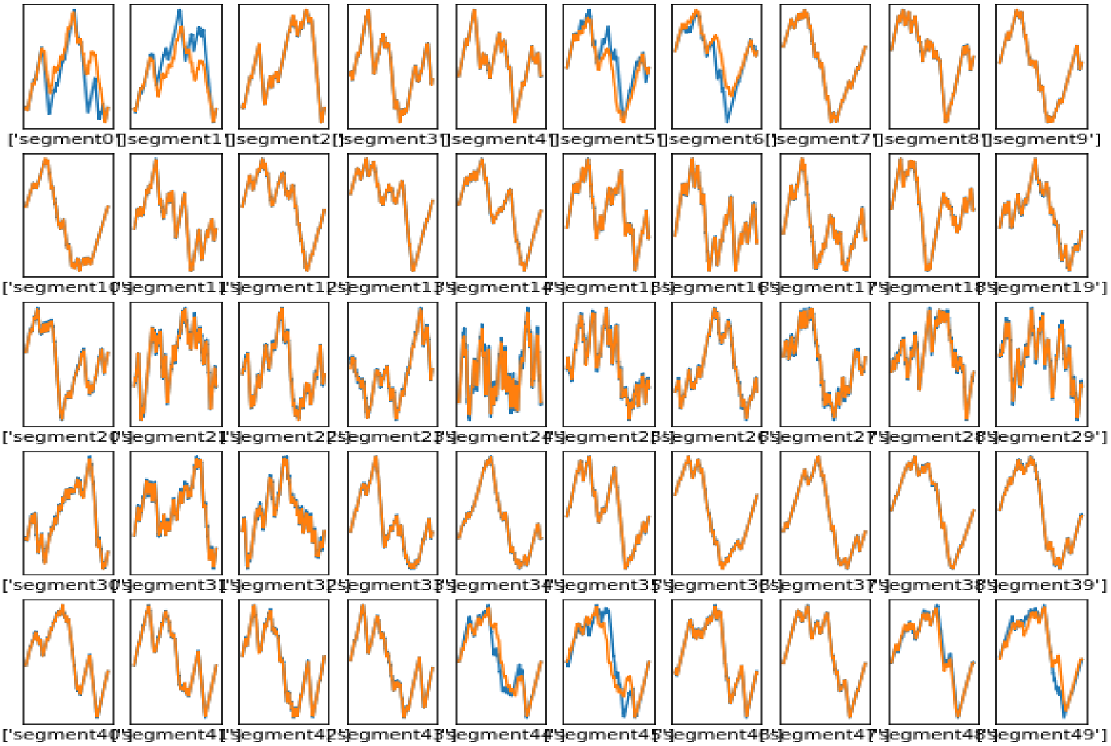
Convolutional neural networks (CNN) have been very successfully applied to image processing analysis. It was suggested to see how they would works at estimating warping function. In phase four, we implemented a 1-D convolutional neural network ($1-DCNN$) to attempt to achieve voice transformation. Convolutional neural network models were developed to operate exclusively on 2D data such as images and videos, in which the model learns an internal representation of a two-dimensional input, in a process referred to as feature learning. This same process can be applied to one-dimensional sequences of data, such as time series data.

The first task in building 1-D CNN based voice transformation system is to find an optimal architecture for 1-D CNN. To experiment with different CNN architectures we considered, the source male time aligned spectral features are used as an input to the CNN and the warping paths $\mathbf{a}_F$ and $\mathbf{a}_F$ are used as an output to the CNN in two main architectures ($1/IP-1/OP$ and $3/IP-1/OP$). The source male time aligned spectral features were computed by first computing the spectral feature information for the selected first 600 phrases form the CMU-ARCTIC database and then applying the two-level DW to achieve the time alignment and saving warping paths ($\mathbf{a}_F$ and $\mathbf{a}_F$) to be used in the training process. After a lot of experiments on different number of layers (four, five, six, seven,eight and more layers) with different number of neurons for each layer and according to the mean square values, MCD values, quality of the final sound and the spectrogram figures, the selected 1-D CNN for the $1/IP-1/OP$ architecture is seven layer network and

