IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 19 February 2024, accepted 10 March 2024, date of publication 4 April 2024, date of current version 12 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3384973

RESEARCH ARTICLE

An Investigation of Fundamental Frequency Pattern Prediction for Japanese Electrolaryngeal Speech Enhancement Based on Frame-Wise Phoneme Representations

MOHAMMAD ESHGHI[1] **AND TOMOKI TODA**[2]**, (Senior Member, IEEE)**
[1]Graduate School of Information Science, Nagoya University, Nagoya 464-8601, Japan
[2]Information Technology Center, Nagoya University, Nagoya 464-8601, Japan

Corresponding author: Mohammad Eshghi (mohammad.eshghi@g.sp.m.is.nagoya-u.ac.jp)

This work was partially supported by JST, CREST, and JPMJCR19A3.

ABSTRACT Total laryngectomy (TL) is as a well-established treatment for advanced laryngeal malignancies, entailing the complete removal of the larynx. Speech rehabilitation following TL is crucial for improving the quality of life and facilitating social reintegration. Electrolaryngeal (EL) speech, a widely used voice restoration technique utilizing external excitation signals, often produces artificial and monotonous sound quality. Efforts to enhance EL speech include the application of statistical voice conversion and neural approaches to speech enhancement. These approaches typically aim to map spectral features into acoustic characteristics, including the fundamental frequency (F_0). However, challenges arise owing to substantial discrepancies and pattern differences between extracted features for EL and normal speech, compounded by limited clinical training data. To address this issue, we explored F_0 pattern prediction based on frame-wise phoneme information using bidirectional long short-term memory recurrent neural networks. Beyond direct predictions based on phoneme labels, we expanded our analysis to include real-valued phoneme embeddings and conducted predictions for clustered embeddings representing low-dimensional input representations. Our findings demonstrate that both regression and classification predictive modeling can map frame-wise phoneme information into natural F_0 patterns. Additionally, phoneme labels can be considered as shared features between EL and normal speech, allowing for improved prediction accuracies by incorporating phoneme information from normal speech into the training sets for EL speech. Furthermore, by learning phoneme embeddings and creating input features for F_0 prediction based on the clustering of these embeddings, accurate F_0 patterns can be predicted, and the challenge of finding a strategy to reduce the dimensionality of the input features can be effectively alleviated.

INDEX TERMS Electrolaryngeal speech, fundamental frequency prediction, phoneme labels, phoneme embeddings, speech enhancement.

I. INTRODUCTION

The larynx, or the voice box, is the region that connects the throat to the trachea, located in the front of the neck. It contains the vocal folds, which open for breathing, close

The associate editor coordinating the review of this manuscript and approving it for publication was Manuel Rosa-Zurera.

to protect the trachea during swallowing, and vibrate for voice production. Laryngeal cancer arises when cancerous cells develop in the laryngeal tissues. Depending on the type, extent, and location of the disease, respiration and speech production mechanisms may be partially or totally disrupted. The primary medical treatment for glottic cancer, involving uncontrolled cell growth on the vocal folds, is the excision of the affected vocal folds and surrounding tissues, a procedure known as total laryngectomy (TL) [1], [2], [3], [4], [5], and patients who have undergone this operation are called laryngectomees.

Following TL, the pharynx is decoupled from the trachea, and inhalation and exhalation occur through an opening in the trachea called the *tracheostoma* or *stoma* [1]. Owing to these anatomical changes, individuals lose the ability to speak normally, and the fundamental frequency (F_0) generation mechanism is entirely compromised.

As a direct consequence of TL, individuals often experience a decrease in social interaction, diminished interest in in-person engagements, and a gradual withdrawal from society [6]. Therefore, effective voice restoration becomes paramount for enhancing the quality of life (QoL) of laryngectomees and reintegrating them into their pretreatment world. Over the past 150 years [3], three voice restoration techniques have evolved: esophageal speech (ES), electrolaryngeal speech (EL), and tracheoesophageal speech through a tracheoesophageal puncture (TEP).

Although TEP speech is considered the gold standard, often resulting in clear voices [5], ES speech and EL speech are particularly popular in Japan [7]. ES speech involves ingesting air into the esophagus and deliberately belching it up to make the tissue of the pharyngo-esophageal segment in the pharynx vibrate [1], [8], [9]. ES speech has a belching-like sound quality and is challenging to learn, although it does not require any devices or procedures. EL speech, on the other hand, involves using a handheld battery-operated device, an *electrolarynx* or an *artificial larynx*, held against the skin of the neck or cheek. This device transmits single-tone excitation signals through the skin into either the pharynx or the oral cavity [2]. Patients alter these signals with their lips, tongue, and mouth to generate speech.

Although EL speech has the advantages of being non-invasive and immediately usable following TL by patients with moderate electrolarynx proficiency, it is characterized by a mechanical and buzzy speech quality. On the one hand, generating sufficiently loud sound source signals to produce audible EL speech causes degradation in intelligibility owing to a buzzing noise from the electrolarynx, known as directly radiated EL noise [10]. On the other hand, the electrolarynx lacks the ability to generate natural F_0 patterns corresponding to linguistic content, resulting in a deficiency in both suprasegmental and voice-related segmental characteristics [9], [11]. Therefore, producing natural-sounding and expressive EL speech remains a significant challenge, necessitating further research.

Various strategies have been employed to enhance EL speech quality. In previous studies, F_0 and voicing state control using signal processing approaches have been explored [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22]. Detecting specific biosignals from sensors with different modalities and translating them into F_0 control signals have always been the main objectives of such studies. However, these efforts have faced limitations in providing fine control over pitch regulations to produce a wide range of intonations, resulting in inadequate speech quality. A different approach involves EL speech enhancement based on the statistical voice conversion (VC) technique [23], [24]. This approach utilizes a conversion function learned from parallel training data to modify the acoustic features of EL speech without manipulating linguistic information. Although the data-driven VC-based approach offers more sophisticated acoustic modifications, concerns arise regarding prediction accuracies and potential oversmoothing of parameters [24]. Despite these challenges, ongoing research in this area aims to bridge the gap in achieving enhanced EL speech that closely resembles naturally produced speech by vocal folds.

Recently, deep learning (DL) has surpassed traditional machine learning (ML) approaches in various applications, leveraging its capacity to learn representations with different abstraction levels from unengineered input features [25]. This feature-learning capability enables DL models to capture information on complex macro- and microstructures and learn the underlying mappings between these structures and target variables, resulting in significant speedups compared with conventional methods. DL-based techniques have demonstrated success in alaryngeal speech enhancement [8], [26], [27], [28], [29], notably in mapping EL speech spectral features into natural F_0 patterns using recurrent neural networks (RNNs) [8], [28]. Furthermore, sequence-to-sequence (seq2seq) modeling for EL-to-normal speech (EL2SP) conversion has been proposed [30], [31] using text-to-speech (TTS) pretraining [32] and the attention-based encoder–decoder framework [33].

Despite the universal approximation theorem asserting that neural networks can approximate any function with arbitrary accuracy [34], the complexity of learning the mapping function in DL models should not be underestimated. DL models typically involve thousands to millions of parameters learned during the training phase, making large datasets necessary for successful training. However, gathering extensive data, especially in clinical settings with limited groups such as laryngectomees, can be challenging and time-consuming. Additionally, DL models excel in recognizing patterns when training data is devoid of irregularities, raising concerns when dealing with alaryngeal voice spectra where structures for certain phonemes often collapse owing to production difficulties. Consequently, models must accommodate significant pattern alterations, and training with inadequate and unstructured data may lead to inaccurate mapping functions.

In contrast to conventional TTS systems reliant on statistical parametric speech synthesis (SPSS) [35], which require substantial domain expertise and extensive text analysis, neural end-to-end TTS systems such as Deep Voice 3 [36], Char2Wav [37], Tacotron [38], and Tacotron2 [39] accurately estimate acoustic features on the basis of normalized character sequences. Motivated by recent advancements in neural speech synthesis, we investigated natural F_0 pattern prediction for EL speech based on frame-wise phoneme information in our previous works [40], [41]. We demonstrated that F_0 prediction benefits from existing dependencies between neighboring phonemes in phoneme concatenations. However, we did not specifically explore how the type of target F_0 pattern can impact prediction accuracies, a critical consideration as it specifies the type of mapping function. Although a regression problem is typically solved to learn a continuous mapping function for F_0 prediction, it is also possible to quantize target F_0 patterns into discrete intervals and solve a classification problem to obtain a discrete mapping function. In this study, we examined the significance of the mapping function type for predicting natural F_0 patterns from frame-wise phoneme information and present a comparative analysis of predictive modeling methods' performances. We also provided a thorough explanation of the process of turning phoneme labels into phoneme embeddings and clustering them to create low-dimensional input features for real-world F_0 prediction. Specifically, we made the following contributions:

1) We demonstrated that, in addition to being able to predict F_0 patterns using regression predictive modeling, classification predictive modeling can also achieve this goal.
2) We developed a prediction system based on the bidirectional long short-term memory (BiLSTM) architecture that leverages both regression and classification modeling techniques to convert forced-aligned phoneme labels into natural F_0 patterns.
3) We showed that low-dimensional input features derived from a subset of phoneme labels can also result in accurate predictions.
4) We provided a phoneme embedding-based system that allows shared features of any dimension to be produced for F_0 prediction.
5) We demonstrated that the prediction accuracies for EL speech can be improved by sharing phoneme information between EL and normal speech.

The remainder of this paper is organized as follows. In Section II, we provide an overview of existing works on EL speech naturalness enhancement and summarize their drawbacks and limitations. In Section III, we introduce the idea of predicting natural F_0 patterns from frame-wise phoneme information and describe how F_0 patterns can be predicted using regression and classification predictive modeling. The experimental conditions and results are described in section IV. Finally, the discussion of findings and conclusion are provided in section V and VI, respectively.

II. RELATED WORKS

A. F_0 AND VOICING STATE CONTROL BASED ON SIGNAL PROCESSING TECHNIQUES

To generate nonmonotonic excitation signals, variations in certain physical quantities or spatial displacements must be correlated with an F_0 control signal. The F_0 patterns can then be modified by the established link either manually, as they are in direct response to user interaction, or fully automatically. Manual methods for controlling F_0 patterns involve tracking hand gestures, measuring pressure, and discerning the direction of thumb movements. Pressure-sensitive artificial excitation signal production, as presented in [15] and [42], utilizes a pressure-sensitive push button to add intonation to utterances. However, this approach has limitations when pressures exceed the push button's feasible reaction range. An F_0 control mechanism with two degrees of freedom (DoFs) has been introduced in [16], where thumb movements in the left and right directions are converted into a control signal. Here, displacement-related variations in the amount of light recorded by a light sensor are transformed into electrical signals (voltages at different levels) that regulate F_0 patterns. Additionally, by using a microswitch, up/down movements are utilized as voice commands to control the electrolarynx on/off times. It has been reported that this system is effective for generating initial-accented utterances and question phrases with rising F_0 patterns at the end. The extension of this control mechanism to an intra-oral pressure sensor can be found in [19].

