10th ISCA Speech Synthesis Workshop
20-22 September 2019, Vienna, Austria

An Investigation of Features for Fundamental Frequency Pattern Prediction in Electrolaryngeal Speech Enhancement

Mohammad Eshghi[1], *Kou Tanaka*[2], *Kazuhiro Kobayashi*[3], *Hirokazu Kameoka*[2], *and Tomoki Toda*[3]

[1]Graduate School of Information Science, Nagoya University, Japan
[2]NTT Communication Science Laboratories, NTT Corporation, Japan
[3]Information Technology Center, Nagoya University, Japan

{mohammad.eshghi, kobayashi.kazuhiro}@g.sp.m.is.nagoya-u.ac.jp,
{tanaka.ko, kameoka.hirokazu}@lab.ntt.co.jp, tomoki@icts.nagoya-u.ac.jp

Abstract

Despite abundance of research, natural voice restoration after total laryngectomy (i. e., removal of the vocal folds of the larynx), has remained a challenge. A typical way of producing a relatively intelligible speech for patients suffering from this inability is to use an electrolarynx. However, the outcome voice sounds artificial and has "robotic" quality owing to constant fundamental frequency (F_0) patterns generated by the electrolarynx. In existing frameworks on natural F_0 patterns prediction, a model is trained on a massive amount of parallel training data to find a mapping that maps spectral features of the source speech into F_0 contours of the target speech. However, creating big datasets for electrolaryngeal (EL) speech is considered as a cumbersome and expensive task. Moreover, EL speech spectral features are significantly different from spectral features of the normal speech, and therefore, it is not straightforward to effectively use easily available normal speech datasets in training of the model for EL speech. Consequently, the quality of the models could be still low due to the lack of sufficient training data. To address this problem, we investigate F_0 pattern prediction based on other features that could be shared between normal speech and EL speech. By using shared input features, we would be to train the prediction model using a large amount of training data. As such features, in this work, we examine F_0 prediction accuracy based on phoneme-related features. The findings show that by considering phoneme labels for both vowels and consonants and one-hot encoding of these labels, we are able to predict F_0 contours with high correlation coefficients.

Index Terms: electrolaryngeal speech, speech enhancement, fundamental frequency pattern prediction, statistical voice conversion, phoneme labels, recurrent neural network

1. Introduction

In human societies, the ability to communicate with others to convey messages and express emotional feelings is of paramount importance. Although different elements can determine the quality of life (QoL) across societies, undoubtedly, the ability to communicate is one of the key factors that can significantly influence the QoL. In general, an speech signal is the product of four systems [1]: 1. respiratory (air generator), 2. phonatory (vibrating apparatus), 3. resonatory (resonance modulator) and 4. articulatory (articulating tract). During the production of voiced speech segments such as voiced consonants and vowels, the air flow expelled from the lungs sets the vocal folds into vibration. These vibrations generate sound waves and would subsequently get modulated by the shape of the vocal tract and articulatory movement. However, in patients with larynx cancer, vocal folds are sometimes completely removed from the larynx (i. e., total laryngectomy), and hence, production of voiced speech segments is impossible. Given that the phonetic system of most languages are notably consisted of voiced consonants and vowels, the absence of this acoustic feature would lead to marked voice abnormalities and decreased intelligibility.

Over the past decades, many voice restoration techniques have been proposed to fill the gap of vibrating apparatus and re-produce speech. Amongst different available techniques, EL speech has been considered as a viable method for producing relatively intelligible voices by laryngectomees. In this method, a battery operated vibrator, called an electrolarynx, is placed against the neck and excitation signals are mechanically generated from outside. Although using the noninvasive electrolarynx is an efficient method to produce speech while patients' oral cavity and articulatory abilities are preserved, the resulting EL speech is typically noisy and unnatural. On the one hand, in order for EL speech to be heard easily, excitation signals must have high intensities. Generating intensified buzzy excitation signals results in intelligibility degradation of the EL speech, because these signals are reflected back and leak outside. On the other hand, by using an electrolarynx, it is not possible to generate natural F0 patterns corresponding to linguistic contents. Hence, EL speech sounds unnatural and has a robotic quality.