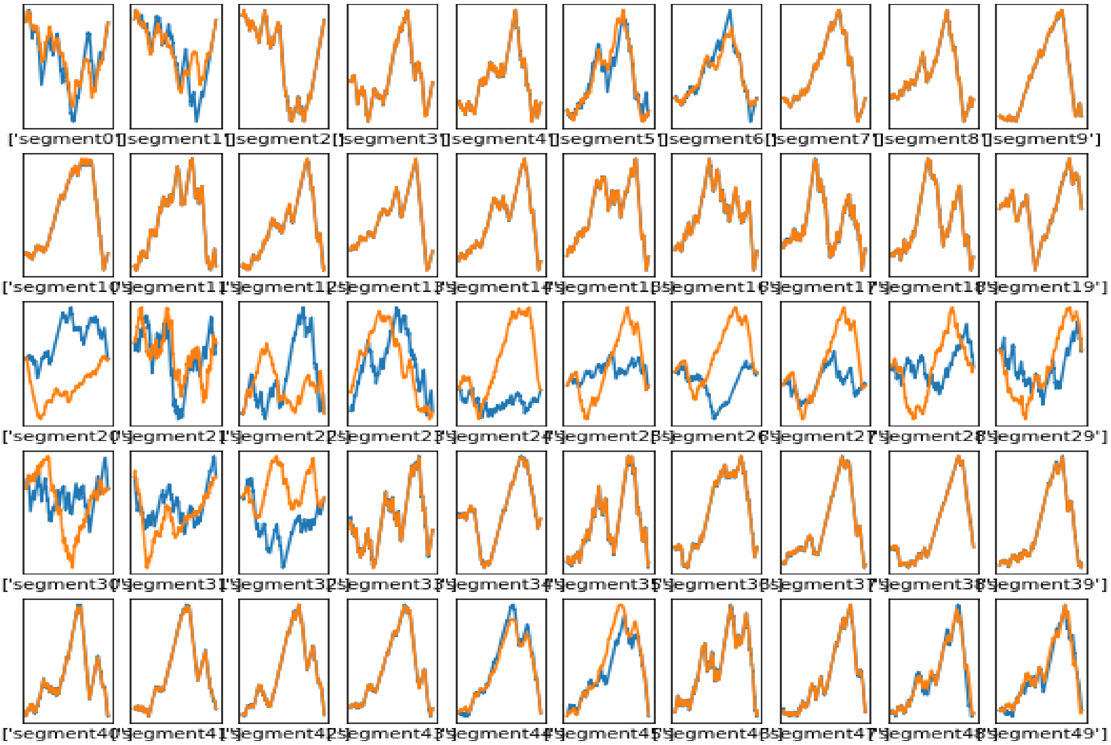


(a) Warping Path **a**, 1000 iterations

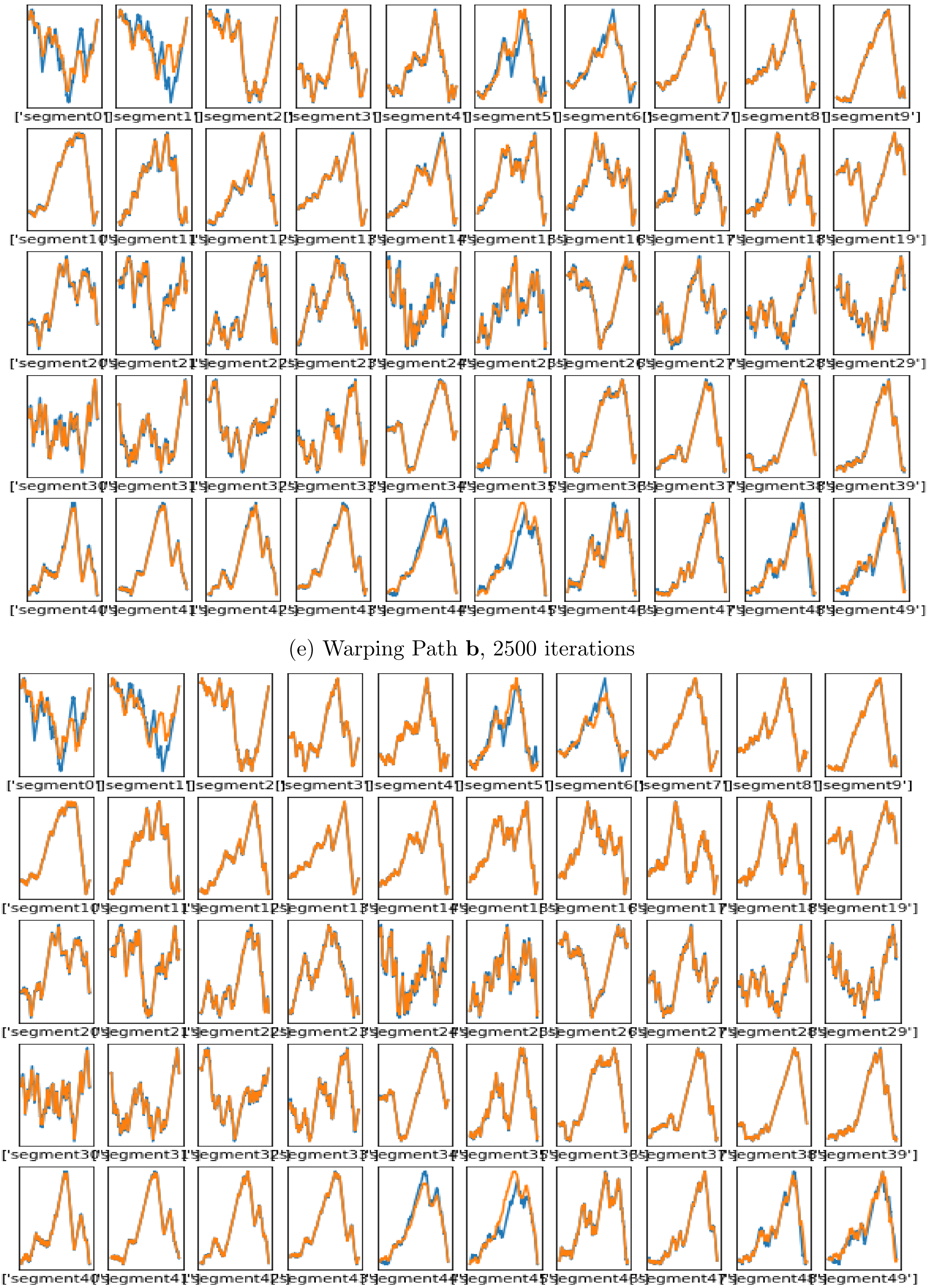
(b) Warping Path **a**, 2500 iterations



(c) Warping Path **a**, 5000 iterations



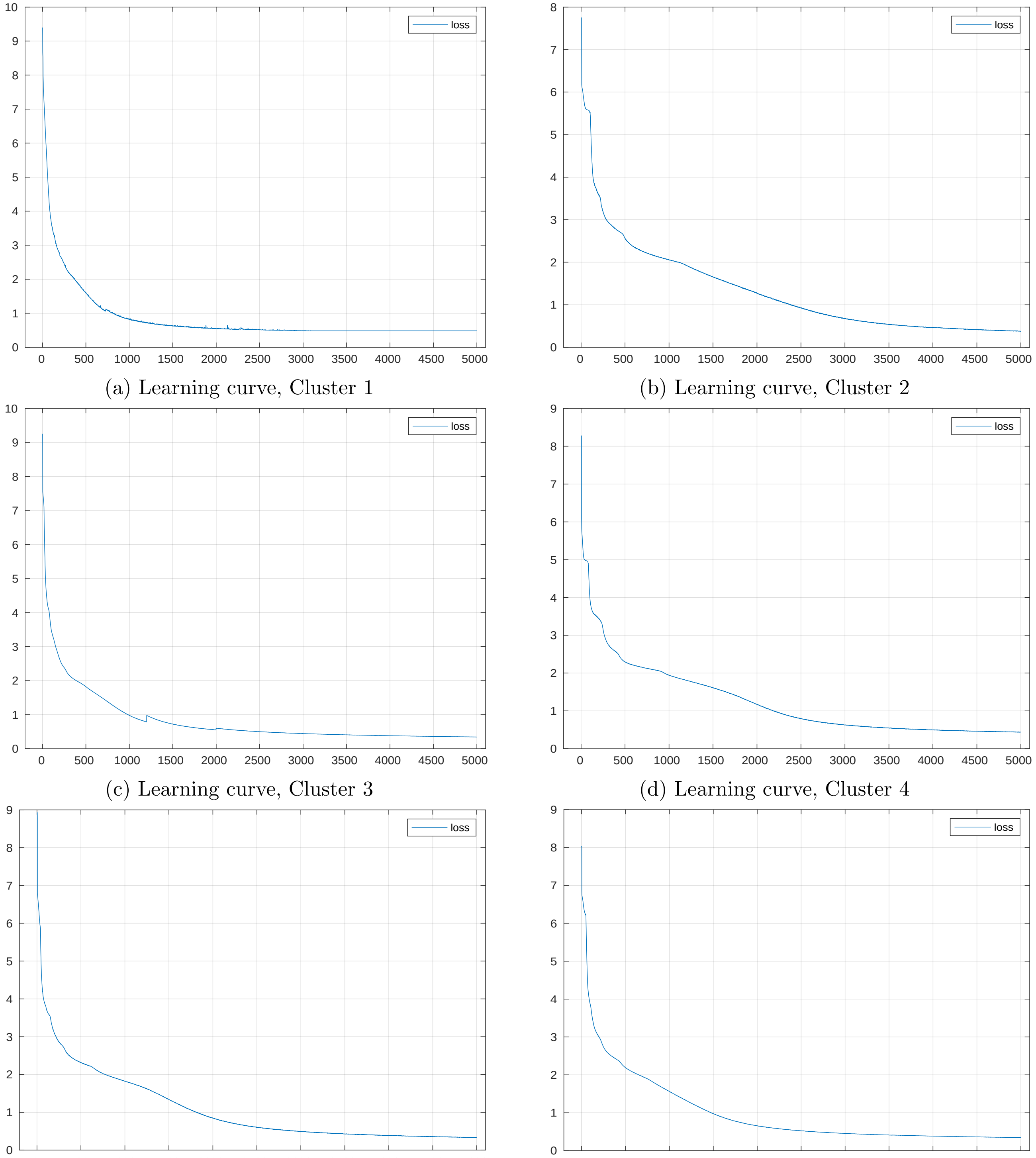
(d) Warping Path **b**, 1000 iterations



(e) Warping Path **b**, 2500 iterations

(f) Warping Path **b**, 5000 iterations

Fig. 4.26: Learned Warping Paths **a** and **b**, Phase Three



(a) Learning curve, Cluster 1

(b) Learning curve, Cluster 2

(c) Learning curve, Cluster 3

(d) Learning curve, Cluster 4

(e) Learning curve, Cluster 5

(f) Learning curve, Cluster 6

Fig. 4.27: Mean Squared Error (Learning Curve) using Cluster Method

five layer for the architecture of $3/IP - 1/OP$ as shown in the fields of phase four in Table 4.1. For instance, 256L F46N(k=3) F100N(k=3) FL D100N 964L means that it's a five-layer network, where the first layer of F64N(k=3) represents the number of output filters used in the convolution operation for this layer (46 filter) with the size of the convolutional window equal to 3 (kernel_size), after CNN the learned features are flattened to one long vector (FL) and pass through a fully connected layer of size 100 neuron before the output layer (946L) used to make a prediction. L represents "linear" output function and N represents "ReLU" output function (activation function).

Figure 4.28 shows a typical spectrogram information for a male speaker (part (a)), a typical spectrogram information for a female speaker (part (b)), a typical spectrogram information for a warped male speaker using $1/IP - 1/OP$ architecture of CNN at 2500 training iterations with a phase reconstruction algorithm (part (c)), a typical spectrogram information for a warped male speaker using $1/IP - 1/OP$ architecture of CNN at 5000 training iterations with a phase reconstruction algorithm for the same phrase used in phase one (part (d)), a typical spectrogram information for a warped male speaker using $3/IP - 1/OP$ architecture of CNN at 2500 training iterations with a phase reconstruction algorithm (part (e)), a typical spectrogram information for a warped male speaker using $3/IP - 1/OP$ architecture of CNN at 5000 training iterations with a phase reconstruction algorithm for the same phrase used in phase one (part (f)). Acoustically, the warped signal in part (f) is clearly looks like the male signal not the female signal, also the final warped sound is more close to the male sound than female sound and with a lot of signal processing artifacts (Press "Play" box to play the transformed voice form male to female sounds using convolutional method (Play)).

