VOICE CONVERSION FOR PERSONS WITH AMYOTROPHIC LATERAL SCLEROSIS

Yunxin Zhao, Mili Kuruvilla-Dugdale, and Minguang Song

***Abstract*— Amyotrophic lateral sclerosis (ALS) results in progressive paralysis of voluntary muscles throughout the body. As speech deteriorates, individuals rely on pre-programmed messages available on commercial speech generating devices to communicate using one of the generic electronic voices on the device. To replace these generic voices and restore vocal identity, our aim was to develop personalized voices for people with ALS via the approach of voice conversion. The task is challenging because very few people have large quantities of their premorbid healthy speech recorded. Therefore, we have to rely on small quantities of dysarthric speech concomitant with an individual's disease stage. Further, progressive fatigue prohibits acquisition of large speech datasets and individuals display a range of dysarthria severities resulting from breathing, voice, articulation, resonance, and prosody disturbances. As the first step to address these problems, we used healthy source speakers and proposed the approach of combining a sparse structured spectral transform with multiple linear regression-based frequency warping prediction for spectral conversion, and interpolating the transformed spectral frames for speech rate modification. Our experimental data included four healthy source speakers from the ARCTIC dataset, and four target ALS speakers with mild to severe dysarthria, forming 16 speaker pairs. Subjective listening evaluations showed that on average, (i) the proposed approach improved speech intelligibility by about 80% over the target speakers' speech, (ii) the converted voice was 3 times more similar to the target speakers' speech than to the source speakers' speech, and (iii) the converted speech quality was close to the MOS scale "good" relative to the source speakers' speech being "excellent."**

***Index Terms*— voice conversion; dysarthria; amyotrophic lateral sclerosis; assistive speech devices**

I. INTRODUCTION

Amyotrophic lateral sclerosis (ALS) is the most common motor neuron disorder worldwide [1] and is characterized by the loss of motor neurons that innervate muscles of speech, swallowing, as well as trunk and limb movement. Typically, individuals only survive 3-5 years after diagnosis and are faced with progressive speech loss, which requires them to rely on speech generating devices for communication. Commercial devices with text-to-speech (TTS) output allow individuals with ALS to communicate verbally [2] and the text is accessed using functioning muscles, such as eye muscles, which usually remain functional till the end of patients' lives [3]. A variety of software options exist for message formulation, such as spelling letter-by-letter, selecting individual words, and choosing partial and/or full messages from a pre-programmed display. Once the message is formulated, the speech output option allows the message to be heard aloud in one of the electronic voices available on the device.

At present, the TTS voices used in speech devices are generic and impersonal. Using synthetic speech that sounds different from the person with ALS can be embarrassing and cause social isolation [4]. It is often infeasible to train a high-quality TTS system for a person with ALS because 1) recorded samples of premorbid healthy speech are usually sparse, 2) large quantities of speech are difficult to record post-diagnosis due to patients' progressive fatigue, and 3) dysarthric impairments in ALS speech interfere with the intelligibility of synthesized speech.

To address these problems, one approach is to collect speech data from a large pool of healthy speakers, to find a good voice match that can be used to build a proxy TTS model for a person with ALS [4]. This requires extensive data collection, and it is difficult to find a well-matched voice due to the complex vocal characteristics that make a voice unique. Another option is to record messages in the person's own voice that can later be uploaded to their speech device and accessed directly via touch or eye gaze for communication. Creating personalized messages using this approach is time and effort intensive, and these recordings need to be completed early on in the disease process before an individual's speech starts to deteriorate.

In the current work, we investigate a novel approach of using voice conversion (VC) to generate personalized voices for speakers with ALS. VC modifies spectral and prosodic features of a source speaker's speech to resynthesize speech that sounds like the voice of a target speaker. In our study, the source speakers are healthy, while the target speakers with ALS display mild to severe dysarthria with varying levels of impairment in breathing, voice, articulation, resonance, and prosody. At this stage of study, using healthy speech instead of TTS as the source allows us to focus on three key issues in VC for ALS speech, that is, to give converted speech (i) intelligibility for comprehension, (ii) good quality for ease of listening and (iii) similarity to the target speaker for voice identity. This choice of speech source will generate fewer outcome variations than a TTS source, which allows for intelligibility, quality, and similarity evaluations to be carried out by a feasible number of listeners. In contrast, investigating

Y. Zhao and M. Song are with the Department of Electrical Engineering and Computer Science, M. Kuruvilla-Dugdale is with the Department of Speech, Language & Hearing Sciences, University of Missouri, Columbia MO 65211 USA(e-mails: zhaoy@missouri.edu, msong@mail.missouri.edu, kuruvillam@health.missouri.edu). This work was supported in part by Missouri Spinal Cord Injury/Disease Research Program. Dr. Kuruvilla-Dugdale's effort was partly supported by the National Institutes of Health [R15 DC016383].

VC together with various TTS methods and their training data would create large outcome variations, placing a significant burden on listeners and resources for subjective evaluations. Messages generated using our VC method could be directly used in speech generating devices instead of pre-programmed messages and would significantly expand the personalized message sets. In fact, ALS users show a preference for communicating via full phrases or sentences to circumvent problems with letter-by-letter spelling when using eye gaze [3]. That said, our VC method can be readily applied to the TTS source in the next stage of study, especially as the quality of deep learning-based TTS is now approaching the quality of human speech (single voice in read style) [5].

Although different from the problem we address here, for summaries about VC between healthy speakers, please refer to the VC challenges of 2016 and 2018 [6,7]. For VC involving speakers with speech disabilities, only a few efforts are reported pertaining to voice personalization. One study used statistical Eigenvoice conversion to convert alaryngeal source speech (esophageal, electrolaryngeal, or body-conducted silent electrolaryngeal speech) to healthy target speech and improve voice individuality of laryngectomees [8]. In another study researchers first constructed a spectral dictionary consisting of the vowels of a source speaker with cerebral palsy and the consonants of a healthy target speaker, and then used exemplar-based VC to produce speech with improved intelligibility [9]. Speech parameter modification has also been studied to improve intelligibility of dysarthric speech due to Friedreich's ataxia [10], where vowel duration and formants from a dysarthric speaker were modified to match those of a non-dysarthric speaker. In a recent deep-learning based VC [11], a generative adversarial network was used to improve speech intelligibility of one patient post-orofacial surgery. The method was shown to improve articulation clarity and preserve the patient's voice characteristics to some extent, but not fully retain the speech linguistic content.