Fully automated scenarios involve sensing certain physiological parameters or biosignals to generate a multilevel voltage signal that actuates the electrolarynx. Consequently, the control signal relies on the space–time records of biological events, such as muscle contractions. Obtaining control information from air pressure at the tracheostoma output [12], [13] and neck surface electromyography over laryngeal muscles [17], [20], [22] are popular techniques for recording biological events. In the former technique, an air pressure sensor is placed on the tracheostoma, either by holding it with one hand or by mounting it on a neck brace, to measure the expelled air stream from the lungs. The sensor output is fed into an electrical circuit, which uses it to produce an F_0 value for pitch regulation. However, gaining conscious control over pitch using this technique is challenging, and requires skill and dexterity [43]. In the latter technique, electromyography (EMG) detects the electrical potential or activity generated by laryngeal muscle contractions during vocalization, serving for F_0 control and regulation. This involves either direct detection through intramuscular (invasive) needle electrodes inserted into the cricothyroid (CT) muscle or indirect detection with surface electrodes on the neck skin above the infrahyoid (IH) strap muscles [20], [43]. According to [14], F_0 shows a higher correlation with CT muscle activity levels than with IH strap muscle activity levels. However, for noninvasive recordings of voice-related activation patterns, IH strap muscle activations are commonly used to extract the predictive relationships between F_0 and laryngeal muscle

activations. EMG-based F_0 regulation faces two primary challenges [43] as follows: (1) precise pitch control for creating a broad range of intonations is challenging and (2) recorded EMG signals are individual-specific and session-dependent owing to variations in electrode placement, as well as differences in electrode, skin, and muscle conditions across sessions [26]. Consequently, building a universal transform function between EMG activities and F_0 is challenging.

B. VC-BASED STATISTICAL F_0 PREDICTION AND VOICING STATE CONTROL

While keeping linguistic information intact, VC-based systems aim to modify speech parameters by using a statistical features mapping process. The modifications of acoustic features are based on nonlinear mapping functions learned in a data-driven manner using both signal processing and ML techniques. Gaussian mixture model (GMM)-based F_0 modeling [24], [44] is a well-developed predictive modeling technique wherein, on the basis of a parallel dataset and a supervised training procedure, natural F_0 patterns are predicted.

Before training, the corresponding utterances in the dataset are time-aligned using a dynamic time-warping (DTW) algorithm to obtain joint feature vectors. During training, a GMM models the joint probability density of input and output parameter vectors, and the expectation-maximization (EM) algorithm estimates the model parameters. In the prediction stage, the input features extracted from EL speech are mapped to the most likely target F_0 pattern using the maximum likelihood parameter generation (MLPG) technique.

Common input features in VC systems include mel-cepstral (MCEP) coefficients (in segmental or nonsegmental form), mel-frequency cepstral coefficients (MFCC), and EMG activations. For instance, F_0 prediction based on segmental MCEP coefficients has been discussed in [44], whereas F_0 prediction based on MFCCs is available in [9], [45], and [46]. Additionally, F_0 and voicing state estimation from EMG data have been studied by Nakamura et al. [47] and Janke and Diener [27].

Despite significantly improving the quality of EL speech, GMM-based F_0 prediction has two major shortcomings. Modeling and prediction errors are concerns, and parameters estimated by the MLPG technique often exhibit an oversmoothing tendency. Oversmoothing reduces the naturalness of the enhanced voices, whereas prediction errors immediately affect speech intelligibility. Considering that the global variance estimation proposed by Toda et al. [24] can significantly reduce the oversmoothing issue, it is clear that the primary goal should be to minimize modeling and prediction errors. Therefore, it is crucial to consider more accurate mapping functions with greater capabilities to extract predictive relationships between source and target features.

Shallow and deep neural networks have attracted particular attention in statistical VC and speech enhancement in recent years. What distinguishes neural networks from other ML techniques is their hierarchical structure, allowing them to acquire increasingly complex feature representations from the provided input features. Virtually all current F_0 pattern prediction systems have replaced their mapping functions with neural networks owing to this task-specific feature-learning characteristic. For example, RNNs have been used to predict natural F_0 patterns for alaryngeal voices on the basis of conventional spectral features in [8] and [28]. Similarly, Diener et al. [29] developed a prediction system based on electromyographic signals.

To address the challenges of EL speech enhancement, both frame-to-frame and seq-to-seq [48] mapping paradigms can be applied. Seq-to-seq VC models, utilizing an attention-based encoder–decoder architecture [33], can perform representation learning and alignment simultaneously [30], capturing long-term dependencies such as prosody and speaker identity [32]. Some research studies have demonstrated the potential of using TTS pretraining in conjunction with seq-to-seq modeling for EL speech enhancement [30], [31]. However, most seq-to-seq models require a substantial amount of high-quality parallel training data. As it is difficult to create large-scale, high-quality parallel datasets for EL speech, using seq-to-seq models trained on insufficient data can lead to skipped phonemes and mispronunciations [49]. Consequently, the frame-to-frame mapping paradigm, such as long short-term memory (LSTM) based feature mapping [50], remains highly favored for EL speech enhancement. Furthermore, the computational complexity of frame-wise models is lower than that of the seq-to-seq models.

Regardless of the chosen mapping paradigm, insufficient training data with irregular patterns and dissimilar structures can compromise the task-oriented feature-learning characteristic and mapping power of neural networks, leading to limited EL speech enhancement results. In the next section, we introduce the concept of using frame-wise phoneme information to generate features with fewer disparities between EL and normal speech. We further explain how this approach enables the sharing of phoneme information between EL and normal speech, enhancing F_0 prediction accuracies for EL speech.

III. F_0 PREDICTION BASED ON FRAME-WISE PHONEME INFORMATION

A. OVERVIEW

In contrast to traditional TTS systems, recent neural end-to-end TTS systems such as Deep Voice 3 [36], Char2Wav [37], Tacotron [38], and Tacotron2 [39] accurately estimate acoustic features on the basis of normalized character sequences. In this section, we introduce a fully neural method to predict F_0 patterns for EL speech based on phoneme labels. By replacing acoustic features with linguistic features, specifically phoneme labels, we can establish a common representation for any statements or sentences laryngectomees struggle to utter and normal speakers effortlessly produce. Essentially, similar character sequences

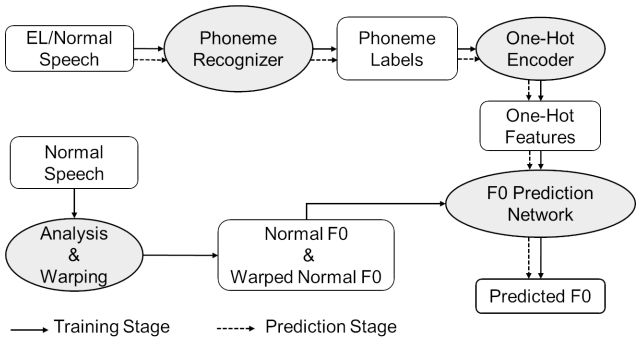


FIGURE 1. Mapping of the phoneme labels into target F_0 patterns using regression predictive modeling.

representing phoneme labels are acquired when both EL and normal speakers utter a shared sentence using the exact word sequence (or transcript). It is important to note that these acquired character sequences may have different temporal patterns attributable to variations in phoneme durations and pause positions. Leveraging these normalized character sequences and the strong dependencies between adjacent phonemes in phoneme concatenations, we aim to predict natural F_0 patterns.

Fig. 1 illustrates the system for predicting F_0 patterns based on phoneme labels using conventional regression predictive modeling. For this system, utterance pairs of EL and normal speech with their corresponding orthographic transcriptions are prepared in advance. A phoneme recognizer then extracts forced-aligned phoneme labels frame-by-frame, and these labels undergo one-hot encoding to transform them into numerical representations suitable for machine and deep learning algorithms. The one-hot encoded features are fed into the prediction network, constructed with BiLSTM layers [51], enabling the network to learn the underlying mapping between the target F_0 patterns and the provided categorical input features. After network training, any arbitrary input sequence of phoneme labels can be mapped into a natural F_0 pattern.

Moreover, it is important to note that coarticulation typically occurs in the production of adjacent sounds, demonstrating a temporal overlap in the articulatory gesture and phonetic representation of the sounds. Incorporating BiLSTM layers into prediction network enhances the model's ability to capture and leverage these intricate temporal relationships and improves the prediction accuracies. Depending on the type of target F_0 pattern used as ground truth data during network training, two distinct predictive modeling methods for approximating the mapping functions can be considered, namely, (1) regression modeling and (2) classification modeling. In the following section, we provide a detailed explanation of these predictive modeling methods.

B. F_0 PREDICTION APPROACHES

1) REGRESSION PREDICTIVE MODELING

In supervised training of the prediction network, the network must be provided with target F_0 patterns or ground truth data during the training phase. Predictive modeling is defined as the problem of developing a model for approximating the optimal mapping function from input features to target F_0 patterns. Knowing that the speech analysis for extracting excitation features yields one value per frame, which is either F_0 in Hz or 0 for segments of speech that are unvoiced, the goal of regression predictive modeling is to predict a real value at each prediction step. As a result, the skill of the model must be reported as an error in such predictions. The root mean squared error (RMSE), which minimizes the residuals between the predicted values and the ground truth data, is the most commonly used error measure in practice.

Most of the time, continuous response variables are employed in regression modeling to simplify and reduce the complexity of the data space. As F_0 is always equal to zero for unvoiced and silent frames, many discontinuities are observed in the extracted F0 patterns by vocoders. Interpolating F_0 values in the unvoiced regions using the spline interpolation method is a standard method of converting F_0 patterns to a continuous form. Furthermore, because it is difficult to precisely model and reproduce rapid movements in the F_0 patterns, also referred to as microprosody [52], it is frequently recommended to filter them out. Thus, the two preprocessing procedures for creating continuous F_0 patterns are spline interpolation and low-pass filtering.

2) CLASSIFICATION PREDICTIVE MODELING

In contrast to regression modeling where target F_0 values form a continuous output variable, in classification predictive modeling, output variables are discrete classes or levels. To put it differently, classification modeling turns the task of predicting a real value for F_0 into predicting the most-probable quantization level where the F_0 falls onto [29]. Therefore, by solving a multilevel classification problem, one can approximate the mapping function from input features to discretized target F0 patterns.

In multilevel classification problems, the model predicts a continuous value as the probability of a given input being quantized to each output level. Considering these probabilities, the most probable level can be simply determined on the basis of the output response with the highest value. By comparing the predicted outputs in one-hot form with the one-hot encoded target F_0 patterns, one can measure the skill of the model in terms of classification accuracy, which is the percentage of the correctly quantized samples out of all predictions made.