Figure 4.29 shows how the convolutional ANN has learned the warping paths **a** and **b** at various segments of the selected speech signal after different number of training iterations. Figure 4.29(a),(b) and (c) show how the clustering ANN learned the warping path **a** at 1000, 2500 and 5000 training iterations, respectively. Figure 4.29(d), (e) and (f) show the learned warping path **b** at 1000, 2500 and 5000 training iterations, respectively. Also, the DW path is shown in blue and the NN-learned path is in orange. Figure 4.30(a) shows the mean square error between the true values of the warping paths (**a** and **b**) and the estimated one for the CNN of $1/IP - 1/OP4$ architecture and Figure 4.30(b) show the learning curve for the CNN of the architecture of $3/IP - 1/OP$. The MCD values for this phase were reported in phase four field in table 4.1.

From Figure 4.30, the value of the mean square error (MSE) at 5000 training iteration is around 5.4 for $1/Ip - 1/OP$ architecture and its fixed at this value, while for the $3/IP - 1/OP$ the MSE is around 2.3 and its fixed at this value. On the other side, MSE values for the clustering method are much lower than these values for the CNN. Also by looking through table 4.1, the best MCD values is the one coming form the clustering method. Table 4.2 show the MCD values for the sound produced by using two-level dynamic warping (direct method) with and without using phase reconstruction. The MCD value for the direct method with phase reconstruction algorithm is lower than the MCD value of the direct method without phase reconstruction, this support our hypothesis of using phase reconstruction algorithm. From table 4.1 and table 4.2, we can see MCD value of the clustering method is the closet one to the MCD value of the direct method with phase reconstruction. According to these results, our experiments showing that unfortunately the work of CNN and the work in phase two are not suited to this particular problem and the work in phase three is the best work for this problem.

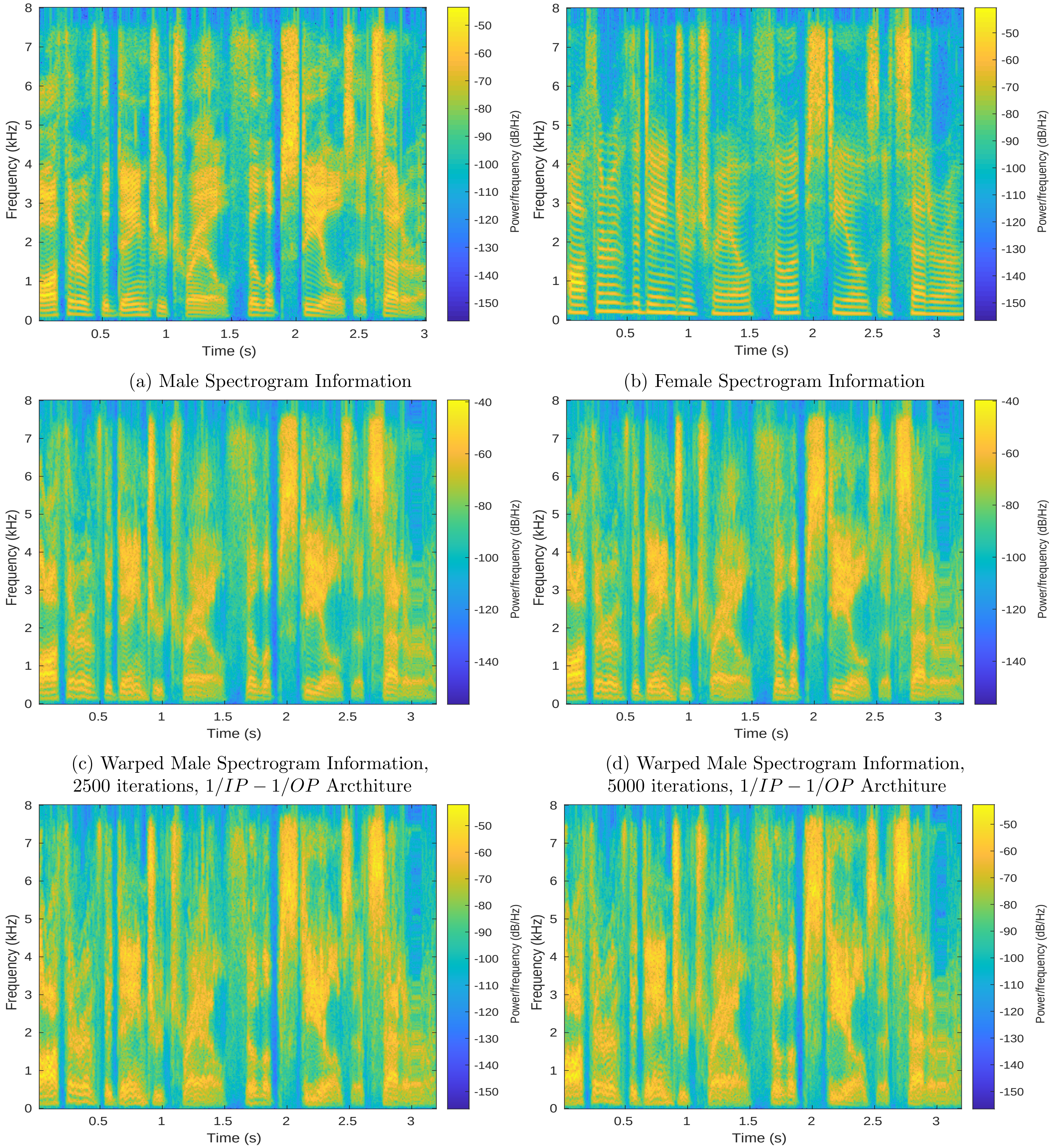
In order to show that if our clustering method of ANN based transformation can be considered as a good method to voice transformation, we have provided comparison between the result of our clustering method with the results in the literature as follows: From table 4.1, it is possible to observe that the values of distortion (MCD=2.8 dB) for the clustering method is lower than the ones presented in [82], which reported MCD of around 5.5 dB, the result from [82] is relatively to a male to female conversion. The performance of our method is better than the performance of the work reported in [44], the reported value of MCD is around 6.55 dB, the result form [44] is based on voice transformation using NN. The work of voice transformation using ANN in [67] reported MCD value of around 6.1 dB, which is higher than our value. The process of VT using ANN is shown in Algorithm 4.3. Table 4.3 shows different phrases as examples for our method of voice transformation form male to female speakers with the MCD values.

No.	Two-level DW method	MCD[db]
1	DW, with phase reconstruction	2.4
2	DW, without phase reconstruction	6.3

Table 4.2: MCD's Obtained for Two-Level DW with and without Phase Reconstruction

No.	Phrase	Male sound	Female sound	Transformed sound	MCD[dB]
1	Not at this particular case, Tom, apologized Whittemore	Play	Play	Play	2.6006
2	For the twentieth time that evening the two men shook hands	Play	Play	Play	2.64
3	Lord, but I'm glad to see you again, Phil	Play	Play	Play	2.57
4	Will we ever forget it	Play	Play	Play	2.44
5	God bless 'em, I hope I'll go on seeing them forever	Play	Play	Play	2.59
6	And you always want to evening the two men shook hands	Play	Play	Play	2.601