Considering the issues of dysarthria and limited speech data, we opt to adapt our recent VC method, sparse structured spectral transform (SSST) [12], to the current task for ALS speech. SSST learns simultaneous frequency warping and spectral shaping on high-dimensional STRAIGHT spectra [13] from a small amount of parallel source-target speech. In our previous study that involved 12 pairs of healthy speakers from the ARCTIC dataset [14], SSST was judged favorably in speech quality and voice similarity in comparison with three other established VC approaches: generic exemplar [12], Gaussian mixture model (GMM), and GMM with global variance maximization [15]. Unlike in [8,9] where the source speakers had speech impairments, the source speakers in the current study are healthy and the target speakers have ALS. Even though our target speech is impaired, SSST can largely maintain the intelligibility of the source speech because the spectral transforms are defined for broad phonetic classes instead of individual phonetic units. Considering that ALS speech as well as speech from the elderly generally have a slower-than-normal speech rate [16], we also investigate speech rate modification via interpolations on the converted spectral frames. The rationale is that modifying speech rate to a reasonable extent to match that of the target ALS speaker may help enhance voice similarity. In general, speech rate conversion is an essential component of prosody conversion in VC, but it has not gained as much attention as the other conversion aspects because the speech rates of healthy speakers used in VC are often similar. We used the STRAIGHT vocoder [13] to generate speech waveforms after the spectral and prosody conversion. Our approach has been evaluated on 4 source ARCTIC speakers and 4 target speakers with ALS, i.e., 16 source-target speaker pairs. Subjective listening evaluations were performed using transcription-based intelligibility, goodness measures of quality, and voice similarity to the target.

Our current work contributes to voice personalization for ALS users in three aspects: conceptual, technical, and experimental. *Conceptually*, our approach departs from the commonly held belief that the target speech must be intelligible in order for the voice-converted speech to be intelligible. Our insight is that performing VC at a coarse spectral level can strike a good balance between capturing voice similarity and preserving speech intelligibility. *Technically*, the level of conversion can be controlled by the number of structured sparse spectral transforms (SSST) [12] that are integrated probabilistically. SSST was developed for healthy-to-healthy VC, and when used for healthy-to-ALS VC, it is difficult to estimate the source-to-target frequency warping parameter if the dysarthria is severe. To overcome this difficulty, we proposed a novel approach to predict the frequency warping parameter from the $\log(F_0)$s of the source and target speakers by learning a multiple linear regression function from healthy-to-healthy VC, and then transferring the regression function to the healthy-to-ALS VC task. We further developed a novel frame interpolation method for speech rate conversion that is often ignored in healthy-to-healthy VC, but it is important for capturing ALS voice characteristics due to the speakers' speech rate reductions. *Experimentally*, using our own patient pool, we collected ALS speech data that cover the full range of dysarthria severity from mild to severe, and conducted comprehensive subjective tests on naturalness, similarity to target, and intelligibility. To the best of our knowledge, this is the first study of this kind.

In Section II, the SSST-based VC method [12] is summarized. In Section III, the multiple linear regression based frequency warping prediction and the frame interpolation based speech rate conversion are described. In Section IV, experiment setup and evaluation results are discussed. In Section V, conclusions are made.

II. SSST-BASED VOICE CONVERSION

Spectral conversion involves performing frequency warping and spectral shaping on a source speaker's speech to match the characteristics of a target speaker's speech. Denote a magnitude (envelope) spectrum of a source by $\mathbf{x} = [x_1 \cdots x_d]^T$ and that of a target by $\mathbf{y} = [y_1 \cdots y_d]^T$, with d the number of features. The conversion can be expressed as

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2019.2961844, IEEE Journal of Biomedical and Health Informatics

$$\hat{y}_i = \sum_{j:\ \varphi_\alpha(j)=i} w_{\varphi_\alpha(j),j} x_j, \qquad i = 1, \cdots, d, \tag{1}$$

where $\varphi_\alpha(.)$ warps the *j*-th frequency point of the source to the *i*-th frequency point of the target, and $w_{\phi_\alpha(j),j}$ scales the contribution of x_j to $\hat{y}_i$. In [12], Eq. (1) is implemented by a linear transform $\hat{\mathbf{y}} = \mathbf{W}\mathbf{x}$, with $\mathbf{W} = \left[w_{i,j} \right]_{d \times d}$ a sparse, nonnegative matrix that allows only relevant source spectral components to contribute to a target spectral component.

A. Sparse transform estimation

In SSST, a sparse transform is realized by embedding a region-of-support (ROS) in $\mathbf{W}$. The ROS is formed according to a source-to-target frequency warping constraint for each speaker pair, where an ROS includes the feasible warping points $\{(i, j)\}$ and excludes the infeasible points, with the range of feasibility defined by the physical anatomy of human vocal tracts. The elements of $\mathbf{W}$ within the ROS are initialized to ones and those outside are initialized to zeros. $\mathbf{W}$ is iteratively optimized by the multiplicative parameter update algorithm (MPUA) of nonnegative matrix factorization (NMF) [17] to refine the weights within the ROS for spectral shaping [12].

A bilinear function [18] is used in [12] to define frequency warping paths P_α:

$$\varphi_\alpha(\omega) = \omega + 2\tan^{-1}\left(\frac{(1-\alpha)\sin(\omega)}{1-(1-\alpha)\cos(\omega)} \right) \tag{2}$$

where ω is angular frequency, α is a warping parameter, with $\alpha < 1$ warps frequency from low to high as in male-to-female conversion, and $\alpha > 1$ warps frequency from high to low as in female-to-male conversion. By quantizing α within a feasible range $[\alpha_{\min}, \alpha_{\max}]$ with a step $\Delta\alpha$, a set of warping paths can be enumerated. The source-target frequency pairs on P_α form a tight ROS, and the so initialized transform is denoted by $\mathbf{W}_\alpha^{(0)}$. Given a pair of speakers with their temporally aligned spectral matrices $\mathbf{D_A}$ and $\mathbf{D_B}$, α is optimized by

$$\alpha^* = \arg\min_\alpha \| \mathbf{D}_B - \hat{\mathbf{W}}_\alpha \mathbf{D}_A \|_F \tag{3}$$

with $\|.\|_F$ the Frobenius norm, and $\hat{\mathbf{W}}_\alpha$ as estimated by MPUA. Based on α^*, an ROS $\Omega(\alpha^*)$ is formed by multiple P_α's, $\alpha \in \left[\alpha^* - \Delta\alpha,\ \alpha^* + \Delta\alpha \right]$, to initialize $\mathbf{W}_{\Omega(\alpha^*)}^{(0)}$, and MPUA is applied again to obtain the optimized transform $\hat{\mathbf{W}}_{\Omega(\alpha^*)}$. In the conversion stage, the source spectral matrix $\mathbf{X}$ is multiplied by $\hat{\mathbf{W}}_{\Omega(\alpha^*)}$ to produce the converted spectral matrix $\hat{\mathbf{Y}}$.