Quantized target F_0 patterns for the classification modeling are created on the basis of the median-cut algorithm, which was first presented by Paul Heckbert [53] for color image quantization to display high-quality reproductions of color images with short frame buffers. The first step in quantizing F_0 patterns to N bits, or equivalently to 2^N levels, is to take into account all of the training utterances and concatenate their F_0 patterns to create a large "*interval*" or long sequence

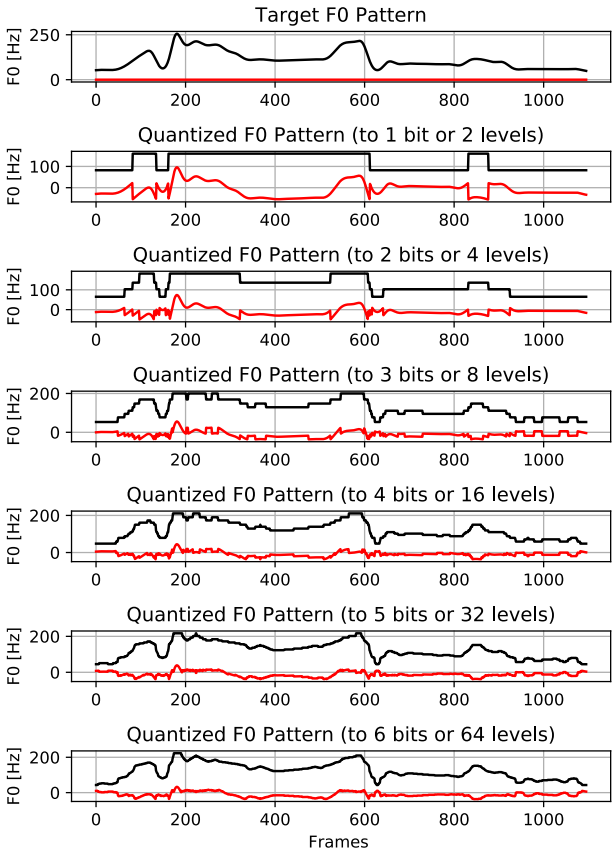


FIGURE 2. Quantization of the target F_0 pattern using median-cut algorithm for classification predictive modeling. The quantization error signal, also known as the unintended quantization noise, is shown by the solid red line. The quantization noise decreases as the number of quantization levels increases.

that contains all the existing target F_0 values. After sorting the values in this interval in ascending order, tuples of the form ($F0\_value$, $corresponding\_index$) are created. From this point on, as long as the total number of quantization levels is not reached, the largest remaining interval is split along its median into two subintervals, and the values in each subinterval are quantized to the mean value of that subinterval. Finally, the categorical output variables for the network training are obtained by the one-hot encoding of the quantized F_0 patterns. Note that both the quantization levels and the one-hot codes must be stored in a lookup table (or code book) to reverse this process and decode the encoded data. This will make it possible to implement an efficient lookup table decoding (LTD) routine.

Fig. 2 illustrates the quantization of a ground truth F_0 pattern for an EL utterance to 2, 4, 8, 16, 32, and 64 quantization levels. As we can see, the quantization process introduces a certain amount of error or distortion. This error is the quantization error or the unintended quantization noise. The quantization noise decreases as the number of quantization levels increases. However, a larger number of levels requires a larger number of bits for one-hot encoding of the F_0 values, which leads to high-dimensional output features. For practical applications, quantizing F_0 values to five bits is sufficient, and an average correlation of 0.99 is found between the quantized F_0 patterns and the reference ones [29].

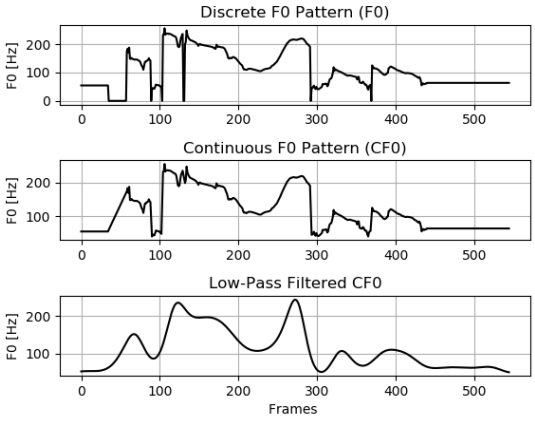
3) TIME-FRAME ALIGNMENT BETWEEN EL AND NORMAL SPEECH

Before we can train the prediction models, the created F_0 patterns for EL and normal speech must be time-aligned. In most VC systems, time alignment between source and target speakers is achieved by applying an automatic DTW algorithm to spectral features, i.e., to MCEP coefficients. This procedure produces accurate alignment results if comparable patterns with no abnormalities can be identified in the spectral features. The inability of laryngectomees to produce certain phonemes, which results in collapsed regions in the spectra of alaryngeal voices, makes it clear that noniterative time alignment cannot precisely match the acoustic features for EL speech with the extracted ones for normal speech. To address this problem, a multistep iterative method for approximating the time-warping functions has been proposed [54], by which the alignment is refined by minimizing the overall difference between the target and converted feature vectors.

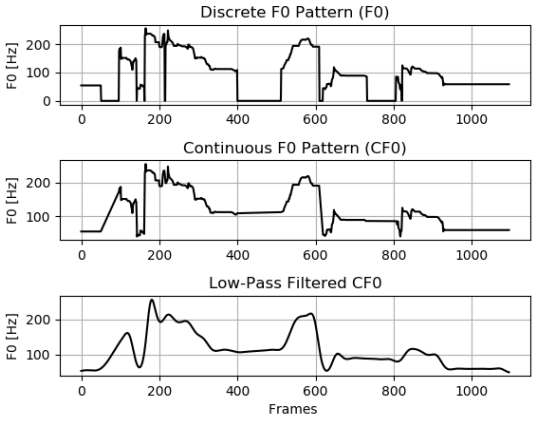
Aside from collapsed regions in the spectrum, the alignment accuracy is adversely affected by the presence of short pauses (SPs) in EL speech. Laryngectomees utter long sentences quite slowly and with many SPs in between to make them intelligible. Since normal speech does not contain these pauses, a significant mismatch between EL and normal speech emerges, which can considerably increase the complexity of the alignment process. To solve this mismatch problem, we present a novel warping method that utilizes both acoustic and phoneme-level linguistic features. Given an utterance pair of EL and normal speech, the alignment is carried out label by label, taking into account the sequences of forced-aligned phoneme labels. Every time, a tuple is created using a label from the EL speech sequence and the corresponding label from the normal speech sequence. If the labels in this tuple are the same, then the iterative alignment procedure described in [54] is applied to the matching spectral features to these labels to minimize the MCEP distortion. If these labels are not the same, a flag for an SP in the EL speech is raised. In such cases, target features at the position corresponding to the SP index are zero-padded. This process is repeated over all possible tuples, and the resulting warping paths are stacked together to form the ultimate warping path. After warping, the last steps in creating the final continuous F_0 patterns are low-pass filtering and spline interpolation. In Fig. 3, we summarize the stages involved in preparing these target F_0 patterns.

C. PHONEME REPRESENTATIONS ON PREDICTING F_0 PATTERNS

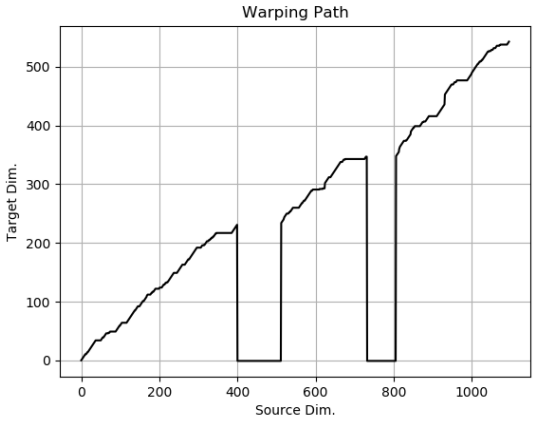
Phoneme representations are essential in predicting F_0 patterns, as phonemes serve as perceptually distinct units of sound in a language, distinguishing one word from another in terms of both pronunciation and meaning. Various languages



(a) **Low-pass filtered F_0 pattern for normal speech.**



(b) **Prepared target F_0 pattern for EL speech.**



(c) **Warping path obtained after time alignment. Zero-valued sections correspond to short pauses in EL speech.**

FIGURE 3. Preparation of ground truth F_0 patterns. Time alignment is achieved using both acoustic and phoneme-level linguistic features.

have different sets of phonemes, and legitimate words in a language are formed from a specific set of allowed phoneme labels. For instance, English has 44 contrastive sound units, composed of 20 vowels and 24 consonants [55]. Consonants are pronounced by impeding the airflow through the mouth [56], whereas vowels are produced by allowing the air to flow from the lungs through the mouth with minimal obstruction and without audible friction.

Japanese, being a syllabic language, comprises five vowel sounds represented by the Roman alphabet letters /a/, /e/, /i/, /o/, and /u/ [57]. Each vowel has short and long forms and can appear in the initial, medial, or final positions of words [58], [59]. The syllabic structure in Japanese typically follows the pattern of consonant plus vowel or vowel alone [57], resulting in open syllables that do not end with a consonant. Unlike English, Japanese does not have diphthongs (at phonological level); instead, it only includes monophthongs in its vowel inventory [60]. This lack of diphthongs allows vowels to occur sequentially, and two successive vowel sounds are considered two separate syllables. Understanding the phonemic structure of a language is crucial to predicting F_0 patterns, as it provides the basis for mapping linguistic features to acoustic features in the context of speech synthesis. The knowledge of phonemes and their characteristics aids in accurately modeling the relationships between phonetic elements and prosodic features, contributing to the naturalness of synthesized speech.

The utilization of frame-wise phoneme information for F_0 pattern prediction can be explored through the following three approaches:

1) One-hot encoding of all phoneme labels: This approach involves representing each phoneme label as a one-hot encoded vector. One-hot encoding is a binary representation where each phoneme corresponds to a unique binary pattern. This approach considers prediction based on the smallest sound units into which a word can be divided.
2) One-hot encoding of speech syllables: In this approach, syllables, which are larger sound units than phonemes, are considered. Each syllable is represented by a one-hot encoded vector.
3) One-hot encoding of all vowels: This approach focuses on encoding only vowel sounds, as syllables are typically made up of one vowel sound. Each vowel is represented by a one-hot encoded vector.

By training prediction models on these different sets of features and evaluating their accuracies, one can assess the importance of various linguistic information for F_0 pattern prediction. Additionally, we can investigate if accurate F_0 prediction is achievable using a subset of phoneme labels or low-dimensional input features. Successful F_0 prediction from these features suggests the potential for simplifying the structure of the required automatic speech recognition (ASR) system. Specifically, instead of aligning audio recordings with original transcriptions, the recordings could be aligned with modified or simplified transcriptions, reducing processing time and recognition errors. It is important to note that no ASR system is employed in this study, and the speech waveforms are forced-aligned to orthographic transcriptions to automatically generate phoneme-level segmentation.