To produce natural-sounding EL speech, it is required to predict and control the underlying F_0 contours of the EL speech. Traditionally, statistical voice conversion (VC) [2, 3] has been applied to this prediction task. This technique aims to predict natural F_0 contours based on the statistics extracted from a parallel dataset consisting of utterance pairs of EL speech and normal speech. In [4], Nakamura *et al.* have proposed an speaking-aid system using VC, in which segmental feature vectors of spectra of the EL speech were used to predict natural F_0 contours. They have furthermore introduced an EL-air speech system in which F_0 contours can be controlled by using an air-pressure sensor. In [5], the authors have developed a real-time statistical F_0 contour prediction system for vibration control of the electrolarynx. This system, in turn, uses segmental spectral features to predict F_0 contours, and moreover, predicts forthcoming F_0 values to control F_0 patterns of the excitation signals. In a recent work, Kobayashi *et al.* [6]

10.21437/SSW.2019-45

have used a system based on deep neural networks (DNNs) to map segmental features into target F_0 contours. Even though these systems have improved perceived naturalness of the EL speech, the predicted F_0 patterns still deviate from the target ones and they are not able to present the prosodic system of the language.

Recent advances in Text-to-Speech (TTS) [7, 8] and voice conversion [9] systems, have made it possible to generate natural F_0 contours from phoneme-related features and synthesize speech with human quality. Inspired by these systems, in this work, we investigate F_0 prediction accuracy based on phoneme labels to examine whether or not these labels could be considered as shared features between normal and EL speeches. We hypothesize that the application of EL speech spectral features cannot result in high prediction accuracy when a small amount of parallel training data is available. Since EL speech spectral features are significantly different from spectral features of the normal speech, it is not straightforward to effectively use easily available normal speech datasets for training of the model for the EL speech. Consequently, the quality of the models could be low due to the lack of sufficient training data. By using shared input features, we would be to train the prediction model using a large amount of training data. As such features, in this work, we examine F_0 prediction accuracy based on phoneme-related features. The findings show that by considering phoneme labels for both vowels and consonants and one-hot encoding of these labels, we are able to predict F_0 contours with high correlation coefficients. Furthermore, even if we only consider the occurrence times for possible phoneme combinations in an utterance, comparable prediction accuracy as in the case based on segmental spectral features can be obtained.

2. Related works

In the literature of EL speech enhancement, statistical F_0 prediction based on Gaussian mixture models (GMMs) [3, 4, 5], and F_0 prediction using neural networks [6] have been proposed for enhancing naturalness of the EL speech. In statistical F_0 prediction, a parallel dataset consisting of utterance pairs of EL speech and normal speech is developed in advance and a two-step training-prediction process is performed to predict F_0 contours from segmental spectral features. In the training step, the joint probability density function for acoustic features of the EL speech and normal speech is modeled with a GMM. This GMM is then trained based on the expectation-maximization (EM) algorithm to optimize the model parameters. In the prediction step, segmental spectral features of the EL speech are mapped into the most likely F_0 sequence of the normal speech based on the maximum likelihood parameter generation (MLPG) technique.

GMM-based F_0 prediction can result in more natural F_0 patterns. However, due to modeling and conversion errors and also inherit characteristics of the EL speech spectral features, intelligibility degradations can be easily perceived in the synthesizing voices, especially for the tonal languages such as Japanese and Mandarin. Therefore, to further increase the complexity of the prediction model, F_0 pattern prediction based on DNNs [6] has been used. This method follows similar principles as in the GMM-based F_0 prediction. However, in the training step, instead of using EM algorithm to optimize GMM parameters, the parameters of the prediction network (weights and biases) are optimized using back-propagation through time (BPTT) with any optimization technique such as stochastic gradient descent (SGD).

DNNs can be considered as universal function approximators with the capability to learn the underlying mapping between input features and desired output feature in a supervised format. Therefore, by using deep models, the network is able to learn higher-level features that could be beneficial for understanding and modeling the relationships between acoustic features of the EL speech and normal speech. However, the prerequisite of an accurate F_0 prediction based on DNNs is the availability of a large amount of training data. Because the the existing EL speech datasets contain very limited number of utterances, it is hard to train these models, and therefore, the network may fail to learn an accurate mapping between segmental spectral features of the EL speech and F_0 contours of the normal speech.