B. Mixture sparse spectral transform

In [12], SSST includes a mixture of transforms to model the spectral mapping properties for different phonetic classes implicitly, and the ROS initialization is applied to all the transforms in a mixture. To do so, the source mel frequency cepstral coefficients (MFCC) features are used to estimate a Gaussian mixture model (GMM), and the mixture clustering structure is transferred to the temporally aligned magnitude spectral matrices $\mathbf{D}_A$ and $\mathbf{D}_B$ to obtain *M* pairs of matrices $\left\{ \left(\mathbf{D}_{A,m}, \mathbf{D}_{B,m} \right), m = 1, \cdots, M \right\}$, which are used to derive mixture-specific transforms $\hat{\mathbf{W}}_{\Omega(\alpha^*),\mathbf{m}}, m = 1, \cdots, M$. At the conversion stage, the posterior probabilities of the *m*-th mixture given the *t*-th MFCC frame, $\gamma_{m,t}$, provide data-dependent weights on $\hat{\mathbf{W}}_{\Omega(\alpha^*),\mathbf{m}}$ to convert $\mathbf{x_t}$ of the source to $\hat{\mathbf{y}}_\mathbf{t}$ of the target in the magnitude spectral domain:

$$\hat{\mathbf{y}}_\mathbf{t} = \sum_{m=1}^{M} \gamma_{m,t} \hat{\mathbf{W}}_{\Omega(\alpha^*),\mathbf{m}} \mathbf{x_t}. \tag{4}$$

C. Exemplar-based spectral conversion

As in [12], we include a probabilistic exemplar-based spectral conversion method for comparative evaluations. The considerations are threefold: 1) the exemplar VC approach performed better than the VC methods of GMM and GMM with global variance maximization in our study of [12], 2) it is also based on NMF as SSST, and 3) as the exemplars are target speech spectra, it may better preserve the voice characteristics of target speakers. By taking the GMM-clustered aligned data matrices $\left\{ \left(\mathbf{D}_{A,m}, \mathbf{D}_{B,m} \right), m = 1, \cdots, M \right\}$ as the exemplar dictionaries, a source spectral matrix $\mathbf{X}$ is factorized in *M* ways as $\mathbf{X} \approx \mathbf{D}_{A,m} \hat{\mathbf{H}}_m$, $m = 1, \cdots, M$, and the activations $\hat{\mathbf{H}}_m$ are transferred to the target dictionaries $\mathbf{D}_{B,m}$ to produce the converted spectral matrix

$$\hat{\mathbf{Y}} = \sum_{m=1}^{M} \mathbf{D}_{B,m} \hat{\mathbf{H}}_m \mathbf{P}_m \tag{5}$$

where $\mathbf{P_m} = \mathrm{diag}(\gamma_{m,1}, \cdots, \gamma_{m,N})$. It is worth noting that this exemplar method is only a generic one. By incorporating additional enhancements [19-20], such as contextual frames, exemplar-based VC performance may be improved, but they are beyond the scope of the current work.

Because the mixture SSST method (Eq. (4)) emphasizes the role of the transform function $\mathbf{W}$ in the first matrix factor of NMF, and the exemplar method (Eq.(5)) emphasizes the role of the activation function $\mathbf{H}$ in the second matrix factor of NMF, they are referred to as WNMF and HNMF, respectively in the subsequent discussions. It is worth noting that WNMF has a low computation complexity at the voice conversion stage, where for each speech utterance WNMF performs direct matrix multiplications instead of iterative NMF as required by HNMF.

III. Voice Conversion For Target Speakers With ALS

Estimating frequency warping parameters by the method of Eq. (3) becomes difficult when a target speaker's dysarthria is severe. In this section, we formulate our new approach based on multiple linear regression to predicting the parameters for target speakers with ALS, and describe our method of spectral frame interpolation for speech rate modification.

A. Spectral conversion

When the target speech dysarthria is severe, the Frobenius norm minimization in Eq. (3) may not result in a good match

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2019.2961844, IEEE Journal of Biomedical and Health Informatics

between the target and the transformed spectral matrices $\mathbf{D}_B$ and $\hat{\mathbf{W}}_\alpha \mathbf{D}_A$, rendering the estimated α^* erroneous. We propose a solution to this problem based on two considerations. First, the extent of frequency warping between two speakers is dependent on their F_0s. For example, the F_0 of a male speaker is in general lower than that of a female speaker, and for male-to-female spectral conversion we need to warp frequency in the low-to-high direction. Second, in dysarthric speech, although spectral details may be lost, voiced speech segments are still available to allow for pitch or F_0 estimations. Furthermore, by performing SSST on a pair of healthy speakers, we can estimate a warping parameter and obtain a 3-tuple sample {source F_0, target F_0, warping parameter α^*}, and from multiple healthy speaker pairs, we can generate a set of 3-tuple data samples. From this sample set, we can estimate a multiple linear regression (MLP) [21] function for predicting the warping parameter α^*. In this way, given the F_0s of a healthy source and an ALS target speaker pair, the warping parameter can be predicted without resorting to Eq. (3).