1) PHONEME EMBEDDINGS

The following are three drawbacks to using one-hot codes to represent categorical values. (1) One-hot codes are

unlearned representations created by assigning a single 1 and a series of 0s to every observation. As a result, related categories are not positioned adjacent to one another in the vector space since the relationships between them are entirely disregarded. (2) For variables with a large number of unique categories, the dimensionality of the encoded vectors increases significantly. Training neural networks on high-dimensional representations increases computational costs and memory demands. (3) The naive allocation of one-hot codes to every unique category does not provide us with any information on how to convert high-dimensional vectors into informative low-dimensional representations. Thus, trial and error will be the main method for producing input features with reduced dimensions.

These drawbacks of one-hot encoding are overcome by neural network embeddings. Embeddings are real-valued vectors obtained by mapping discrete variables to a continuous vector space. They allow the representation of high-dimensional data in a low-dimensional space, making it possible to capture the underlying relationships between data points. One of the main benefits of using embeddings is that they can be learned and reused across models. Additionally, as embeddings are real-valued vectors, it is possible to find similar embeddings and group them into distinct clusters by calculating the distance or similarity between them.

The process of learning phoneme embeddings is combined with the F_0 prediction model and can be considered a simple matrix multiplication in which each phoneme is mapped into a dense vector. To learn phoneme embeddings, we first define a dictionary $\mathcal{D}$ and create entries for each phoneme label. Assuming that there are K unique phoneme labels, then $\mathcal{D}$ contains K *key-value* pairs. For each key, i.e., each unique phoneme label, a tuple formed by an integer number and a one-hot code is stored as the value. Whereas the integer number corresponds to the index of the label after ordering the phoneme labels in alphabetical order, the one-hot code represents the encoding of that label with a single 1 and $K-1$ 0s. These one-hot codes in the dictionary are used to define a matrix $\mathbf{D}$ as follows:

$$\mathbf{D}_{K\times K} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}, \quad (1)$$

where each row corresponds to a *key* in $\mathcal{D}$. Next, a simple network with K inputs, one input for every entry in $\mathcal{D}$, is utilized as the embedding layer on the front end of the F_0 prediction model. These inputs are then connected to activation functions within this layer. The activation functions employ the identity function $f(x) = x$. Therefore, the input value is the same as the output value. In other words, these activation functions serve the purpose of providing a platform for addition without altering the input values. The number of activation functions determines how many numbers are associated with each phoneme label, i.e., the dimension of phoneme embeddings. The weights on these connections eventually represent the phoneme embeddings. For simplicity and visualization purposes, we associate two numbers with each phoneme label. Thus, two linear activation functions are used, as depicted in Fig. 4.

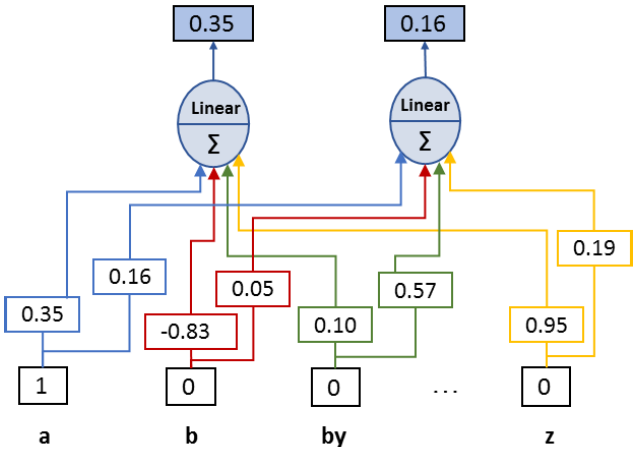


FIGURE 4. An embedding layer is used to map one-hot encoded phoneme labels into real-valued phoneme embeddings. The number of activation functions corresponds to the desired dimension of the phoneme embeddings, representing how many numbers are associated with each phoneme label. When an input value is zero, the layer sends only zeros to the activation functions; conversely, when it is one, the corresponding weight values are directed to the activation functions. As an example, the mapping of the one-hot encoded phoneme label /a/ into phoneme embeddings can be observed here.

It is evident that when an input value is zero, it only transmits zeros to the activation functions, and when it is one, the corresponding weight values are directed to the activation functions. In Fig. 4, the weights start with random values and are optimized through backpropagation, one step at a time, during the training stage when the network predicts F_0 values frame by frame (refer to Fig. 5). By utilizing the outputs of the activation functions, i.e., the resulting embeddings, as inputs to the F_0 prediction model and making predictions, we repeatedly update the weights on all connections until the network converges. Once the network is trained, the weights on the connections between the input layer and the linear activation functions become the real-valued phoneme embeddings. As the embeddings are learned concurrently with the F_0 prediction model, the resulting embeddings are tailored for the natural F_0 prediction task. This represents the main advantage of this approach in learning phoneme embeddings. It is important to note that we employ the same embedding network for each input phoneme label. This means that regardless of how long the input sequence is, we simply copy and use the exact same embedding network for each phoneme label. This provides us with identical embeddings for the same phoneme labels and the flexibility to handle input sequences of varying lengths.

2) CLUSTERING OF PHONEME EMBEDDINGS TO DEFINE PHONEME CATEGORIES

Replacing discrete phoneme labels with continuous phoneme embeddings increases space utilization and reduces

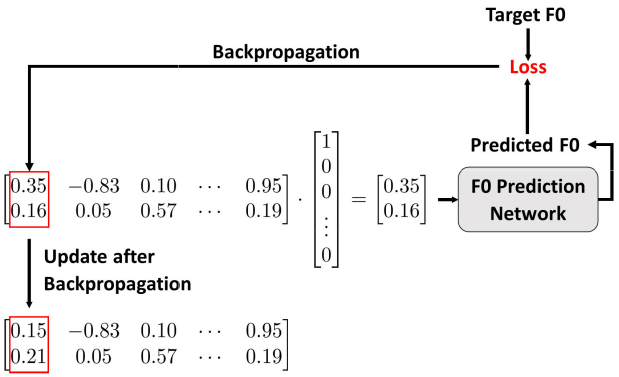


FIGURE 5. The process of learning phoneme embeddings is combined with the F_0 prediction model and can be considered as a straightforward matrix multiplication where each phoneme is mapped into a dense vector.

unnecessary parameters. In the multidimensional space of the embeddings, similar representations corresponding to phoneme labels with similar characteristics are found adjacent to each other. This allows us to group similar embeddings and, as a result, reduce the dimension of the features for $F0$ prediction. To identify the most similar embeddings and combine them into an appropriate number of clusters, a clustering process must be considered. In the realm of unsupervised learning, k-means clustering and hierarchical clustering are two widely used methods for grouping data points into discrete clusters. Whereas k-means clustering divides data into a predefined number of clusters, hierarchical clustering creates a hierarchical tree-like structure to represent the relationships between the clusters.

By focusing on the fast and straightforward k-means clustering algorithm, where the number of clusters is a hyperparameter that controls the dimension of the resulting features and by converting the assigned cluster labels to individual phoneme embeddings to one-hot codes, we can create new one-hot features for F_0 prediction. Therefore, unlike the approach of not employing phoneme embedding where a clear strategy for constructing low-dimensional input characteristics does not exist, learning and clustering of the phoneme embedding can lead to the creation of low-dimensional input representations in an efficient and general manner.

IV. EXPERIMENTAL EVALUATION

A. EXPERIMENTAL CONDITIONS

1) DATASET

Utterances in set $\mathcal{A}$ of the ATR speech dataset [61] were used to train and evaluate the systems in our experiments. The ATR dataset consists of 503 phonetically balanced Japanese sentences organized into 10 sets with 50 utterances each, except for the 10^{th} set that has 53 utterances, which are distinguished by alphabet letters. The existing sentences in set $\mathcal{A}$ were uttered by a Japanese male speaker with and without an electrolarynx to form a parallel dataset for EL and normal speech. The acoustic features, including F_0 values (in Hz) and spectral envelope, were extracted by the STRAIGHT analysis system [62]. The first 25 MCEP coefficients extracted for both speech types were used as the spectral features for time warping. Forced-aligned phoneme labels were also obtained by using the open-source Julius speech recognition system [63].

TABLE 1. Average correlation coefficients ($\bar{r}$) between ground truth and quantized F_0 patterns for normal speech training utterances for various quantization bit depths.

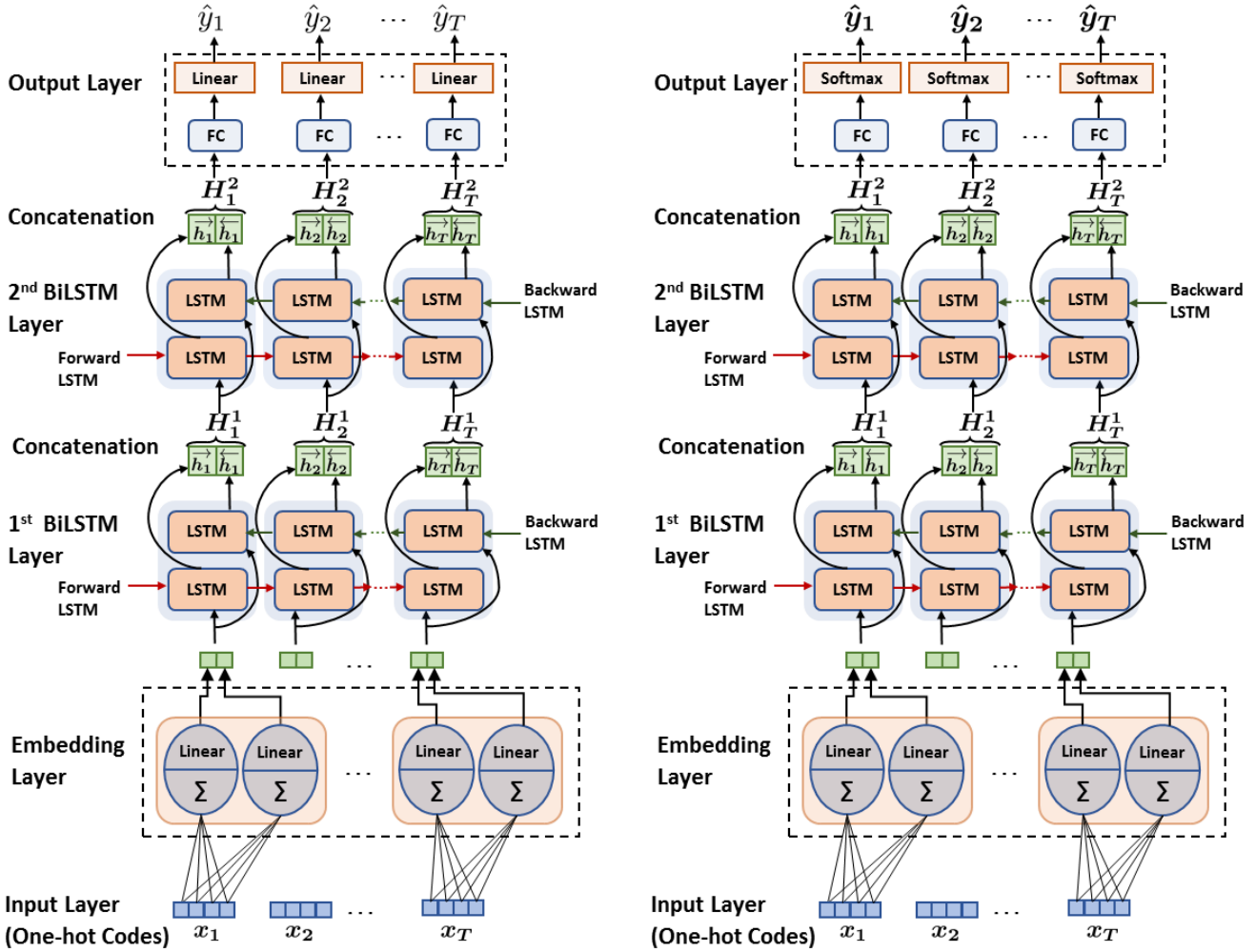
	Bit Depth						
	1	2	3	4	5	6	7
$\bar{r}$	0.83	0.95	0.98	0.98	**0.99**	**0.99**	**0.99**