Table 4.3: Examples of different voice transformation sounds



(a) Male Spectrogram Information

(b) Female Spectrogram Information

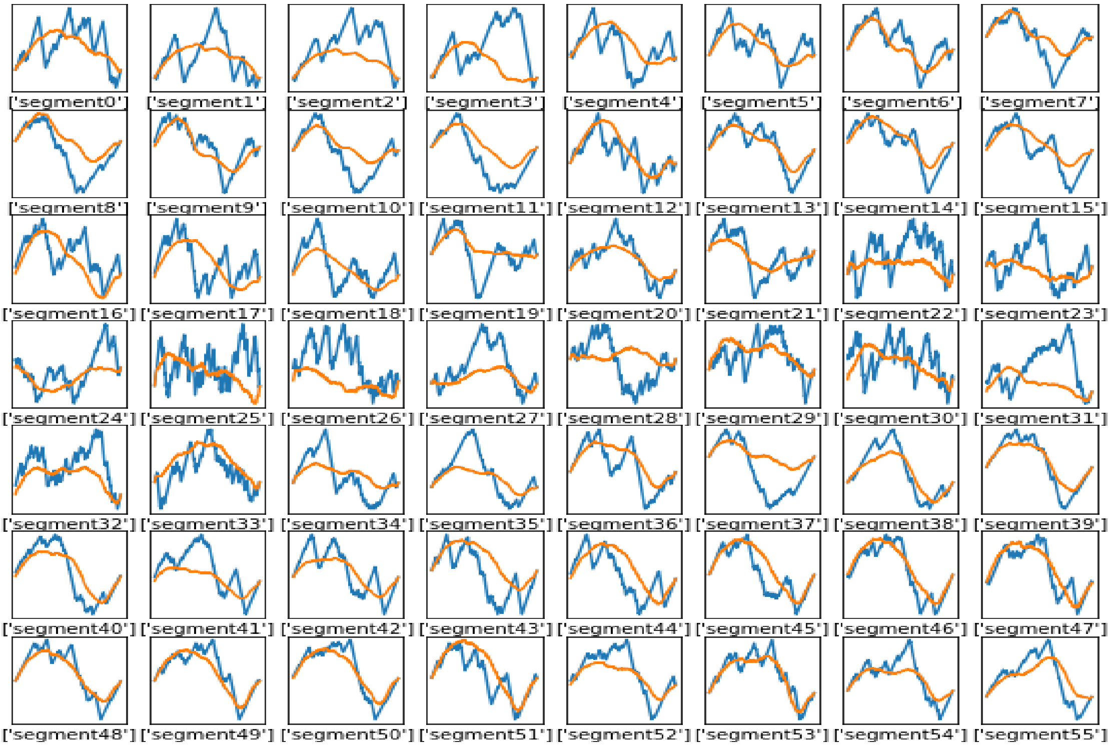
(c) Warped Male Spectrogram Information, 2500 iterations, $1/IP - 1/OP$ Arcthiture

(d) Warped Male Spectrogram Information, 5000 iterations, $1/IP - 1/OP$ Arcthiture

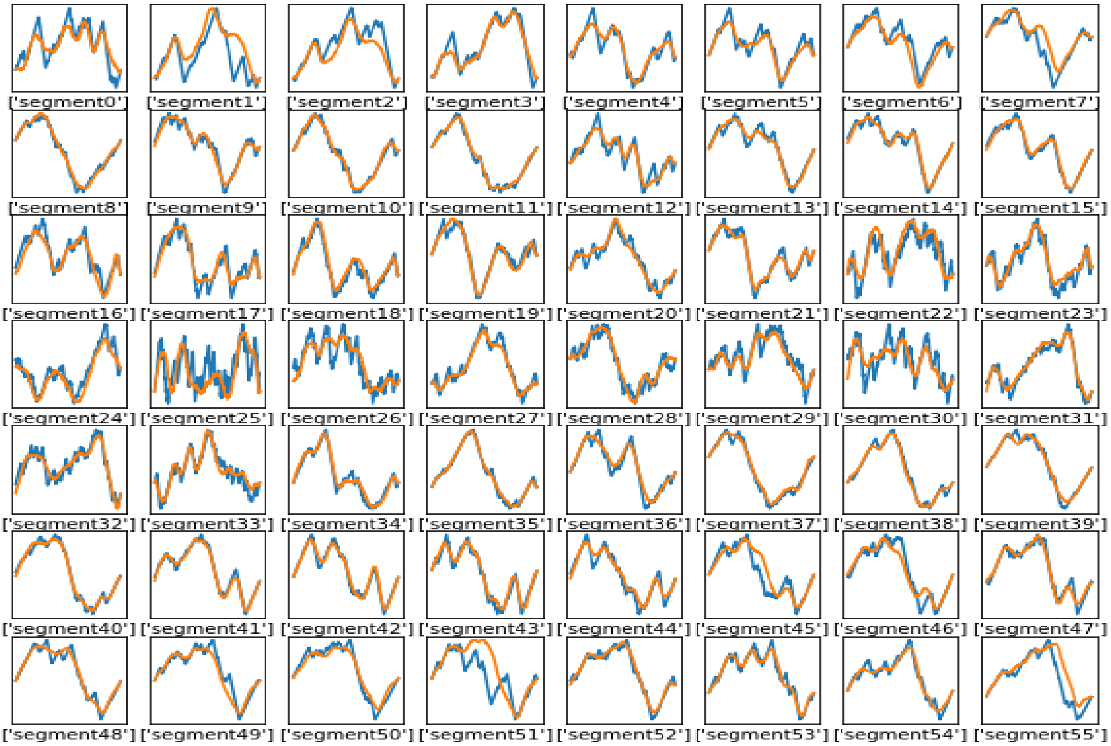
(e) Warped Male Spectrogram Information, 2500 iterations, $3/IP - 1/OP$ Arcthiture

(f) Warped Male Spectrogram Information, 5000 iterations, $3/IP - 1/OP$ Arcthiture

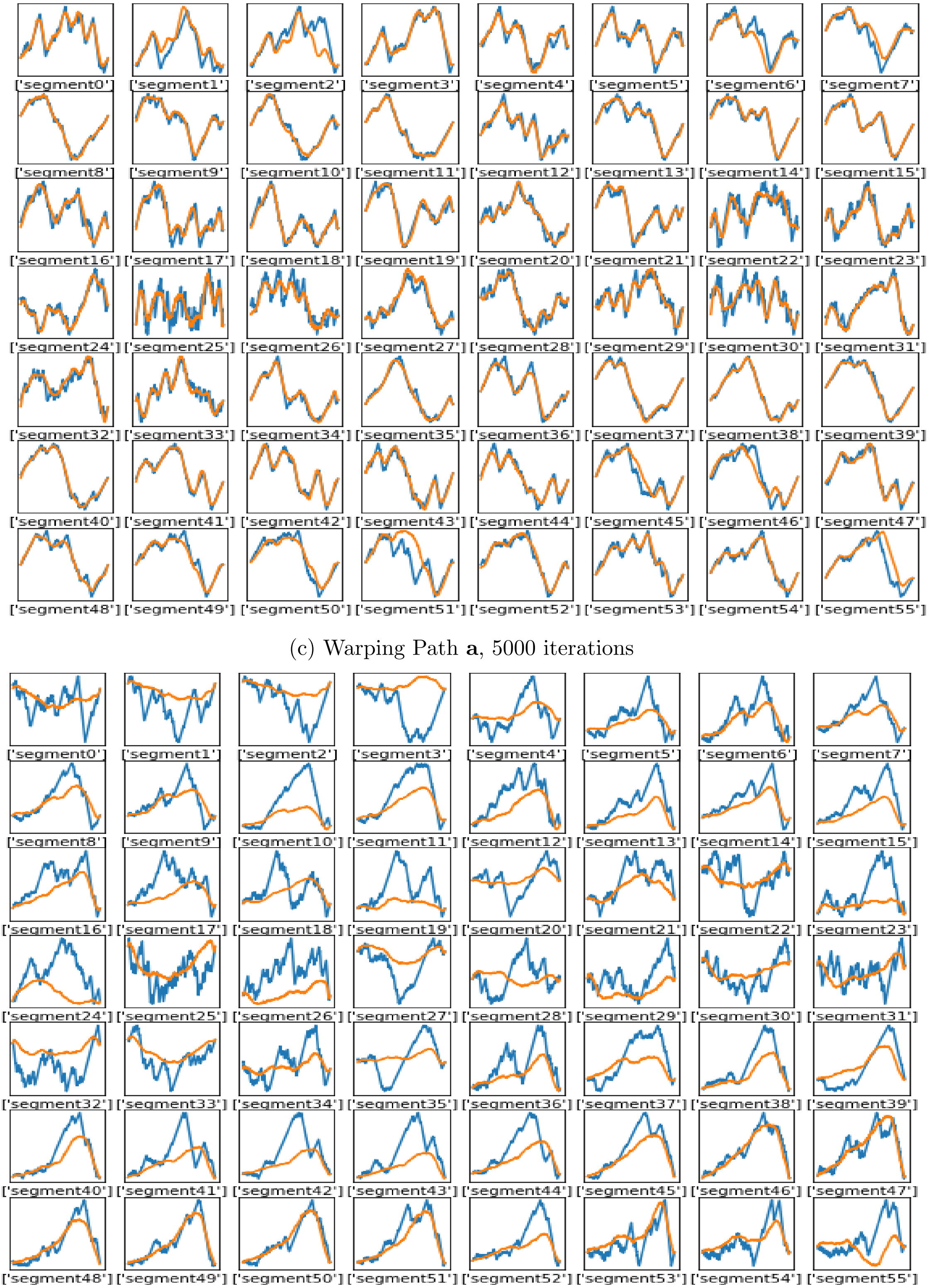
Fig. 4.28: Spectrogram Information for Warped Male Speaker, Phase Four