In our regression formulation, the $\log(F_0)$s of the source and target speakers are the explanatory variables, and the parameter α^* is the response variable. For the *q*-th pair source and target speakers, denote their average $\log(F_0)$s by $\bar{f}_{s,q}$ and $\bar{f}_{t,q}$, respectively, and their frequency warping parameter α^* by α_q (obtained by Eq. (3)). The regression function is then

$$\alpha_q \approx \beta_1 + \beta_2 \bar{f}_{s,q} + \beta_3 \bar{f}_{t,q}$$

Assume a total of *Q* source-target speaker pairs. Let $\mathbf{X}$ be the extended explanatory data matrix, with the *i*-th row defined by the $\log(F_0)$s of the *i*-th speaker pair, $\mathbf{y}$ be the response vector, with the *i*-th element defined by the predicted warping parameter for the *i*-th speaker pair, and $\boldsymbol{\beta}$ be the regression parameter vector:

$$\mathbf{X} = \begin{bmatrix} 1 & \bar{f}_{s,1} & \bar{f}_{t,1} \\ \vdots & \vdots & \vdots \\ 1 & \bar{f}_{s,Q} & \bar{f}_{t,Q} \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_Q \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

Minimizing the total squared regression error $\min_{\boldsymbol{\beta}} \| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \|^2$ gives the regression parameter $\hat{\boldsymbol{\beta}} = [\hat{\beta}_1\ \hat{\beta}_2\ \hat{\beta}_3]^T$ as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{6}$$

where the superscript ^{T} denotes vector or matrix transpose.

Given a novel pair of healthy-source and ALS-target speakers, with their average $\log(F_0)$s being $\bar{f}_s$ and $\bar{f}_t$, respectively, the warping parameter $\hat{\alpha}^*$ is predicted as:

$$\hat{\alpha}^* \approx \hat{\beta}_1 + \hat{\beta}_2 \bar{f}_s + \hat{\beta}_3 \bar{f}_t \tag{7}$$

Based on the predicted $\hat{\alpha}^*$, the ROS $\Omega(\hat{\alpha}^*)$ can then be formed for SSST as discussed in Section II.A.

B. Speech rate conversion

The converted speech rate is made to approximate a target speaker's average speech rate. The source speakers' speech rates are similar, and so an average rate is used as the source speech rate. To reduce speech rate, we time-stretch the converted STRAIGHT spectral sequence through spectral frame interpolation. This method is easy to implement, without the need for pitch synchronous time scale modification as in the approach of PSOLA [22].

Let the spectral sequence prior to the rate modification be, $\mathbf{Y}(t)$, $t = 1, \cdots, T$, where the index t represents the spectrum time positions normalized by the frame shift time. To stretch the duration of the spectral sequence by a factor of $\tau > 1$, we first reposition the $\mathbf{Y}$ spectra at $1, 1+\tau, \cdots, 1+(T-1)\tau$ to increase the gap between adjacent frames by a factor of τ, and then interpolate the spectra to produce a spectral sequence $\mathbf{Y}_{CR}(k)$, $k = 1, \cdots, K$, with $K = \lfloor 1+\tau(T-1) \rfloor$, at the original frame rate. This time-scale modification procedure is illustrated in Fig. 1 for $\tau = 1.5$ (discussed in details below). At the resampled position k, the spectrum $\mathbf{Y}_{CR}(k)$ is interpolated by the two $\mathbf{Y}$ spectra whose stretched positions are immediately before and after k, corresponding to the unstretched time positions $\tilde{t}_l(k)$ and $\tilde{t}_r(k)$:

$$\tilde{t}_l(k) = \lfloor 1+(k-1)/\tau \rfloor, \quad \tilde{t}_r(k) = \lceil 1+(k-1)/\tau \rceil$$

where $\lfloor\ \rfloor$ and $\lceil\ \rceil$ are the floor and ceiling operators, respectively. The computation for $\mathbf{Y}_{CR}(k)$, $k = 1, \cdots, K$, is

$$\mathbf{Y}_{CR}(k) = \lambda_k^l \mathbf{Y}(\tilde{t}_l(k)) + \lambda_k^r \mathbf{Y}(\tilde{t}_r(k)),$$
$$\text{with } \lambda_k^l \geq 0, \lambda_k^r \geq 0, \text{ and } \lambda_k^l + \lambda_k^r = 1.$$

The interpolation weights λ_k^l and λ_k^r are defined to be inversely proportional to the absolute difference of k to the two stretched positions $1+(\tilde{t}_l(k)-1)\tau$ and $1+(\tilde{t}_r(k)-1)\tau$, respectively, which gives $\lambda_k^r = (k-1-(\tilde{t}_l(k)-1)\tau)/\tau$ and $\lambda_k^l = 1-\lambda_k^r$. Note that when $\tilde{t}_l(k) = \tilde{t}_r(k)$, $\mathbf{Y}_{CR}(k) = \mathbf{Y}(\tilde{t}_l(k))$.

To synthesize the rate-modified speech, this interpolation procedure is also applied to pitch, voiced-unvoiced, and aperiodicity frame features. These rate-modified feature frames are provided to the STRAIGHT vocoder to generate the converted speech waveform.

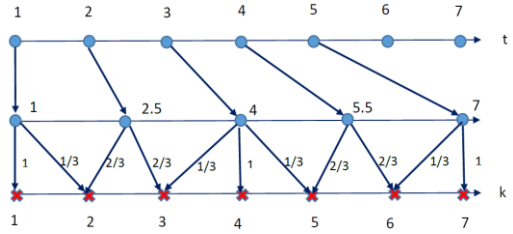

Fig. 1 *Spectral frame interpolation-based speech rate modification*

In Fig. 1, the top axis shows the positions of the original speech frames at one time unit apart, and the middle axis shows the mapped positions of these frames after time stretching. For example, the frame at the original positions 4 and 5 are mapped to the positions 5.5 (i.e., 1+3*1.5) and 7 (i.e., 1+4*1.5) after time stretching. The bottom axis shows the interpolated frame positions, again at one time unit apart. For example, the interpolated frame 6 is between the stretched frame positions 5.5 and 7 with their corresponding unstretched positions as $\tilde{t}_l(6) = 4$ and $\tilde{t}_r(6) = 5$. The interpolation weights for the frame 6 is then $\lambda_6^r = 1/3$ and $\lambda_6^l = 2/3$, and the interpolated spectrum

becomes $\mathbf{Y}_{CR}(k) = (2\mathbf{Y}(4) + \mathbf{Y}(5))/3$. On the other hand, the interpolated frame 7 aligns exactly with the stretched frame position 7 that has the unstretched position of $\tilde{t}_l(7) = \tilde{t}_r(7) = 5$, which gives $\lambda_7^r = 0$ and $\lambda_7^l = 1$, and thus $\mathbf{Y}_{CR}(7) = \mathbf{Y}(5)$.