2) NETWORK ARCHITECTURE

As depicted in Fig. 6, the architecture of the prediction models consisted of a stack of two BiLSTM layers followed by a fully connected (FC) dense layer. The number of hidden units of the LSTM layers in each direction was set to 64, which corresponded to 128 in both directions as the output dimension. For the regression model, the identity function was used as the activation function of the output layer, and the mean squared error (MSE) loss function was selected as the objective loss function to estimate the prediction errors. For the classification model, five bits were found to be sufficient on the basis of the average correlations obtained between the ground truth F_0 patterns and quantized ones for normal speech training utterances at various quantization bit depths (see Table 1). Hence, the 5-bit quantization of the target F_0 patterns was considered, and the softmax function was used as the activation function of the output layer to produce 32 class labels. The class with the highest probability according to the softmax output for each frame was always chosen to be decoded to a numerical F_0 value. Then, the cross-entropy (CE) loss function was employed to iteratively improve the classification accuracies. Both F_0 prediction models were implemented in Keras, with Google's TensorFlow serving as the back-end engine.

3) EXPERIMENTS

As the baseline system, we considered natural F_0 pattern prediction based on the MCEP coefficients extracted from the smoothed spectral envelopes. Subsequently, we introduced two additional systems to explore F_0 pattern prediction based on frame-wise phoneme information. In the first system, forced-aligned phoneme labels were used as input features for the F_0 prediction models, and the prediction accuracies were assessed on the basis of three distinct ways of conveying information through the input features. These features included one-hot codes representing (1) the full set of phoneme labels, (2) only the set of vowel labels, where consonants were substituted with the first succeeding vowel label in the transcriptions and (3) the speech syllables. In the second system, an embedding layer with two activation functions was added to the front-end of the prediction models



(a) **For regression predictive modeling.**

(b) **For classification predictive modeling.**

FIGURE 6. **Architectures of F_0 prediction networks based on BiLSTM layers.**

(i.e., for regression and classification predictive modelings), right before the first BiLSTM layers. These structures were trained separately to learn the mapping functions for transforming phoneme labels into natural F_0 patterns through transitioning from a two-dimensional embedding space. Once training was completed, one-hot encoded phoneme labels were fed frame-by-frame into the embedding layers, and the outputs of these layers were stored as the corresponding learned embedding vectors for F_0 prediction. This process was applied to the entire dataset, encompassing both EL and normal speech. Subsequently, the embedding vectors of the training utterances for both speech types were pooled together, and a k-means clustering model was trained to group embeddings into 8, 15, 22, 29, and 36 clusters on the basis of their proximity to the cluster centroids. Given that there were 36 forced-aligned phoneme labels in our dataset, the maximum number of clusters was set to 36. After clustering, embeddings within individual clusters were converted to one-hot codes. Finally, the embedding layers were detached, and the remaining prediction models were trained once again from scratch using the resulted one-hot features with their respective dimensions.

Using back-propagation through time (BPTT) and the Adam optimizer [64], we optimized the parameters of the prediction models (weights and biases) in all systems for utterance batches of size 32. The learning rate α, β_1 and β_2 were set to 0.001, 0.9 and 0.999, respectively. The dataset was divided into training, validation and evaluation sets for each type of speech. Thirty percent of the utterances were randomly chosen for the validation set and utilized for hyperparameter tuning. From the remaining utterances, 10 were randomly selected as the evaluation set to assess the final models' performance and calculate the accuracies of the predicted F_0 patterns compared with the ground truth data.

The overall accuracy was determined in terms of correlation over voiced frames by the fivefold cross-validation method. To ensure a fair comparison between the networks for regression and classification predictive modelings, the

low-pass-filtered continuous F_0 patterns were consistently used as the reference patterns. In other words, the decoded F_0 patterns from the classification predictive modeling were low-pass-filtered and compared with the same ground truth data used for the regression predictive modeling, not with the low-pass-filtered versions of the quantized target F_0 patterns. The ground truth continuous F_0 patterns were normalized to zero mean and unit variance using the statistics of the training sets.

Pearson's product-moment correlation was used to assess the strength and direction of the associations between the target and the predicted F_0 patterns. Let $\boldsymbol{y}$ and $\boldsymbol{y}'$ denote the target and corresponding predicted F_0 patterns, respectively. The Pearson's correlation coefficient r is defined as [65]

$$r_{\boldsymbol{y}\boldsymbol{y}'} = \frac{\mathrm{cov}(\boldsymbol{y}, \boldsymbol{y}')}{\sigma_{\boldsymbol{y}}\sigma_{\boldsymbol{y}'}} = \frac{\sum_{i=1}^{n} (y_i - \overline{\boldsymbol{y}}) \left(y'_i - \overline{\boldsymbol{y}'}\right)}{\sqrt{\sum_{i=1}^{n} (y_i - \overline{\boldsymbol{y}})^2} \sqrt{\sum_{i=1}^{n} \left(y'_i - \overline{\boldsymbol{y}'}\right)^2}}, \quad (2)$$

where $\overline{\boldsymbol{y}}$ and $\overline{\boldsymbol{y}'}$ are the respective mean values and n is the total length when only voiced frames are considered. Moreover, y_i and y'_i are the i^{th} values in the target and predicted F_0 patterns. In general, r ranges from -1 to 1, where values closer to -1 indicate a strong negative correlation, and those closer to 1 indicate a strong positive correlation. However, when dealing with the accuracies of the predicted F_0 patterns, we aim to maximize the positive correlation between the target and predicted F_0 patterns.

B. EXPERIMENTAL RESULTS

Average F_0 correlation coefficients for the baseline and the other two systems are presented in Table 2. It can be seen from the first row of the table that, regardless of the type of predictive modeling, F_0 patterns with high accuracy ($\bar{r} \approx 0.95$) can be predicted for normal speech on the basis of the spectral features collected from a limited number of utterances. This highlights the existence of guiding acoustic cues in the spectrum of normal speech and their leading role in predicting paralinguistic features. As for EL speech, even though the spectral envelope varies according to individual phonemes, it is the absence of these acoustic cues in the spectrum that makes the prediction task very challenging and causes a significant drop in the average r value.

Moving on to the rows representing F_0 pattern prediction based on the phoneme labels indicates the following interesting findings: (1) By utilizing the full set of phoneme labels and allocating unique one-hot codes to them, we can successfully predict F_0 patterns for both EL and normal speech, with average r values comparable to those calculated for the predictions based on the spectral features. (2) It is evident that the existing gap between the average r values for EL and normal speech has been reduced (compare, for instance, 0.61 vs 0.95 for the MCEP coefficients in the case of regression predictive modeling with 0.60 vs 0.69 for the set of all phoneme labels). This indicates that, as opposed to the spectral features where they are substantially discrepant, the obtained features out of phoneme labels are less discrepant because they are formed out of exactly the same utterances with exactly the same linguistic content. Hence, they can be considered a type of shared features between EL and normal speech. However, owing to the fact that EL speech is produced at a low rate with many short pauses, i.e., lower articulation speed, there are still some discrepancies in these features. Having shared features can make it possible to provide the networks with more training data and let them improve their performance in approximating the unknown mapping functions. This can be observed in the last column of the table where the obtained average r values for EL speech have been improved by training the prediction networks using training datasets of EL and normal speech jointly.

Substituting consonants with the first succeeding vowel label in the transcriptions and using low-dimensional input representations for the prediction networks, where the dimension was equal to the cardinality of the vowel set, reveals that despite hindering the networks from information related to the consonants, no degradation in the average r value for EL speech was observed. However, for normal speech, a decrease in the average r value took place. This decrease might have occurred owing to a local minimum problem, where the prediction networks might have become stuck in a suboptimal solution during training, possibly owing to the shape of the error function.

Considering prediction accuracies for the speech syllables, it is evident that the estimates of the F_0 values were not satisfactory for either EL speech or healthy speech. This can be explained by the fact that when prediction networks are trained on speech syllables, we no longer consider the original phoneme labels or a subset of them. Instead, we consider speech boundaries. That is, rather than mapping sequences of phoneme labels into natural F_0 patterns, extremely low-dimensional features representing the type of individual speech section or syllable, whether a syllable is formed out of a vowel or a consonant followed by a vowel, are mapped into F_0 patterns. This information is evidently insufficient for predicting accurate F_0 patterns. Note that $\bar{r} = 0.56$ and $\bar{r} = 0.51$ were obtained when the prediction models were trained on a very limited amount of training data. These values were improved by using shared features from both speech types and increasing the number of training utterances.

As shown in Table 2, using phoneme embeddings with the highest dimension (the last row of the table), corresponding to defining as many distinct clusters as the cardinality of the set of all phoneme labels (36 clusters in our study, as depicted in Figs. 7a and 8a), yields average r values that are (almost) equal to the ones presented in the second row of the table, regardless of the type of predictive modeling for both EL and normal speech. This was expected because, clearly, allocating 36 labels to 36 clusters does not change the conditions. Focusing on the case where embeddings were

TABLE 2. Correlation coefficients for baseline and all other defined systems using regression and classification predictive modeling for various types of input feature. For the classification case, the target F_0 patterns were quantized to 5 bits (32 levels).