(a) Warping Path **a**, 1000 iterations

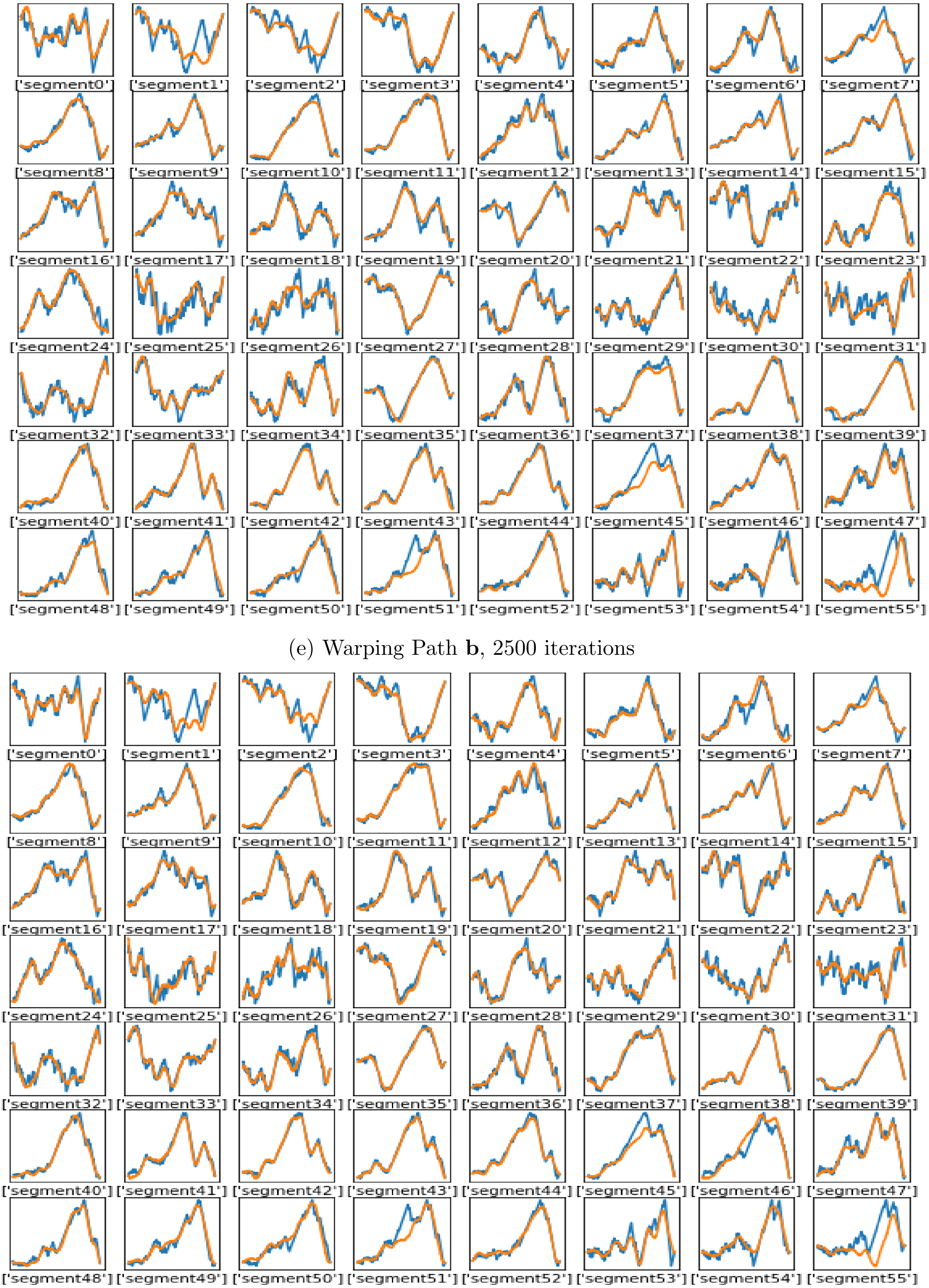


(b) Warping Path **a**, 2500 iterations



(c) Warping Path **a**, 5000 iterations

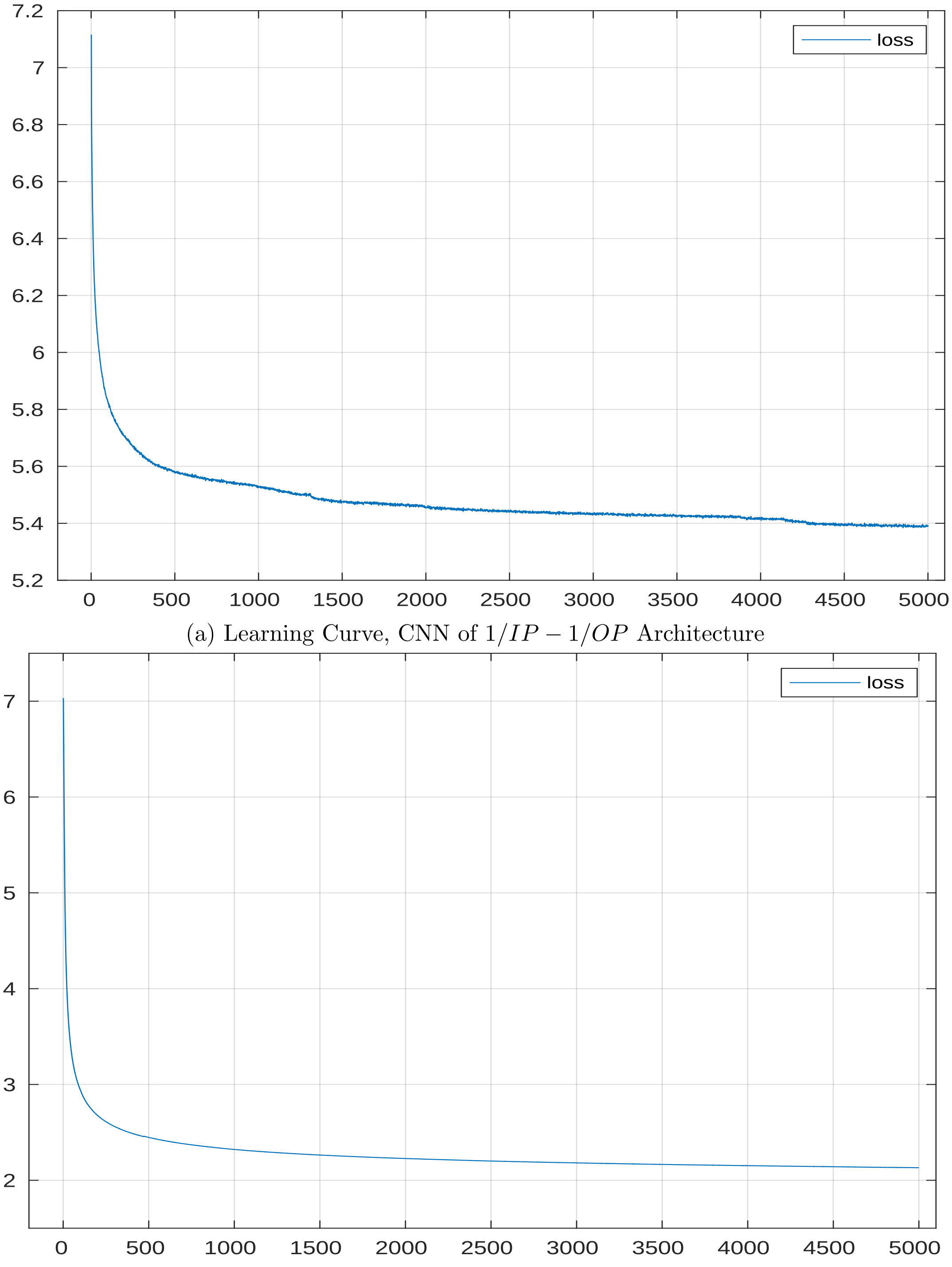
(d) Warping Path **b**, 1000 iterations



(e) Warping Path **b**, 2500 iterations

(f) Warping Path **b**, 5000 iterations

Fig. 4.29: Learned Warping Paths **a** and **b**, Phase Four



(a) Learning Curve, CNN of $1/IP - 1/OP$ Architecture

(b) Learning Curve, CNN of $3/IP - 1/OP$ Architecture

Fig. 4.30: Mean Squared Error (Learning Curve), Phase Four

Algorithm 4.3 Whole Voice Transformation Process using Artificial Neural Network (ANN)

```
Input:
        Male speech information, $source\_speech$, $n = 600$ phrases
        Female speech information, $target\_speech$, $n = 600$ phrases
Output:
        Warped speech
Begin
        For $i = 1$ to $n$
            Function: Identifying Starting and ending points
                $source\_speech\_clip(i) = identify\_start\_end\_function(source\_speech(i))$
                $target\_speech\_clip(i) = identify\_start\_end\_function(target\_speech(i))$
            Function: Spectral feature extraction
                $\mathcal{S}_1(i) = spect\_function(source\_speech\_clip(i))$
                $\mathcal{S}_2(i) = spect\_function(target\_speech\_clip(i))$
            Function: Algorithm 4.1
                $[(\mathbf{a}_T, \mathbf{b}_T), (\mathbf{a}_F, \mathbf{b}_F)](i) = dtw\_algorithm(\mathcal{S}_2(i) = target\_spect, \mathcal{S}_1(i) = source\_spect)$
            Function: Time Alignment step
                $\mathcal{S}_{1_T}(:, \mathbf{a}_T)(i) = \mathcal{S}_1(:, \mathbf{b}_T)(i)$
        end
        Finding: $max_{index}$ in whole $\mathbf{a}_F$ and $\mathbf{b}_F$
        Interpolation Step: $\tilde{\mathbf{a}}$ and $\tilde{\mathbf{b}}$
        Neural Net Step:
            ANN Input: $\mathcal{S}_{1_T}$
            ANN Output: $\tilde{\mathbf{a}}_F$ and $\tilde{\mathbf{b}}_F$
            ANN training Output: predict $\tilde{\mathbf{a}}_{F,estimated}$ and $\tilde{\mathbf{b}}_{F,estimated}$ for one phrase
            Re-interpolation step: $\tilde{\mathbf{a}}_{F,estimated,revs}$ and $\tilde{\mathbf{b}}_{F,estimated,revs}$
        Function: Spectral Alignment step
                For $i_F = 1$ to $size((\mathcal{S}_{1_T}), 2)$
                    $\mathcal{S}_{1_{TF,ANN}}(i, \tilde{\mathbf{a}}_{F,estimated,revs}) = \mathcal{S}_{1_T}(i, \tilde{\mathbf{b}}_{F,estimated,revs})$
                End $i_F$
        Function: Phase Reconstruction Step
                $warped\_speech = phase\_algorithm(\mathcal{S}_{1_{TF,ANN}})$
        Play $warped\_speech$
```

CHAPTER 5

CONCLUSION AND FUTURE WORKS

5.1 Conclusion

In this Dissertation, a new two-level dynamic warping algorithm is presented, where an outer-level warping process does temporal alignment (Dynamic Time Warping, DTW), which temporally aligns block of features to compensate for tempo differences such as different speech rate. This outer-level warping process invokes an inner-level warping process (Dynamic Frequency Warping, DFW) to achieve spectral alignment based on spectral information of block of features to reduce or eliminate the spectral variations of the speech and to compensate for speaker differences in the spectral domain. After careful examination of the literature, it is clear that this DW has not been previously developed. This two level dynamic warping is applied in this dissertation to two applications. In the first application, we applied this algorithm to a dysarthric speech. The two-level DW algorithm used in a training tool to compare the production of a dysarthric speech with the imitation attempt of a healthy speaker that eventually be used to provide the learner with real-time feedback regarding the accuracy of their imitation attempts during training. The second application is voice transformation. Voice transformation is achieved using this algorithm, where the outer level temporally align blocks of speech invokes an inner warping process, which spectrally aligns based on magnitude spectra. This process of voice transformation involves only spectral magnitudes information, and has been found to introduce significant deleterious signal processing artifacts. To avoid this issue, phase reconstruction a;algorithm was used to improve the quality of the transformed speech. In summary, the following is a list of contributions of each chapter in this dissertation.

- Chapter 2

  - A new two-level dynamic warping algorithm was proposed to achieve warping process on both temporal domain and spectral domain.

- Chapter 3