C. Pitch conversion

Given a pair of source and target speakers, denote respectively their $log(F_0)$ frame values by f_s and f_t (frame indices are omitted), their sample means by $\bar{f}_s$ and $\bar{f}_t$, and their sample standard deviations by σ_s and σ_t. The source-to-target conversion on $log(F_0)s$ is defined by the linear transform

$$f_t = (f_s - \bar{f}_s)\sigma_t / \sigma_s + \bar{f}_t$$

which matches the converted $\log(F_0)'s$ mean and standard deviation with those of the target speaker [23].

D. Procedure

The above described multiple linear regression for ROS estimation and spectral frame interpolation for speech rate modification are integrated into the training and conversion stages of SSST for VC. A block diagram illustrating the spectral and speech rate conversion processes is shown in Fig. 2.

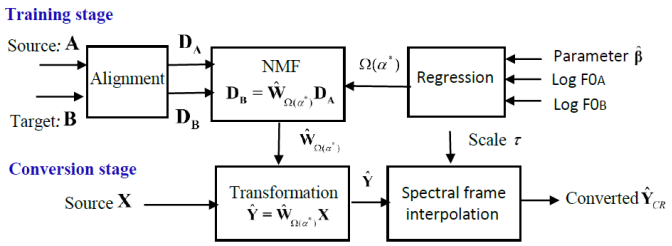

Fig. 2 *Integration of MLR, speech rate modification, and SSST.*

In Fig. 2, the source-target pair A-B data have two parallel streams, MFCC and STRAIGHT spectral envelopes, with the short-time analysis frame rate synchronized between the two streams. The MFCC stream is used for temporal alignment of the source and target frames by dynamic time warping (DTW) [24], which is computationally efficient due to the low dimensionality of MFCC. The alignment time indices are transferred to the spectral envelope stream to derive the aligned exemplar spectral matrices $\mathbf{D}_A$ and $\mathbf{D}_B$, where the aligned frames with mismatched voicing features are removed.

IV. Experimental Evaluations

Our dataset consisted of four target speakers with ALS (one female: tf1, and three males: tm1, tm2, and tm3), and four source speakers from the CMU-ARCTIC database [14] (two females: clb and slt, and two males: bdl and rms), all being American English speakers. The target speakers covered the full range of dysarthria, including mild, moderate, moderate-to-severe, and severe, for tm1, tf1, tm2 and tm3, respectively. The VC task covered the 16 speaker pairs. In our previous work, ARCTIC-to-ARCTIC conversions were performed on the same four ARCTIC speakers with 12 speaker pairs (refer to [12] for details). To record ALS speech data, each speaker was asked to read ARCTIC sentences a0001 through a0040, with the recording sampling rate of 16 kHz. Sentences a0001~a0020 were used for training, and a0021~a0040 for conversion testing.

The toolbox Tandem-Straight [13] was used to generate envelope spectra and pitch parameters and to synthesize converted speech. In speech analysis, the DFT size was 1024, and the frame shift was 4 ms. The HTK toolkit [25] was used to generate MFCC features (13 MFCCs, 13 $\Delta's$, 13 $\Delta\Delta's$) at the same frame rate as the STRAIGHT spectra. The GMM size that specifies the number of Gaussian densities in a mixture [12] was set to 5, considering the notion of 5 broad phonetic classes and the reduced phonetic distinctions in dysarthric speech. Spectral conversion was performed by WNMF integrated with multiple linear regression as discussed in Section III.A, as well as by HNMF. For each method, speech rate was modified as described in Section III.B. To simplify notation, we continue to use WNMF and HNMF to represent the two approaches after MLR or speech rate modification.

A. Preliminary evaluations

Frequency warping prediction

Table I *Frequency warping parameters by MLR (α^*_{MLR}) and by Factorization Error Minimization (α^*_{FEM}) for 16 speaker pairs as α^*_{MLR} / α^*_{FEM} ; inside brackets are average* $\log(F_0)$s *of the individual speakers.*

Target / Source		F	M		
		tf1 (2.3396)	tm1 (2.0901)	tm2 (2.0669)	tm3 (2.0833)
F	clb (2.2627)	0.96/1.00	1.06/1.08	1.08/1.04	1.06/**0.72**
	slt (2.2742)	0.96/1.00	1.06/1.08	1.07/1.09	1.07/**0.72**
M	bdl (2.0762)	0.88/0.92	0.97/1.04	1.00/0.96	0.99/0.96
	rms (1.9667)	0.84/0.88	0.94/1.00	0.95/0.96	0.94/0.96

The $\log(F_0)$s and the frequency warping parameters determined for the 12 ARCTIC speaker pairs in [12] were first used to estimate the regression parameter vector $\hat{\boldsymbol{\beta}}$ (Eq.(6)). Then, for each source-target speaker pair in the current task, the warping parameter $\hat{\alpha}^*$ was computed by MLR (Eq.(7)). The 16 $\hat{\alpha}^*$ values predicted for the 16 speaker pairs are shown in Table I, where inside the brackets are the average $\log(F_0)$ values of the individual speakers. For comparison, the selected warping parameters by minimizing factorization error (Eq. (3)) are also included. It is observed that the MLR predicted $\hat{\alpha}^*$ values were consistent with their expected low-to-high and high-to-low roles in frequency warping. An informal listening also confirmed their feasibility in VC. For the target speakers tf1, tm1, and tm2 (mild to moderate-severe dysarthria), the MLR-predicted parameters were close to the respective values estimated by the selection method. For tm3 (severe dysarthria), however, while regression reasonably predicted the $\hat{\alpha}^*$ values, the selection method failed for the two female source speakers, resulting in highly distorted, child-like voices. It is also worth noting the computational advantage of MLR (Eq. (7)) over the factorization error minimization (Eq.(3)).