Input Features \ Prediction Type		EL Speech		Normal Speech		Joint (EL + Normal)	
		Regression $\overline{r}$	Classification $\overline{r}$	Regression $\overline{r}$	Classification $\overline{r}$	Regression $\overline{r}$	Classification $\overline{r}$
Spectral Features	MCEP	0.61	0.60	0.95	0.94	-	-
Phoneme Labels	All Labels	0.60	0.61	0.69	0.68	**0.65**	**0.65**
	All Vowels	0.60	0.61	0.60	0.58	**0.61**	**0.63**
	Speech Syllables	0.56	0.51	0.62	0.61	**0.57**	**0.59**
Clustered Phoneme Embeddings	8 Clusters	0.56	0.56	0.65	0.68	**0.65**	**0.66**
	15 Clusters	0.61	0.60	0.71	0.67	**0.63**	**0.64**
	22 Clusters	0.58	0.59	0.67	0.71	**0.66**	**0.66**
	29 Clusters	0.59	0.61	0.68	0.72	**0.63**	**0.64**
	36 Clusters	0.59	0.61	0.68	0.68	**0.64**	**0.65**

allocated to 29 clusters (see Figs. 7b and 8b), we observe that although the dimension of the input features has been reduced by seven units, no changes in the prediction behavior of the models are observed. It turns out that the most accurate F_0 patterns are not necessarily predicted when individual clusters are considered; it could also be achieved when some nearby clusters with nearly similar data points are merged together. This confirms that learning phoneme embeddings and creating input features for F_0 prediction based on the clustering of these embeddings can effectively reduce the difficulty of finding a strategy for lowering the dimension of the input features. In Tables 3 and 7, we have summarized phonemes in these individual 29 clusters for both modeling techniques.

It is appealing to see that by classifying phoneme embeddings into 15 clusters, which is less than 50% of the total number of phoneme labels, as depicted in Figs. 7d and 8d, no degradation in average r values is observed. This indicates that even though, after applying k-means clustering, the data points with similar characteristics are aggregated close to each other and form distinct clusters, the intercluster dissimilarities are so small that they can be easily neglected among some neighboring clusters. Hence, by allowing similar clusters to merge and dissimilar clusters to separate, the embedding space can be partitioned into fewer but very large clusters, or more precisely, hyperclusters. This significantly reduces the dimension of the input representations and embeds very useful information on vowel–consonant combinations and the proximities among them in the embedding space to the resulting features.

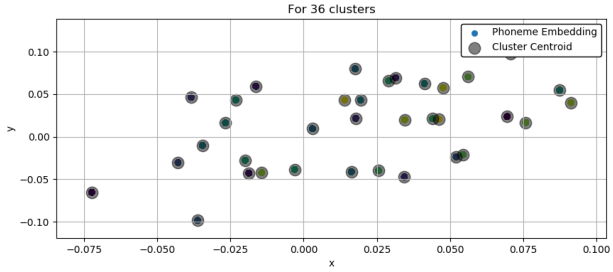
Considering prediction accuracies in the case of clustering embeddings into eight clusters, as illustrated in Figs. 7e and 8e, we can observe a drop in the average r values for both EL and normal speech. This indicates that a minimum number of clusters is always required to separate the most dominant information for the prediction task from the irrelevant and misleading information that markedly degrades the prediction accuracies. For instance, as shown in Tables 6 and 10, no distinct clusters have been allocated to silence (sil). Hence, it is absolutely important not to violate this limit on the minimum number of clusters and wisely make a trade-off between the number of clusters and the resulting correlations. Lastly, it is worth mentioning that, irrespective of the type of predictive modeling, the prediction accuracies have been effectively improved by joint training. Therefore, even in the case of having eight clusters, the resulting r values are comparable to those for a larger number of clusters.

TABLE 3. Phonemes in individual clusters (for 29 clusters, regression).

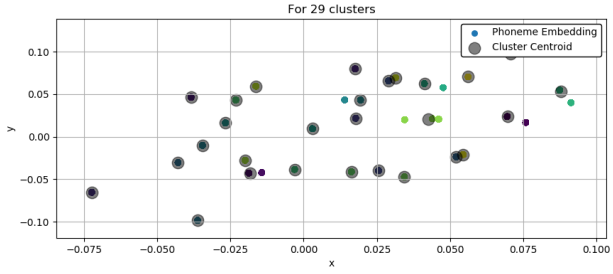
	Cluster Index							
	1	2	3	4	5	6	7	8
Phoneme Labels	f k	g py	d	j	N gy	ch	hy my ry	n
	Cluster Index							
	9	10	11	12	13	14	15	16
Phoneme Labels	sil	i	e	q	h	r	m	z
	Cluster Index							
	17	18	19	20	21	22	23	24
Phoneme Labels	b	dy p	sh	a	o	ny u	w	s
	Cluster Index							
	25	26	27	28	29			
Phoneme Labels	t	ts	ky	by	y			

TABLE 4. Phonemes in individual clusters (for 22 clusters, regression).

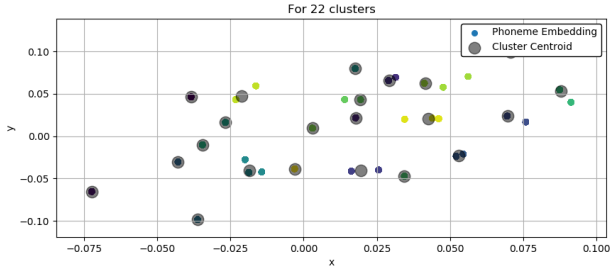
	Cluster Index							
	1	2	3	4	5	6	7	8
Phoneme Labels	f k ts	g py	d	j ky	N by gy	ch t	hy my ry	n w
	Cluster Index							
	9	10	11	12	13	14	15	16
Phoneme Labels	sil	i	e	q	h	r s	m	z
	Cluster Index							
	17	18	19	20	21	22		
Phoneme Labels	b	dy p	sh	a	o y	ny u		



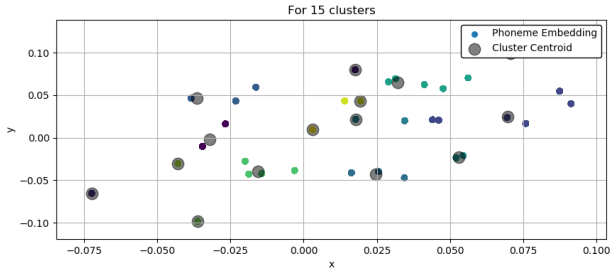
(a) **Clustering into 36 clusters.**



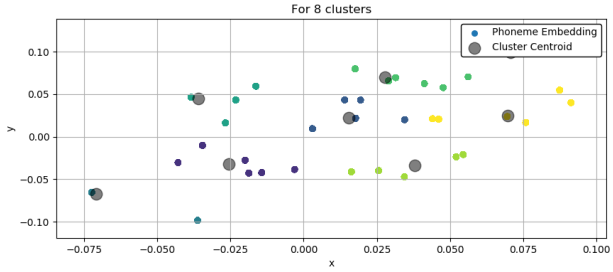
(b) **Clustering into 29 clusters.**



(c) **Clustering into 22 clusters.**

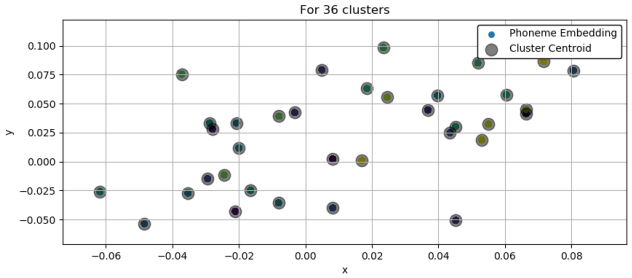


(d) **Clustering into 15 clusters.**

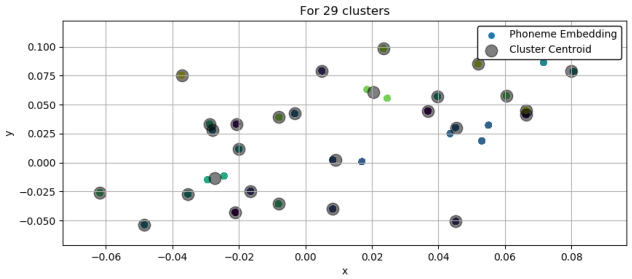


(e) **Clustering into 8 clusters.**

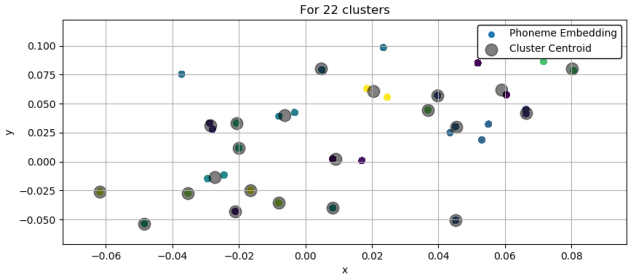
FIGURE 7. **Clustering of phoneme embeddings into 8, 15, 22, 29, and 36 clusters. The 2D embeddings were learned on the basis of the regression predictive modeling.**



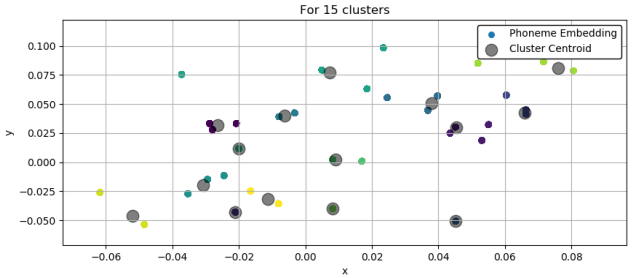
(a) **Clustering into 36 clusters.**



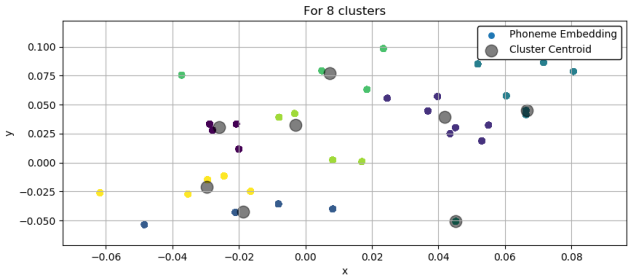
(b) **Clustering into 29 clusters.**



(c) **Clustering into 22 clusters.**



(d) **Clustering into 15 clusters.**



(e) **Clustering into 8 clusters.**

FIGURE 8. **Clustering of phoneme embeddings into 8, 15, 22, 29, and 36 clusters. The 2D embeddings were learned on the basis of the classification predictive modeling.**

V. DISCUSSION

Successful reintegration to society is critical if we are to improve the QoL of laryngectomees. EL speech serves as a voice restoration technique after TL surgery, offering reasonable intelligibility but artificial sound quality. The objective

TABLE 5. Phonemes in individual clusters (for 15 clusters, regression).

	Cluster Index							
	1	2	3	4	5	6	7	8
Phoneme Labels	b f k ts	g py	d	a j ky	N by gy o y	ch sh t	dy hy ny p ry u	n w
	Cluster Index							
	9	10	11	12	13	14	15	
Phoneme Labels	sil	i my	e	q z	h	r s	m	

TABLE 6. Phonemes in individual clusters (for 8 clusters, regression).