  - The warping algorithm was used as a training tool that eventually be used to provide real time feedback regarding the accuracy of the imitation process.
  - A clinical experiment was performed on the speech data to determine if the speech feature vectors and dynamic warping (DW) are able to distinguish between healthy subjects reading a phrase in their ”own voice” and healthy subjects imitating that same phrase produced by a speaker with dysarthria.

- Chapter 4
  - The warping algorithm was used to a achieve voice transformation.
  - Applying phase reconstruction algorithm to the transformed voice to improve the quality of the voice.
  - Artificial neural network was applied to train the spectral information obtained from dynamic warping algorithm to assist in voice transformation.

5.2 Future Work

5.2.1 Dysarthric Imitation Application

The clinical test performed in chapter 2 was consider as an initial step in the development and evaluation of the proposed learning tool. This study is preparatory to the longer-term objectives of this research, which is to determine if training with a tool which provides visual feedback about the accuracy of an imitation attempt is able to improve a listeners ability to understand dysarthric speech. In other words, does this tool assist with imitation accuracy and does this tool elevate intelligibility improvements relative to imitation only?. Successful demonstration of intelligibility may lead to clinical tools that may find widespread use.

5.2.2 Voice Transformation Application

The voice transformation achieved in this dissertation focus on finding a mapping function form the spectral information of the source and target speakers. Future work will focus on how to improve the mapping by exploring and eliminating the causes of the signal processing artifacts.

This specific transformation done here is the most traditional case of voice transformation when the source speaker and target speaker are speaking the same language. Another item of interest as a future work is whether this warping method can be used to produce different language accents or if this method can be able to do cross lingual voice transformation, which means when the source speaker and the target speaker are spoke different languages.

REFERENCES

[1] G. A. Ten Holt, M. J. Reinders, and E. Hendriks, "Multi-dimensional dynamic time warping for gesture recognition," in *Thirteenth Annual Conference of the Advanced School for Computing and Imaging*, vol. 300, 2007, p. 1.

[2] J. R. Deller, J. G. Proakis, and J. H. Hansen, *Discrete-time processing of speech signals.* Macmillan Publishing Company, 2000.

[3] T. W. Parsons, *Voice and speech processing.* McGraw-Hill (New York [ua]), 1987.

[4] P. Mardziel, "Improved two-dimensional warping," *Worcester Polytechnic Institute*, Aug. 2005.

[5] E. P. Neuburg, "Dynamic frequency warping, the dual of dynamic time warping," *The Journal of the Acoustical Society of America*, vol. 81, no. S1, pp. S94–S94, 1987.

[6] D. J. Berndt and J. Clifford, "Finding patterns in time series: A dynamic programming approach," in *Advances in Knowledge Discovery and Data Mining.* American Association for Artificial Intelligence, 1996, p. 229–248.

[7] Z. Huang and L. Hou, "Speaker normalization using dynamic frequency warping," in *2008 International Conference on Audio, Language and Image Processing.* IEEE, 2008, pp. 1091–1095.

[8] K. Paliwal and W. Ainsworth, "Dynamic frequency warping for speaker adaptation in automatic speech recognition," *Journal of Phonetics*, vol. 13, no. 2, pp. 123–134, 1985.

[9] J. Black, "Multiple-choice intelligibility tests," *Journal of Speech and Hearing Disorders*, vol. 22, pp. 213–235, Jun. 1957.

[10] S. A. Borrie, M. Shäfer, and J. M. Liss, "Perceptual learning of dysarthric speech: A review of experimental studies," *Journal of Speech, Language, and Hearing Research*, vol. 55, pp. 290–305, Feb. 2012.

[11] S. A. Borrie, M. J. McAuliffe, J. M. Liss, G. A. O'Beirne, and T. Anderson, "The role of linguistic and indexical information in improved recognition of dysarthric speech," *Journal of Acoustical Society of America*, vol. 133, pp. 474–482, Jan. 2013.

[12] S. A. Borrie and M. Shäfer, "The role of somatosensory information in speech perception: Imitation improves recognition of disordered speech," *Journal of Speech, Language, and Hearing Research*, vol. 58, pp. 1708–1716, Dec. 2015.

[13] K. Lansford, S. Borrie, and L. Bystricky, "Use of crowdsourcing to asses the ecological volidity of perceptual training in dysarthria," *American Journal of Speech-Language Pathology*, vol. 25, pp. 233–239, May 2016.