Speech rate modification

Denote the durations of the i-th target and source utterances by $T_{\text{target},i}$ and $T_{source,i}$, respectively, with the durations including between-word pauses. Considering the fact that the source speakers' speech rates were similar, for each utterance i, the

average duration, $\bar{T}_{\text{source},i}$, was computed over the four source speakers. Then, given a target speaker, the speech rate conversion factor τ was approximated by an averaged duration ratio $\tau = \frac{1}{N}\sum_{i=1}^{N}\frac{T_{\text{target},i}}{\bar{T}_{\text{source},t}}$, with N the total pairs of utterances. For the target speakers tf1, tm1, tm2, and tm3, the ratios were approximately 1.8, 1.2, 2.0, and 2.0, respectively, where for tm3, the ratio was capped at 2 so that the converted speech articulation rate was not overly reduced. In Table II, the speech durations totaled over the 40 ARCTIC sentences are provided for the 8 speakers. It is observed that the duration of tm3 was much longer than the other ALS speakers, as a result of many more prolonged pauses in tm3's speech.

Table II *Durations (sec.) of 40 sentences of the source and target speakers*

ARCTIC	bdl: 119.52	clb: 140.45	slt: 119.03	rms: 123.12
ALS	tm1: 154.67	fm1: 219.87	tm2: 244.67	tm3: 400.56

Prior to conducting formal listening tests, the authors completed an informal auditory-perceptual evaluation on the effect of speech rate modification on speech quality and voice similarity. The pilot results suggested that the proposed rate modification significantly enhanced voice similarity to the target speakers at a moderate cost of reduced quality relative to the unmodified rate samples. This tradeoff was considered worthwhile under our overall objective of voice personalization, and thus the rate-modified speech was used in the formal evaluations.

It is worth noting that our previous evaluations of SSST [12] on healthy-to-healthy VC produced good scores for similarity to target even without speech rate conversion. In the current healthy-to-ALS VC task, it appears that in addition to spectra and pitch, listeners tend to associate the slower-than-normal speech rate of the target speaker with voice similarity. From a perceptual standpoint this is a reasonable strategy for similarity judgments and suggests that similarity scores may worsen if the source speech rate is applied to VC for ALS speakers. In general, for healthy-to-ALS VC, conversions on speech spectra, pitch, and rate are all needed, but the interaction between speech rate conversion and the other conversions needs to be examined systematically. However, it is beyond the scope of the current study to subjectively evaluate the interactions among the conversion factors.

B. Formal evaluations

Subjective listening tests on speech quality, intelligibility, and voice similarity were conducted to assess the effectiveness of the proposed VC approach. Speech quality was measured by mean opinion score (MOS) on a scale of 1 to 5, with 1=bad, 2=fair, 3=good, 4=very good, 5=excellent. Speech intelligibility was measured by word error rate, with errors including insertion, deletion, and substitution. Voice similarity was measured by an ABX test, where each converted speech sample was compared against a source and a target sample, and the listeners were asked to judge whether the converted sample was more similar to the source speaker or to the target speaker, with "Equal" for indistinguishability.

Speech quality and voice similarity were evaluated by 40 listeners whereas speech intelligibility was evaluated only by 10 experienced listeners, because for the latter test listeners were required to orthographically transcribe what they heard, which was complex and time consuming. The generic exemplar method HNMF (Section II.C) was included for comparison. The procedures and outcomes are detailed below.

Test 1. Speech quality

Each listener listened to a total of 64 audio samples: 16 (speaker pairs) × 4 (source, target, WNMF, HNMF). Given a speaker pair, 4 audio samples were played in a random order (in different sentence texts): source, target, WNMF, and HNMF, and the listener gave MOS scores to each sample. The individual and average scores for the target speakers with ALS are summarized in Table III. The scores were normalized relative to fixing the source score to 5.

The MOS scores varied largely among the four target speakers, in agreement with their dysarthria severity. The VC speech MOS showed positive correlations with the target speech MOS. The average MOS of WNMF approached the level "good," while the MOS of HNMF approached the level "fair" and the HNMF score was below WNMF by more than 1 point. Based on the paired Student t-test [26], the difference in the average MOS scores between Source vs. Target, Target vs. WNMF, Target vs. HNMF, and WNMF vs. HNMF were all statistically significant at the level of $p \le 0.005$.

Table III *MOS evaluation (40 listeners)*

Target Speakers	Dysarthria Severity	Target	WNMF	HNMF	Source
tm1	mild	5.0000	3.3273	2.2974	-
tf1	moderate	4.3207	2.8589	1.7967	-
tm2	mod.-sev.	3.7147	2.8591	1.5828	-
tm3	severe	3.4010	2.7023	1.4260	-
Average		4.1152	2.9269	1.7624	5.0000

Note. mod.-sev. = moderate-to-severe dysarthria

A separate preference test was also conducted on speech quality to compare the two VC methods. Each listener listened to a total of 32 audio samples: 16 (speaker pairs) × 2 (WNMF, HNMF). Given a speaker pair, 2 audio samples were played in a random order (in different sentence texts): WNMF and HNMF, and the listener was asked to judge which sample had a better quality, with "Equal" for indistinguishability. The listeners' ratings are summarized as percentages in Table IV, where the "Equal" case is omitted because it can be inferred from the fact that "preference to WNMF" plus "preference to HNMF" plus "Equal" totals 100%. There was an increased preference to WNMF up until the level of moderate-severe dysarthria. On average, WNMF led HNMF by a large absolute preference margin of over 80%. Both the MOS and preference tests indicated that the speech quality of the WNMF method was significantly better than the HNMF method.

Table IV *Speech quality preference evaluation (40 listeners)*

Target Speakers	Dysarthria Severity	WNMF	HNMF
tm1	mild	84.38%	8.12%
tf1	moderate	89.38%	6.25%
tm2	moderate-severe	90.62%	5.62%
tm3	severe	90.00%	6.88%
Average		88.59%	6.72%

Test 2. Voice similarity

Voice similarity to the target was evaluated on WNMF and HNMF. For each conversion method, each listener listened to a total of 96 audio samples: 16 (speaker pairs) × 3 (source, target, WNMF) and 16 (speaker pairs) × 3 (source, target, HNMF). Given a speaker pair and a VC method, three audio samples consisting of source, target, and voice-converted were played (in different sentence texts): the source and target playback order was randomized, and the voice-converted sample was always played the last. The listener chose whether the converted sample was more similar to the source or to the target, with "Equal" for indistinguishability. The test results are summarized as percentages below in Table V, where the "Equal" case is omitted for a similar reason as in Table IV.