	Cluster Index							
	1	2	3	4	5	6	7	8
Phoneme Labels	b f h k q ts	g i m my py	d sil	a j ky z	N by e gy o y	ch r s sh t	dy hy ny p ry u	n w

TABLE 7. Phonemes in individual clusters (for 29 clusters, classification).

	Cluster Index							
	1	2	3	4	5	6	7	8
Phoneme Labels	hy ky ny u	sil	n	q	by	my y	b ch	a
	Cluster Index							
	9	10	11	12	13	14	15	16
Phoneme Labels	N	d	h	py z	s	t	m	f
	Cluster Index							
	17	18	19	20	21	22	23	24
Phoneme Labels	gy	p w	k	i	g	j	sh	r
	Cluster Index							
	25	26	27	28	29			
Phoneme Labels	o	ts	e	dy	ry			

TABLE 8. Phonemes in individual clusters (for 22 clusters, classification).

	Cluster Index							
	1	2	3	4	5	6	7	8
Phoneme Labels	hy ky ny u	sil	n o	q	by r ry	my y	b ch	a e
	Cluster Index							
	9	10	11	12	13	14	15	16
Phoneme Labels	N	d	h	dy py z	s sh	t	m	f
	Cluster Index							
	17	18	19	20	21	22		
Phoneme Labels	gy ts	p w	k	i	g	j		

TABLE 9. Phonemes in individual clusters (for 15 clusters, classification).

	Cluster Index							
	1	2	3	4	5	6	7	8
Phoneme Labels	hy ky ny u	sil	n o ts	q	by p r ry	my y	b ch g	a e i
	Cluster Index							
	9	10	11	12	13	14	15	
Phoneme Labels	N k w	d f	h j	dy gy py z	s sh	t	m	

TABLE 10. Phonemes in individual clusters (for 8 clusters, classification).

	Cluster Index							
	1	2	3	4	5	6	7	8
Phoneme Labels	N hy k ky ny u w	d h sil t	dy gy n o py ts z	q	by p r ry	my s sh y	b ch f g j	a e i m

of this study was to explore F_0 prediction utilizing frame-wise phoneme information. This investigation encompassed the prediction of both continuous F_0 patterns, approached as a regression problem, and discretized F_0 patterns, achieved through a multilevel classification problem employing the median-cut algorithm for quantization. After aligning forced-aligned phoneme sequences of EL and normal speech considering both acoustic and phoneme-level linguistic information, we implemented a prediction system based on BiLSTM RNNs. This system was designed to map one-hot encoded phoneme labels into F_0 patterns.

Our study demonstrated the capability of our system to predict F_0 patterns not only on the basis of the full set of phoneme labels but also using a subset of phoneme labels, which is equivalent to employing low-dimensional input representations. Additionally, our investigation revealed that frame-wise phoneme information can be regarded as shared features between EL and normal speech. Consequently, by augmenting training sets for EL speech with phoneme information from normal speech, we observed improved prediction accuracies. Beyond direct predictions based on phoneme labels, we expanded our analysis to include real-valued phoneme embeddings in a continuous vector space. Furthermore, we conducted predictions for clustered embeddings representing low-dimensional input representations. This confirmed that the learning of phoneme embeddings and the creation of input features for F_0 prediction based on the clustering of these embeddings can effectively alleviate the challenge of finding a strategy to reduce the dimension of the input features. The exploration of different

clustering algorithms and selection of the most effective algorithm remain intriguing avenues for future research.

The most important factors determining the prediction accuracies are how input features are created, the type of information represented by the input features, and the dimensionality of the input features. Considering learning and clustering of the phoneme embeddings as a practical approach to creating such features, it is crucial to identify both the optimal clustering algorithm and the most effective dimensionality for the resultant features. Although we acknowledge that k-means may not be the ultimate choice for phoneme clustering in our system, its effectiveness stands out, evident in comparable correlation coefficients with and without clustering in both regression and classification predictive modeling. The exploration of a range of k values enabled us to carefully navigate a trade-off between the dimensionality of input features and the corresponding prediction accuracies.

Assuming that the underlying distributions of the training and test data are similar, which is a fundamental assumption in ML, the robustness of the proposed F_0 prediction system based on classification predictive modeling is significantly affected by the bit depth used for quantizing the target F0 patterns. The bit depth has a direct impact on both the quantization noise and the dynamic range of the quantized F_0 patterns. Using a larger number of bits facilitates a finer definition of levels, leading to reduced quantization errors.

Limiting ourselves to a minimal amount of EL speech training data served as a strategic choice. This decision aimed to demonstrate that if the system excels in predicting accurate F_0 patterns in these challenging approaches, it should logically exhibit even better performance in simpler situations with a more substantial amount of training data. Leveraging frame-wise phoneme information as a bridge between EL and normal speech, we can augment the training data with normal utterances in a parallel or nonparallel manner. Although data augmentation helps in learning good data representations during training and improves F_0 prediction accuracies, it cannot completely make up for the lack of original EL data. Therefore, to ensure robustness, we employed a rigorous fivefold cross-validation approach to evaluate the performance of our F_0 prediction models. This helped prevent overfitting and allowed for the estimation of the generalized performance.

Finally, in the context of developing health-related artificial intelligence (AI) systems, particularly those geared towards speech rehabilitation after TL, it is crucial to underscore several ethical considerations that should underpin their implementation. Privacy and informed consent are paramount, requiring the secure handling of patient data and explicit consent for its utilization. The principles of transparency and explainability are equally crucial, ensuring that patients fully comprehend the purpose, capabilities, and limitations of such AI systems. To mitigate bias, it is essential to use diverse datasets for training and conduct regular assessments to identify and rectify potential biases in predictions. Upholding patient autonomy and empowerment is critical, emphasizing individuals' control over their data and decisions regarding AI-based interventions. Collaborating closely with healthcare professionals is a key strategy to ensure that AI enhances rather than replaces their expertise.

VI. CONCLUSION

We demonstrated that we can successfully predict natural F_0 patterns using frame-wise phoneme information. The type of target F_0 pattern does not substantially impact the prediction accuracies. Hence, we can use either regression or classification predictive modeling techniques with equal preference to map frame-wise phoneme information into natural F_0 patterns. By predicting F_0 patterns based on phoneme information, we can reduce the existing gap between the prediction accuracies for EL and normal speech. Moreover, we can consider input features created from phoneme information as a sort of shared features between EL and normal speech and further improve the prediction accuracies by training the prediction networks using training datasets of EL and normal speech jointly. Additionally, learning phoneme embeddings and creating input features for F_0 prediction based on the clustering of these embeddings can effectively reduce the difficulty of finding a strategy for lowering the dimension of the input features.

ACKNOWLEDGMENT

The authors express their gratitude to Dr. Hirokazu Kameoka, Dr. Kazuhiro Kobayashi, and Dr. Kou Tanaka for their generous support and helpful discussions.

REFERENCES

[1] *Speech & Swallow Therapy*. Accessed: Mar. 29, 2024. [Online]. Available: https://thancguide.org/cancer-basics/after-treatment/rehabilitation/speech-swallow-therapy/

[2] R. Kaye, C. G. Tang, and C. F. Sinclair, "The electrolarynx: Voice restoration after total laryngectomy," *Med. Devices, Evidence Res.*, vol. 10, pp. 133–140, Jun. 2017.

[3] S. Bien, A. Rinaldo, C. E. Silver, J. J. Fagan, L. W. Pratt, C. Tarnowska, E. Towpik, N. Weir, B. J. Folz, and A. Ferlito, "History of voice rehabilitation following laryngectomy," *Laryngoscope*, vol. 118, no. 3, pp. 453–458, Mar. 2008.

[4] K. J. Lorenz, "Rehabilitation after total laryngectomy—A tribute to the pioneers of voice restoration in the last two centuries," *Frontiers Med.*, vol. 4, p. 81, Jun. 2017, doi: 10.3389/fmed.2017.00081.

[5] B. Thiagarajan, "Voice rehabilitation following total laryngectomy," *Online J. Otolaryngol.*, vol. 5, pp. 48–59, Aug. 2014.

[6] E. Babin, D. Beynier, D. Le Gall, and M. Hitier, "Psychosocial quality of life in patients after total laryngectomy," *Rev. Laryngol. Otol. Rhinol.*, vol. 130, no. 1, pp. 29–34, Feb. 2009.

[7] 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, Language Process.*, vol. 22, no. 1, pp. 172–183, Jan. 2014.

[8] L. Serrano, D. Tavarez, X. Sarasola, S. Raman, I. Saratxaga, E. Navas, and I. Hernaez, "LSTM based voice conversion for laryngectomees," in *Proc. IberSPEECH*, Barcelona, Spain, Nov. 2018, pp. 122–126.

[9] A. K. Fuchs, M. Hagmüller, and G. Kubin, "The new bionic electrolarynx speech system," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 5, pp. 952–961, Aug. 2016.

[10] G. S. Meltzner and R. E. Hillman, "Impact of aberrant acoustic properties on the perception of sound quality in electrolarynx speech," *J. Speech, Lang., Hearing Res.*, vol. 48, no. 4, pp. 766–779, Aug. 2005.