[14] S. A. Borrie and M. Shäfer, "Effects of lexical and somatosensory feedback on long-term improvements in intelligibility of dysarthric speech," *Journal of Speech, Language, and Hearing Research*, vol. 60, pp. 2151–2158, Aug. 2017.

[15] E. Helander, J. Schwarz, J. Nurminen, H. Silen, and M. Gabbouj, "On the impact of alignment on voice conversion performance," *International Speech Communication Association*, vol. 1, pp. 1453–1456, Sep. 2008.

[16] A.-W. Al-Dulaimi, S. Budge, S. A. Borrie, T. K. Moon, and J. H. Gunther, "A tool for training speech imitation accuracy," in *2018 52nd Asilomar Conference on Signals, Systems, and Computers.* IEEE, 2018, pp. 1086–1090.

[17] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, pp. 43–49, Feb. 1978.

[18] E. Hsu, K. Pulli, and J. Popovic, "Style translation for human motion," *ACM Transactions on Graphics*, vol. 24, pp. 1082–1089, Jul. 2005.

[19] K. Kulkarni, G. Evangelidis, J. Cech, and R. Horaud, "Continuous action recognition based on sequence alignment," *International Journal of Computer Vision*, pp. 1–32, Jun. 2014.

[20] J. Zhao and L. Itti, "shape DTW: Shape dynamic time warping," *Pattern Recognition*, vol. 74, pp. 171–184, 2018.

[21] E. Keogh and C. A. Ratanamahatana, "Exact indexing of dynamic time warping," *Knowledge and information systems*, vol. 7, no. 3, pp. 358–386, 2005.

[22] M. Vlachos, M. Hadjieleftheriou, D. Gunopulos, and E. Keogh, "Indexing multi-dimensional time-series with support for multiple distance measures," in *Proceedings of the ninth international conference on Knowledge discovery and data mining*, 2003, pp. 216–225.

[23] Z. M. Kovacs-Vajna, "A fingerprint verification system based on triangular matching and dynamic time warping," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1266–1276, 2000.

[24] T. M. Rath and R. Manmatha, "Word image matching using dynamic time warping," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, 2003.

[25] S. Chu, E. Keogh, D. Hart, and M. Pazzani, "Iterative deepening dynamic time warping for time series," in *Proceedings of the 2002 SIAM International Conference on Data Mining*, 2002, pp. 195–212.

[26] M. Shokoohi-Yekta, J. Wang, and E. Keogh, "On the non-trivial generalization of dynamic time warping to the multi-dimensional case," in *Proceedings of the 2015 SIAM international conference on data mining*, 2015, pp. 289–297.

[27] P. Papapetrou, V. Athitsos, M. Potamias, G. Kollios, and D. Gunopulos, “Embedding-based subsequence matching in time-series databases,” *ACM Transactions on Database Systems*, vol. 36, no. 3, pp. 1–39, 2011.

[28] H. Ding, G. Trajcevski, P. Scheuermann, X. Wang, and E. Keogh, “Querying and mining of time series data: experimental comparison of representations and distance measures,” *Proceedings of the VLDB Endowment*, vol. 1, no. 2, pp. 1542–1552, 2008.

[29] E. J. Keogh and M. J. Pazzani, “Derivative dynamic time warping,” in *Proceedings of the 2001 SIAM international conference on data mining*, 2001, pp. 1–11.

[30] S. Salvador and P. Chan, “Toward accurate dynamic time warping in linear time and space,” *Intelligent Data Analysis*, vol. 11, no. 5, pp. 561–580, 2007.

[31] Z. Huang and L. Hou, “Speaker normalization using dynamic frequency warping,” in *IEEE International Conference on Audio, Language and Image Processing*, 2008, pp. 1091–1095.

[32] T. Toda, H. Saruwatari, and K. Shikano, “Voice conversion algorithm based on gaussian mixture model with dynamic frequency warping of straight spectrum,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings*, vol. 2, 2001, pp. 841–844.

[33] Y. Shou, N. Mamoulis, and D. W. Cheung, “Fast and exact warping of time series using adaptive segmental approximations,” *Machine Learning*, vol. 58, no. 2-3, pp. 231–267, 2005.

[34] J. R. Duffy, *Motor Speech Disorders: Substrates, Differential Diagnosis, and Management*, 3rd ed. St. Louis, MO: Elsevier Mosby, 2013.

[35] S. A. Borrie, M. J. McAuliffe, J. M. Liss, C. Kirk, G. A. O’Beirne, and T. Anderson, “Familiarisation conditions and the mechanisms that underlie improved recognition of dysarthric speech,” *Language and Cognitive Processes*, 2012.

[36] Z. Z. Wu, “Spectral mapping for voice conversion,” Ph.D. dissertation, Nanyang Technological University, Singapore, 2015.

[37] F. Rudzicz, “Production knowledge in the recognition of dysarthric speech. phd thesis,” *Department of Computer Science, University of Toronto*, 2011.

[38] J. L. Flanagan, *Speech analysis synthesis and perception.* Springer Science & Business Media, 2013, vol. 3.

[39] J. B. Allen and L. R. Rabiner, “A unified approach to short-time Fourier analysis and synthesis,” *Proceedings of the IEEE*, vol. 65, no. 11, pp. 1558–1564, 1977.

[40] J. Slifka and T. R. Anderson, “Speaker modification with LPC pole analysis,” vol. 1, 1995, pp. 644–647.

[41] M. Narendranath, H. A. Murthy, S. Rajendran, and B. Yegnanarayana, “Transformation of formants for voice conversion using artificial neural networks,” *Speech Communication*, vol. 16, pp. 207–216, Nov. 1995.

[42] O. Türk, O. Büyük, A. Haznedaroglu, and L. M. Arslan, “Application of voice conversion for cross-language rap singing transformation,” *International Conference on Acoustics, Speech and Signal Processing*, pp. 3597–3600, May 2009.

[43] Y. Stylianou, “Voice transformation: a survey,” *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3585–3588, May 2009.

[44] M. V. Ramos, “Voice conversion with deep learning,” *Tecnico Lisboa Masters thesis*, 2016.

[45] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, “Voice conversion through vector quantization,” *Journal of the Acoustical Society of Japan*, vol. 11, no. 2, pp. 71–76, 1990.

[46] H. Kuwabara, “Quality control of speech by modifying formant frequencies and bandwidth,” *11th Int. Congr. Phonetic Science*, pp. 281–284, 1987.

[47] D. Childers, B. Yegnanarayana, and K. Wu, “Voice conversion: Factors responsible for quality,” in *ICASSP’85. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 10, 1985, pp. 748–751.

[48] H. Valbret, E. Moulines, and J.-P. Tubach, “Voice transformation using PSOLA technique,” *Speech communication*, vol. 11, no. 2-3, pp. 175–187, 1992.

[49] H. Matsumoto and H. Wakita, “Vowel normalization by frequency warped spectral matching,” *Speech Communication*, vol. 5, no. 2, pp. 239–251, 1986.

[50] 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*, vol. 20, no. 4, pp. 1313–1323, 2011.

[51] S. H. Mohammadi and A. Kain, “Transmutative voice conversion,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing.* IEEE, 2013, pp. 6920–6924.

[52] D. Erro, E. Navas, and I. Hernaez, “Parametric voice conversion based on bilinear frequency warping plus amplitude scaling,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 3, pp. 556–566, 2012.

[53] V. Popa, J. Nurminen, and M. Gabbouj, “A novel technique for voice conversion based on style and content decomposition with bilinear models,” in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[54] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification.* John Wiley & Sons, 2012.

[55] D. A. Reynolds and R. C. Rose, “Robust text-independent speaker identification using gaussian mixture speaker models,” *IEEE transactions on speech and audio processing*, vol. 3, no. 1, pp. 72–83, 1995.

[56] R. C. Rose and D. A. Reynolds, "Text independent speaker identification using automatic acoustic segmentation," in *International Conference on Acoustics, Speech, and Signal Processing.* IEEE, 1990, pp. 293–296.

[57] 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*, vol. 15, no. 8, pp. 2222–2235, 2007.

[58] Y. Stylianou, O. Cappe, and E. Moulines, "Statistical methods for voice quality transformation," in *Fourth European Conference on Speech Communication and Technology*, 1995.

[59] A. R. Toth and A. W. Black, "Using articulatory position data in voice transformation," 2007, pp. 182–187.

[60] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98*, vol. 1, 1998, pp. 285–288.

[61] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on speech and audio processing*, vol. 6, no. 2, pp. 131–142, 1998.