Table V *Voice similarity preference evaluation (40 listeners)*

Target Speakers	Dysarthria Severity	Similar to Source		Similar to Target	
		WNMF	HNMF	WNMF	HNMF
tm1	mild	36.55%	18.06%	51.03%	70.14%
tf1	moderate	8.12%	5.00%	90.62%	95.00%
tm2	mod.-sev.	18.24%	3.12%	78.62%	95.00%
tm3	severe	25.00%	2.50%	68.12%	95.62%
Average		21.63%	6.89%	72.60%	89.42%

Note. mod.-sev. = moderate-to-severe dysarthria

The results in Table V suggest that for both VC methods the converted speech sounded significantly more similar to the target than to the source. In WNMF, the mild ALS speaker (tm1) had a smaller preference percentage for the target than the other ALS speakers. One reason could be that the speech rate of tm1 was not very different from the healthy source speakers ($\tau = 1.2$), and thus speech rate conversion did not impact voice similarity for this speaker as much as for the other three speakers ($\tau = 1.8, 2.0, 2.0$). Another observation that may partially account for these results is that the male ALS speakers all had noticeable pathological vocal source characteristics like tremor, jitter, shimmer, and noise. As WNMF performed spectral conversion, it is likely that these vocal source characteristics were not captured, but because tm2 and tm3 benefited more from the speech rate conversion than tm1, their target preference ratings were higher than tm1. The female target speaker (tf1) received the highest target similarity score for WNMF, likely because her nasal speech characteristics were captured well by the spectral transform.

HNMF yielded higher voice similarity to target speakers than WNMF did. A paired Student t-test indicated that the difference in average preference (%) of HNMF vs. WNMF was statistically significant at the level of $p \le 0.005$. This outcome reflected the fact that HNMF used speech spectra of target speakers directly as exemplars, and so it retained the characteristics of the target voice better. It is worth noting that our previous work on VC between healthy speakers, i.e., ARCTIC-to-ARCTIC, showed that listeners judged WNMF to be superior to HNMF in speech quality as well as in voice similarity [12]. A possible explanation for this difference is that in the current study listeners judged voice similarity with reference to dysarthria severity in the target speech, since HNMF retained certain dysarthria characteristics of target speakers while WNMF retained fewer dysarthria characteristics, as evidenced by the MOS results in Tables III and IV above and the intelligibility results in Table VI below.

Test 3. Intelligibility

In this test, each listener listened to a total of 48 audio samples: 16 (speaker pairs) × 3 (target, WNMF, HNMF). Given a speaker pair, 3 audio samples were played in a random order (in different sentence texts): target, WNMF, and HNMF. For each audio sample, the listener typed the words that he or she heard into a computer. To score the listeners' responses, word spelling errors were first corrected, then Edit distance [27] was used to align the words of each typed sentence with the reference sentence, and finally, substitution, insertion, and deletion errors were counted, giving the word error rate (WER):

$$WER = \frac{\#substitutions + \#insertions + \#deletions}{\#reference\ words} \times 100\%$$

The evaluation outcome is summarized below in TABLE VI.

Table VI *Intelligibility evaluation in WER (%) (10 listeners)*

Target Speakers	Dysarthria Severity	Target	WNMF	HNMF
tm1	mild	4.85	3.49	2.91
tf1	moderate	10.21	5.03	15.31
tm2	moderate-severe	40.91	6.23	14.21
tm3	severe	48.55	5.96	36.76
Average		26.13	5.18	17.30

It is clear that moderate to severe dysarthria posed a real challenge to listeners' comprehension. The WER's ranged from 4.85% to 48.55% for mild to severe dysarthria, which correspond to the range of MOS scores in Table III, i.e., lower WERs correspond to higher MOS scores. The method WNMF reduced the average WER to 5.18%, a relative improvement of 80.17% over the target. In comparison, HNMF reduced the average WER to 17.30%, a relative improvement of 33.79% over the target. For the mild dysarthric (tm1), the word error rate was reduced to a similar extent by both conversion methods. At the levels of moderate to severe dysarthria, however, HNMF was much less effective, as target spectra that retained certain dysarthria characteristics were used as exemplars. It is possible that the reductions in WER by HNMF over the target speech for tm2 and tm3 resulted from the slower articulation rate in the converted speech than in the target speech, The more frequent and longer pauses in the ALS speech relative to source speech led to a slower articulation rate in VC speech due to rate conversion, which may have given listeners more time to access lexical items and segment the speech signal at appropriate syntactic and prosodic boundaries, which aid with intelligibility judgments [28]. Additionally, the transformed source pitch profile and the STRAIGHT vocoder may have benefited speech intelligibility by reducing the pathological ALS vocal source characteristics in the converted speech. However, the reason for the increased WER by HNMF in the case of tf1 remains unclear. Based on the paired Student t-test, the difference in the average WERs of WNMF vs. Target, and WNMF vs. HNMF were statistically significant at the level of $p \le 0.005$ and $p \le 0.01$, respectively, and the difference in the average WERs of HNMF vs. Target was statistically significant at the level of $p \le 0.05$.

V. Conclusion

In this study, we investigated using VC to generate personalized speech for target speakers with ALS. To handle the limited quantity and dysarthric quality of target speech data, we integrated MLR-based frequency warping prediction with sparse structured transform for spectral conversion, and utilized spectral frame interpolation for speech rate modification. Subjective listening evaluations were conducted on the ARCTIC-to-ALS VC speech consisting of 16 speaker pairs, with positive results obtained for speech intelligibility, quality, and voice similarity. On average, the proposed voice conversion approach improved speech intelligibility by about 80% over the target ALS speech; the converted voice was over 3 times more similar to the target than to the source; and the converted speech quality approached the MOS scale "good" relative to the source speech being "excellent."

Our current work demonstrates strong support for using VC to assist speech communication for patients with ALS. The key advantage of our VC method is its robustness to the full range of dysarthria severity, yielding a significant gain in intelligibility and a good balance between quality and voice similarity. Our work can be readily applied to our next stage of study, which is using high-quality TTS as the voice source. However, training a TTS voice directly by using the speech data of a person with ALS might be feasible only for mild dysarthria. For more severe dysarthria, the TTS voice personalized to the individual will have inferior intelligibility similar to the ALS speech used for training. Although newer data-driven vocoders, such as WaveNet [29], surpass the source-channel vocoders, such as Tandem-STRAIGHT [13], in speech quality and voice similarity, for target speakers with ALS, training the data-driven vocoder would be just as problematic as training a TTS system with ALS speech. In this regard, robust learning methods are yet to be developed and we leave this issue to a future study.