[11] B. Shute, "Perceptions of artificial larynx reliability according to laryngectomees and speech-language pathologists," Ph.D. dissertation, School Prof. Stud. (SPS), Gonzaga Univ., Spokane, WA, USA, Feb. 2003.
[12] N. Uemi, T. Ifukube, M. Takahashi, and J. Matsushima, "Design of a new electrolarynx having a pitch control function," in *Proc. ROMAN*, Nagoya, Japan, Jul. 1994, pp. 198–203.
[13] N. Uemi, T. Ifukube, M. Takahashi, and J. Matsushima, "Development of an electrolarynx having a pitch frequency control function by using expiration pressure," *JJME*, vol. 33, no. 1, pp. 7–14, 1995.
[14] B. Roubeau, C. Chevrie-Muller, and J. L. Saint Guily, "Electromyographic activity of strap and cricothyroid muscles in pitch change," *Acta Oto-Laryngologica*, vol. 117, no. 3, pp. 459–464, Jan. 1997.
[15] H. Takahashi, M. Nakao, T. Okusa, Y. Hatamura, Y. Kikuchi, and K. Kaga, "Pitch control with finger pressure for electrolaryngeal or intra-mouth vibrating speech," *Jpn. J. Logopedics Phoniatrics*, vol. 42, no. 1, pp. 1–8, 2001.
[16] Y. Kikuchi and H. Kasuya, "Development and evaluation of pitch adjustable electrolarynx," in *Proc. Speech Prosody*, Nara, Japan, Mar. 2004, pp. 1–4.
[17] E. A. Goldstein, J. T. Heaton, J. B. Kobler, G. B. Stanley, and R. E. Hillman, "Design and implementation of a hands-free electrolarynx device controlled by neck strap muscle electromyographic activity," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 2, pp. 325–332, Feb. 2004.
[18] H. Takahashi, M. Nakao, Y. Kikuchi, and K. Kaga, "Alaryngeal speech aid using an intra-oral electrolarynx and a miniature fingertip switch," *Auris Nasus Larynx*, vol. 32, no. 2, pp. 157–162, Jun. 2005.
[19] H. Takahashi, M. Nakao, Y. Kikuchi, and K. Kaga, "Intra-oral pressure–based voicing control of electrolaryngeal speech with Intra-oral vibrator," *J. Voice*, vol. 22, no. 4, pp. 420–429, Jul. 2008.
[20] H. L. Kubert, C. E. Stepp, S. M. Zeitels, J. E. Gooey, M. J. Walsh, S. R. Prakash, R. E. Hillman, and J. T. Heaton, "Electromyographic control of a hands-free electrolarynx using neck strap muscles," *J. Commun. Disorders*, vol. 42, no. 3, pp. 211–225, May 2009.
[21] Y. Saikachi, "Development, perceptual evaluation, and acoustic analysis of amplitude-based F_0 control in electrolarynx speech," Ph.D. dissertation, Harvard-MIT HST, MIT, Cambridge, MA, USA, Sep. 2009.
[22] W. De Armas, K. A. Mamun, and T. Chau, "Vocal frequency estimation and voicing state prediction with surface EMG pattern recognition," *Speech Commun.*, vols. 63–64, pp. 15–26, Sep. 2014.
[23] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 2, pp. 131–142, Mar. 1998.
[24] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 15, no. 8, pp. 2222–2235, Nov. 2007.
[25] J. Schmidt, M. R. G. Marques, S. Botti, and M. A. L. Marques, "Recent advances and applications of machine learning in solid-state materials science," *Npj Comput. Mater.*, vol. 5, no. 1, pp. 1–36, Aug. 2019.
[26] L. Diener, M. Janke, and T. Schultz, "Direct conversion from facial myoelectric signals to speech using deep neural networks," in *Proc. IJCNN*, Killarney, Ireland, Jul. 2015, pp. 1–7.
[27] M. Janke and L. Diener, "EMG-to-speech: Direct generation of speech from facial electromyographic signals," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 12, pp. 2375–2385, Dec. 2017.
[28] K. Kobayashi and T. Toda, "Electrolaryngeal speech enhancement with statistical voice conversion based on CLDNN," in *Proc. 26th Eur. Signal Process. Conf. (EUSIPCO)*, Rome, Italy, Sep. 2018, pp. 2115–2119.
[29] L. Diener, T. Umesh, and T. Schultz, "Improving fundamental frequency generation in EMG-to-speech conversion using a quantization approach," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Singapore, Dec. 2019, pp. 682–689.
[30] M.-C. Yen, W.-C. Huang, K. Kobayashi, Y.-H. Peng, S. Tsai, Y. Tsao, T. Toda, J.-S. Jang, and H.-M. Wang, "Mandarin electrolaryngeal speech voice conversion with sequence-to-sequence modeling," in *Proc. ASRU*, Cartagena, Colombia, Dec. 2021, pp. 650–657.
[31] D. Ma, L. P. Violeta, K. Kobayashi, and T. Toda, "Two-stage training method for Japanese electrolaryngeal speech enhancement based on sequence-to-sequence voice conversion," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Doha, Qatar, Jan. 2023, pp. 949–954.
[32] W.-C. Huang, T. Hayashi, Y.-C. Wu, H. Kameoka, and T. Toda, "Voice transformer network: Sequence-to-sequence voice conversion using transformer with text-to-speech pretraining," in *Proc. INTERSPEECH*, Shanghai, China, Oct. 2020, pp. 4676–4680.
[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.* Long Beach, CA, USA: Curran Associates, vol. 30, 2017, pp. 6000–6010.
[34] P. Kidger and T. Lyons, "Universal approximation with deep narrow networks," 2019, *arXiv:1905.08539*.
[35] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Istanbul, Turkey, Jun. 2000, pp. 1315–1318.
[36] W. Ping, K. Peng, A. Gibiansky, S. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," in *Proc. ICLR*, Vancouver, BC, Canada, Apr. 2018, pp. 510–525.
[37] J. Sotelo, S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. C. Courville, and Y. Bengio, "Char2Wav: End-to-end speech synthesis," in *Proc. ICLR*, Toulon, France, Apr. 2017.
[38] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. Saurous, "TacoTron: Towards end-to-end speech synthesis," in *Proc. INTERSPEECH*, Stockholm, Sweden, 2017, pp. 4006–4010.
[39] J. Shen, R. Pang, R. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions," in *Proc. ICASSP*, Calgary, AB, Canada, Apr. 2018, pp. 4779–4783.
[40] M. Eshghi, K. Tanaka, K. Kobayashi, H. Kameoka, and T. Toda, "An investigation of features for fundamental frequency pattern prediction in electrolaryngeal speech enhancement," in *Proc. 10th ISCA Workshop Speech Synth.*, Vienna, Austria, Sep. 2019, pp. 251–256.
[41] M. Eshghi, K. Kobayashi, K. Tanaka, H. Kameoka, and T. Toda, "Phoneme embeddings on predicting fundamental frequency pattern for electrolaryngeal speech," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Auckland, New Zealand, Dec. 2020, pp. 572–577.
[42] *Provox TruTone Plus ElectroLarynx*. Accessed: Mar. 29, 2024. [Online]. Available: https://www.atosmedical.us/product/provox-trutone-plus-electrolarynx
[43] W. De Armas, "Vocal frequency estimation and voicing state prediction with surface EMG pattern recognition," M.S. thesis, Dept. Biomed. Eng., UToronto, Toronto, ON, Canada, Jul. 2013.
[44] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech," *Speech Commun.*, vol. 54, no. 1, pp. 134–146, Jan. 2012.
[45] X. Shao and B. Milner, "Pitch prediction from MFCC vectors for speech reconstruction," in *Proc. ICASSP*, vol. 1, Montreal, QC, Canada, May 2004, p. 97.
[46] B. Milner, X. Shao, and J. Darch, "Fundamental frequency and voicing prediction from MFCCs for speech reconstruction from unconstrained speech," in *Proc. INTERSPEECH*, Lisbon, Portugal, Sep. 2005, pp. 321–324.
[47] K. Nakamura, M. Janke, M. Wand, and T. Schultz, "Estimation of fundamental frequency from surface electromyographic data: EMG-to-F_0," in *Proc. ICASSP*, Prague, Czech Republic, May 2011, pp. 573–576.
[48] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, Montreal, QC, Canada, Dec. 2014, pp. 3104–3112.
[49] J.-X. Zhang, Z.-H. Ling, Y. Jiang, L.-J. Liu, C. Liang, and L.-R. Dai, "Improving sequence-to-sequence voice conversion by adding text-supervision," in *Proc. ICASSP*, Brighton, U.K., May 2019, pp. 6785–6789.
[50] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
[51] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.
[52] A. Sakurai and K. Hirose, "Detection of phrase boundaries in Japanese by low-pass filtering of fundamental frequency contours," in *Proc. 4th Int. Conf. Spoken Lang. Process.*, vol. 2, Philadelphia, PA, USA, Oct. 1996, pp. 817–820.

[53] P. Heckbert, "Color image quantization for frame buffer display," in *Proc. SIGGRAPH*, Boston, MA, USA, Jul. 1982, pp. 297–307.
[54] T. Toda, M. Nakagiri, and K. Shikano, "Statistical voice conversion techniques for body-conducted unvoiced speech enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 9, pp. 2505–2517, Nov. 2012.
[55] J. C. Richards and R. W. Schmidt, *Longman Dictionary of Language Teaching and Applied Linguistics*, 4th ed. London, U.K.: Pearson, 2010, pp. 417–474.
[56] *Consonant*. Accessed: Mar. 29, 2024. [Online]. Available: https://dictionary.cambridge.org/dictionary/english/consonant
[57] I. Thompson, M. Swan, and B. Smith, *Learner English: A Teacher's Guide to Interference and Other Problems* (Cambridge Handbooks for Language Teachers), 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2001, pp. 296–309.
[58] K. Ohata, "Phonological differences between Japanese and English: Several potentially problematic areas of pronunciation for Japanese ESL/EFL learners," *Asian EFL J.*, vol. 6, pp. 1–19, Jan. 2004.
[59] B. Kavanagh, "The phonemes of Japanese and English: A contrastive analysis study," *J. Aomori Univ. Health Welf.*, vol. 8, no. 2, pp. 283–292, Dec. 2007.
[60] M. Mutsukawa, "The phonology of Japanese," *J. English Linguistics*, vol. 31, no. 1, pp. 264–274, Sep. 2018.
[61] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," *Speech Commun.*, vol. 9, no. 4, pp. 357–363, Aug. 1990.
[62] H. Kawahara, J. Estillc, and O. Fujimurad, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT," in *Proc. MAVEBA*, Florence, Italy, Sep. 2001, pp. 59–64.
[63] A. Lee, T. Kawahara, and K. Shikano, "Julius—An open source real-time large vocabulary recognition engine," in *Proc. EUROSPEECH*, vol. 3, Aalborg, Denmark, Sep. 2001, pp. 1691–1694. [Online]. Available: https://github.com/julius-speech/julius
[64] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[65] D. J. Hermes, "Measuring the perceptual similarity of pitch contours," *J. Speech, Lang., Hearing Res.*, vol. 41, no. 1, pp. 73–82, Feb. 1998.

MOHAMMAD ESHGHI received the B.Sc. degree in electrical engineering from the University of Najafabad, Isfahan, Iran, in 2011, and the M.Sc. degree in electrical engineering from Technische Universität Berlin (TU-Berlin), Germany, in 2015. He is currently pursuing the Ph.D. degree in speech processing with the Graduate School of Information Science, Nagoya University, Japan. His research interests include speech processing, signal processing for medical applications, deep learning, and information theory. He is a Student Member of the International Speech Communication Association (ISCA) and the Acoustical Society of Japan (ASJ).

TOMOKI TODA (Senior Member, IEEE) received the B.E. degree from Nagoya University, Japan, in 1999, and the M.E. and D.E. degrees from Nara Institute of Science and Technology (NAIST), Japan, in 2001 and 2003, respectively. He was a Research Fellow of the Japan Society for the Promotion of Science, from 2003 to 2005. He was an Assistant Professor with NAIST, from 2005 to 2011, where he was an Associate Professor, from 2011 to 2015. Since 2015, he has been a Professor with the Information Technology Center, Nagoya University. His research interests include statistical approaches to speech and audio processing. He has received more than ten article/achievement awards, including the IEEE SPS 2009 Young Author Best Paper Award and the 2013 EURASIP-ISCA Best Paper Award (*Speech Communication* journal).