3. F_0 Prediction Based on One-Hot Encoding of Phoneme Labels

3.1. Indexed speech as basis for F_0 prediction

Although DNNs have shown their potentials in learning highly non-linear and complicated mappings from the space of input features into the underlying space of the output features, we still observe that, as long as predicting F_0 contours for the EL speech is concerned, fairly limited improvements can be achieved. This stems mainly from two facts. One the one hand, EL speech spectral features are significantly different from those of the normal speech, though they are varying according to phonemes. Since electrolarynx always generates excitation signals of constant fundamental frequency independent of speech content, the spectrogram of EL speech does not contain any relevant information about F_0 variations for voiced consonants and vowels. Hence, it is not straightforward to predict accurate F_0 contours from EL speech spectral features. On the other hand, creating datasets with a large amount of utterance pairs of EL speech and normal speech is very costly and time-consuming. Therefore, the quality of the prediction models could be still low due to the lack of sufficient training data.

To tackle these issues, it is necessary to look for some other informative input features. These features should carry useful information about voiced consonants and vowels, and also they should be shared between EL speech and normal speech. By having shared features, we would be able to train F_0 prediction networks for the EL speech using easily and publicly available datasets for normal speech.

Recent advances in TTS [7, 8] and VC [9] systems, have made it possible to generate natural F_0 contours from phoneme-related features and synthesize speech with human quality. The authors in [9] have shown that by utilizing phonetic posteriorgrams (PPGs), it is possible to bridge between speakers and train a deep recurrent neural network (DRNN) that successfully converts PPG of the source speaker into acoustic features of the target speaker using non-parallel datasets. Inspired by this work, in this study, we investigate F_0 pattern prediction for the EL speech based on one-hot encoding of the phoneme labels. PPG is defined as a time-versus-class matrix representing the posterior probabilities of each phonetic class for each specific time frame. Using the same analogy, we define a time-versus-

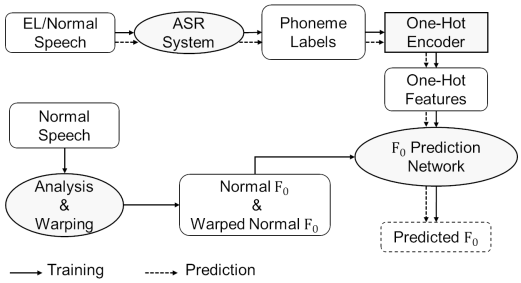


Figure 1: *Block diagram of a system in which F_0 contours are predicted based on phoneme sequence. ASR system in this structure is a frame-by-frame phoneme recognizer.*

phoneme-label matrix and use this as input for the F_0 prediction network. In this matrix, phoneme labels are one-hot encoded, so that we are confident about labels for individual time frames. By considering phoneme sequence as input features, the relationships between adjacent phonemes (temporal information) and phoneme combinations can be utilized to find a mapping that maps phoneme-related features into target F_0 contours.

3.2. Network structure and F_0 prediction procedure

As illustrated in Figure 1, F_0 pattern prediction based on one-hot encoding of the phoneme labels is performed in two steps. In the training step, extracted phoneme labels are one-hot encoded and fed into a network that is trained to learn the mappings between time-versus-phoneme-label matrices and target F_0 contours. In the prediction step, the final model is utilized to predict a sequence of F_0 values for the utterances in the evaluation set. In this system, phoneme labels are extracted frame-by-frame from spectrogram of the utterances using a phoneme recognizer such as DNN-based phoneme posteriorgram estimator. Also, in order to model temporal dynamics of the features within the adjacent frames, recurrent networks (e. g., LSTM [10] or BiLSTM [11]) are used.