[62] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj, "Voice conversion using partial least squares regression," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 912–921, 2010.

[63] A. O. Ayodeji and S. A. Oyetunji, "Voice conversion using coefficient mapping and neural network," in *IEEE International Conference for Students on Applied Engineering (ICSAE)*, 2016, pp. 479–483.

[64] L. Steels, "The artificial life roots of artificial intelligence," *Artificial life*, vol. 1, no. 1_2, pp. 75–110, 1993.

[65] M. Narendranath, H. A. Murthy, S. Rajendran, and B. Yegnanarayana, "Transformation of formants for voice conversion using artificial neural networks," *Speech communication*, vol. 16, no. 2, pp. 207–216, 1995.

[66] T. Watanabe, T. Murakami, M. Namba, T. Hoya, and Y. Ishida, "Transformation of spectral envelope for voice conversion based on radial basis function networks," in *Seventh international conference on spoken language processing*, 2002.

[67] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, "Voice conversion using artificial neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 3893–3896.

[68] 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, pp. 3052–3056.

[69] T. Nakashika, R. Takashima, T. Takiguchi, and Y. Ariki, “Voice conversion in high-order eigen space using deep belief nets.” in *Interspeech*, 2013, pp. 369–372.

[70] A. Kain and M. W. Macon, “Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing.*, vol. 2, 2001, pp. 813–816.

[71] J. Kominek and A. W. Black, “The CMU Arctic speech databases,” in *Fifth ISCA workshop on speech synthesis*, 2004.

[72] I. Saratxaga, I. Hernaez, M. Pucher, E. Navas, and I. Sainz, “Perceptual importance of the phase related information in speech,” in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[73] M. Tong, Z. Bian, X. Li, Q. Dai, and Y. Chen, “Study on phase perception in speech,” *Journal of Electronics*, vol. 20, no. 5, pp. 387–392, 2003.

[74] D.-S. Kim, “Perceptual phase redundancy in speech,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings*, vol. 3, 2000, pp. 1383–1386.

[75] J. Skoglund, W. B. Kleijn, and P. Hedelin, “Audibility of pitch-synchronously modulated noise,” in *IEEE Workshop on Speech Coding for Telecommunications Proceedings. Back to Basics: Attacking Fundamental Problems in Speech Coding*, 1997, pp. 51–52.

[76] D. Griffin and J. Lim, “Signal estimation from modified short-time Fourier transform,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[77] Z. Prusa, “The phase retrieval toolbox,” in *Audio Engineering Society Conference: 2017 AES International Conference on Semantic Audio*, 2017.

[78] B. Sisman, J. Yamagishi, S. King, and H. Li, “An overview of voice conversion and its challenges: From statistical modeling to deep learning,” *arXiv preprint arXiv:2008.03648*, 2020.

[79] N. Embretsén, “Representing voices using convolutional neural network embeddings,” 2019.

[80] W. S. McCulloch and W. Pitts, “A logical calculus of the ideas immanent in nervous activity,” *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115–133, 1943.

[81] D. O. Hebb, *The organization of behavior: A neuropsychological theory.* Psychology Press, 2005.

[82] L. Sun, S. Kang, K. Li, and H. Meng, “Voice conversion using deep bidirectional long short-term memory based recurrent neural networks,” in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP).* IEEE, 2015, pp. 4869–4873.

CURRICULUM VITAE

Al-Waled H. Al-Dulaimi

Carrier Objective

To obtain a university professor position at a public university in Iraq that will enable the use of technical expertise, leadership, and communication, computer, technical and teaching skills.

Education

- BSc in Electrical Engineering, Al-Nahrain University 2000.
- MSc in Electrical Engineering, Al-Nahrain University 2002.

Awards and Honors

- Student engineering second award, second place, provided by Utah Valley University with a generous award in the amount of $200 (Oct. 2020).
- Awarded by Rajendra Prasad and Yasoda Mani Grandhi Endowed Fellowship with a generous amount of $700 (May 2020).
- Awarded travel grant, generously provided by Research and Graduate Studies (RGS) of Utah State University in the amount of $300 to attend and present conference papers (Aug. 2018 and Sep. 2019).
- Outstanding teaching assistant, Utah State University, 2018.
- Student of the semester, Intensive English Language Institute (IELI), Utah State University, Spring 2015.

- Student of the year, Intensive English Language Institute (IELI), Utah State University, 2015.
- A letter of Thanks. By his Excellency the Ministry of Higher Education and Scientific Research, December 2013, Baghdad, Iraq.
- A letter of Thanks. By his Excellency the Ministry of Higher Education and Scientific Research, October 2013, Baghdad, Iraq.
- A letter of Thanks. Senior Deputy/Ministry of Higher Education and Scientific Research, February 2012, Baghdad, Iraq.
- A letter of Thanks. Department of Research and Development. Ministry of Higher Education and Scientific Research, March 2011, Baghdad, Iraq.

Teaching Experience

- Instructor: Mathematical methods for signals and system (course Num. ECE 6030), ECE Dept., Utah State University (USU), Logan, Utah (Spring 2017).
- Teaching Assistant:
  - Deep learning neural networks, Under the Supervision of Dr. Todd Moon, Utah State University (USU), Logan, Utah (Fall 2017).
  - Stochastic Processes, Under the Supervision of Dr. Todd Moon, Utah State University (USU), Logan, Utah (Fall 2017, Fall 2018, Fall 2020).
  - Communication Systems I, Under the Supervision of Dr. Todd Moon, Utah State University (USU), Logan, Utah (Spring 2018, Spring 2020).
  - Mathematical methods for signals and system, Under the Supervision of Dr. Todd Moon, Utah State University (USU), Logan, Utah (Spring 2018, Spring 2019, Spring 2020).
  - Digital Signal and Image Processing, Under the Supervision of Dr. Scott Budge, Utah State University (USU), Logan, Utah (Fall 2018, Fall 2019, Fall 2020).

  - Detection and Estimation Theory, Under the Supervision of Dr. Todd Moon, Utah State University (USU), Logan, Utah (Fall 2018, Fall 2020).
  - Convex Optimization, Under the Supervision of Dr. Jacob Gunther, Utah State University (USU), Logan, Utah (Fall 2018).
  - Real-Time Processors, Under the Supervision of Dr. Scott Budge, Utah State University (USU), Logan, Utah (Spring 2019, Spring 2020).
  - Continuous-Time Systems and Signals, Under the Supervision of Dr. Jacob Gunther, Utah State University (USU), Logan, Utah (Fall 2019, Fall 2020).
  - Discrete-Time Signals and Systems, Under the Supervision of Dr. Jacob Gunther, Utah State University (USU), Logan, Utah (Spring 2020).

- Lecturer:
  - Electronic Communication 1. College of Engineering, Diyala University, Iraq, Spring 2003.
  - Logic Circuits. College of Engineering, Diyala University, Iraq, Spring 2004.
  - Digital Electronics. College of Engineering, Diyala University, Iraq, Spring 2005.

Work Experience

- E-Government department/ Reconstruction and Projects Directorate. Ministry of Higher Education and Scientific Research. Head of E-Government department. Baghdad, Iraq, August 2010 to August 2014.
- Korek company-GSM Mobile Telecommunications Network. Senior radio frequency (RF) planning engineer. Baghdad, Iraq, January 2010 to July 2010.
- Shabakkat company-GSM Mobile Telecommunications Network. Senior radio frequency (RF) Planning Engineer Baghdad, Iraq, April 2009 to December 2009.
- Zain-GSM Mobile Telecommunications Network. Baghdad, Iraq, April 2005 to April 2009.

- RF team leader, August 2008 to April 2009.
- Senior RF planning engineer, June 2007 to August 2008.
- Junior RF planning engineer, April, 2005 to June 2007.

Published Conference Papers

- A.-W. Al-Dulaimi, T. K. Budge, and J. H. Gunther, “Phase effects on speech and its influence on warped speech,” IETC 2020, 2020. Accepted.
- A.-W. Al-Dulaimi, T. K. Moon, and J. H. Gunther, “Voice transformation using two-level dynamic warping,” in 2019 53nd Asilomar Conference on Signals, Systems, and Computers, 2019. Accepted.
- A.-W. Al-Dulaimi, S. Budge, S. A. Borrie, T. K. Moon, and J. H. Gunther, “A tool for training speech imitation accuracy,” in 2018 52nd Asilomar Conference on Signals, Systems, and Computers, pp. 1086–1090, IEEE, 2018. Accepted.