Lastly, subjective listening tests based on auditory-perceptual judgments are prone to high inter-rater variability [30]. Employing a large number of listeners for these tests is costly and infeasible due to listener burden; therefore, objective measures are more desirable for VC, especially when the target speakers have speech impairments. Furthermore, including patients and their families in the assessment of voice similarity and speech quality will likely help improve design decisions and may encourage patients to use their speech generating devices to communicate with their family, caregivers, and health professionals.

References

[1] W. O. McKinley, R. T. Seel, and J. T. Hardman, "Nontraumatic spinal cord injury: incidence, epidemiology, and functional outcome," *Arch. of Phy. Med. and Rehabil.*, vol. 80, no. 6, pp. 619-623, 1999.

[2] H. T. Bunnell, J. Lilley, C. Pennington, B. Moyers, and J. Polikoff, "The ModelTalker system," in *Proc. the Blizzard Challenge Worksho*p, Kyoto, Japan, 2010.

[3] L. J. Ball *et al.*, "Eye gaze access of AAC technology for people with amyotrophic lateral sclerosis," *J. of Med. Speech-Lang. Pathol.*, vol. 18, no. 3, pp. 11-23, 2010.

[4] J. Tamagishi, C. Veaux, S. King, and S. Renals, "Speech synthesis technologies for individuals with vocal disabilities: voice banking and reconstruction," *Acoust. Sci. & Tech*. vol. 33, no. 1, pp. 1-5, 2012.

[5]. J. Shen et al., "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, pp. 4779-4783, Calgary, Canada, 2018.

[6] T. Toda et al., "The voice conversion challenge 2016," in *Proc. Interspeech*, pp. 1632-1636, San Francisco, USA, 2016.

[7] J. Lorenzo-Trueba et al., "The voice conversion challenge 2018: promoting development of parallel and nonparallel methods," in *Proc. Speaker Odyssey 2018 The Speaker and Lang. Recognit. Workshop*, pp. 195-202, Les Sables d'Olonne, France, 2018.

[8] H. Doi, T. Toda, K. Nakamura, H. Saruwatari, and K. Shikano, "Alaryngeal speech enhancement based on one-to-many eigenvoice conversion," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 1, pp. 172-183, 2014

[9] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki, "Individuality-preserving voice conversion for articulation disorders based on non-negative matrix factorization," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, pp. 8037-8040, Vancouver, Canada, 2013.

[10] A. Kain, J.-P. Hosom, X. Niu, J. P. H. van Santen, M. Fried-Oken, and J. Staehely, "Improving the intelligibility of dysarthric speech," *Speech Commun.*, vol. 49, no. 9, pp. 743-759, 2007.

[11] L.-W. Chen, H.-Y. Lee, and Y. Tsao, "Generative adversarial networks for unpaired voice transformation on impaired speech," in *Proc. Interspeech*, pp. 719-723, Graz, Austria, 2019.

[12] Y. Zhao, M. Kuruvilla-Dugdale, and M. Song, "Structured Sparse Spectral Transforms and Structural Measures for Voice Conversion," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 12, pp. 2267-2276, 2018.

[13] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, "TANDEM STRAIGHT: a temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, pp. 3933-3935, Las Vegas, USA, 2008.

[14] J. Kominek and A. W. Black, "CMU ARCTIC database for speech synthesis," *CMU-LTI-03-177*, 2003.

[15] H. Benisty and D. Malah, "Voice conversion using GMM with enhanced global variance," in *Proc. Interspeech*, pp. 669-672, Florence, Italy, 2011.

[16] Y. Yunusova *et al.*, "Profiling speech and pausing in amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD)," *PLoS One*, vol. 11, no.1, pp. 1-18, 2016.

[17] D. D. Lee and H. S. Seung, "Algorithm for nonnegative matrix factorization," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, pp. 556-562, Vancouver, Canada, 2001.

[18] A. V. Oppenheim and D. H. Johnson, "Discrete representation of signals," *Proc. IEEE*, vol. 60, no. 6, pp. 681-691, 1972.

[19] Z. Wu, T. Virtanen, E. S. Chng, and H. Li, "Exemplar based sparse representation with residue compensation for voice conversion," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 10, pp. 1506-1521, 2014.

[20] H. Ming, D. Huang, L. Xie, S. Zhang, M. Dong, and H. Li, "Exemplar-based sparse representation on timbre and prosody for voice conversion," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, pp. 5175-5179, Shanghai, China, 2016.

[21] E. Alpaydin, **Introduction to Machine Learning**, MIT Press, 2004.

[22] H. Valbret, E. Moulines, J. P. Tubach, "Voice transformation using PSOLA techniques," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, pp. I-145-149, San Francisco, USA, 1992.

[23] Y. Stylianous, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech Audio Process.*, vol. 6., no. 2, pp. 131-142, 1998.

[24] L. Rabiner and B.-H. Juang, Fundamental of Speech Recognition, Prentice Hall, 1993.

[25] http://htk.eng.cam.ac.uk/

[26] R. Richard, "Concepts and Applications of Inferential Statistics," http://vassarstats.net/textbook/, 2011.

[27] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals," *Sov. Phys., Doklady*, vol. 10, no. 8, pp. 707-710, 1966.

[28] R. Smiljanić and A. R. Bradlow, "Speaking and hearing clearly: Talker and listener factors in speaking style changes," *Lang. Ling. Compass*, vol. 3, no. 1, pp. 236-264, 2009.

[29] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, T. Toda, "Speaker-dependent WaveNet vocoder," in *Proc. Interspeech*, pp. 1118-1122, Stockholm, Sweden, 2017.

[30] C. Sheard, R. D. Adams, and P. J. Davis, "Reliability and agreement of ratings of ataxic dysarthric speech samples with varying intelligibility," *J. Speech Hearing Res.*, vol. 34, no. 2, 285–293, 1991.