3.3. Different scenarios for predicting F_0 contours based on phoneme labels

In linguistics, phonemes are considered as the atoms of speech. They are the smallest units of spoken sounds capable of distinguishing one word from others and conveying distinct meanings. For predicting F_0 contours based on phoneme labels, it is necessary to investigate which sequence of phonemes can contribute more to prediction accuracy. Since phonemes are divided into vowels and consonants, we can define different scenarios based on this labeling. Furthermore, we need to investigate whether or not the set of all possible phonemes in a language is required for the prediction task. If we could reduce the labels in this set, we would be able to use a frame-by-frame phoneme recognizer with simpler structure. These investigations are language-dependent. In the following, possible scenarios for one-hot encoding of the phoneme labels in Japanese are presented.

(i) **Based on the set of all phoneme labels:** Here, we determine how many unique phoneme labels exist in our dataset of indexed speeches. Then, for every utterance, phoneme durations are calculated to figure out how many frames are grouped under the same phoneme label. For these frames, phoneme label is one-hot encoded. Having done one-hot encoding of the phoneme labels over all frames, these features are fed into the F_0 prediction network.

(ii) **Based on the set of vowel labels:** It is believed than vowels are playing an important role when F_0 prediction is concerned. Hence, here, for every indexed speech, we keep all labels representing unique vowels, and substitute those for consonants with their succeeding vowel. By doing so, we can convert phonemes from being vowels and consonants into vowels only. Moreover, we can reduce the number of labels in the set of all phoneme labels which may help us to use a simpler phoneme recognizer. Finally, vowel durations are calculated and for frames having the same phoneme label, one-hot encoding of the phoneme label is done.

(iii) **Based on the occurrence times for phoneme combinations:** Here, one-hot encoding is done differently. Instead of considering the existing phonemes and their respective durations, in this case, we only focus on exact time instances at which a phoneme combination starts and ends. Since in this study we are investigating Japanese language, a phoneme combination is either a single vowel, or a consonant followed by a vowel. Every individual phoneme combination is then considered as a tuple given by (*start time, content, end time*). In this tuple, *content* is an alias for the name of the combination we no longer care about. Pairs of *start/end* time instances are calculated for all of the existing combinations in the indexed speech, and the final one-hot encoding is done based on an extremely reduced set of labels.

These scenarios have been summarized in Table 1.

Table 1: *Scenarios defined for converting phoneme labels into one-hot features.*

Used labels	Example
1) All phoneme labels	a r a y u r u g e N j i ts u o sp
2) All vowel labels (consonants are substituted by their succeeding vowel.)	a a a u u u u e e i i i u u o sp
3) Vowel (v) or consonant-vowel (cv) Here, only occurrence times are considered.	a ra yu ru ge N ji tsu o sp v cv cv cv cv v cv cv v sil

3.4. Target F_0 preparation

For supervised training of the prediction network, ground truth data or target F_0 contours must be prepared in advance. To do so, time warping is the common technique used to time align input features with desired target features. However, EL speech has many short pauses (SPs) which may either not exist in normal speech, or occur at different positions. (See Figure 2). To diminish these mismatches, we use a warping process that is constrained on phoneme labels. In other words, warping is done phoneme label by phoneme label. We start off by the first pair formed from the first phoneme labels from EL speech and normal speech. If they are similar, then warping is done based on spectral features according to the process described in [12] to minimize mel-cepstral distortion. If they are different, then we know this could have happened because of an SP occurrence in EL speech. In such cases, we zero pad target features, right at

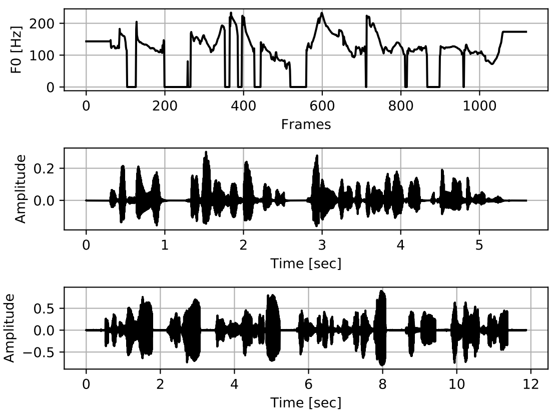


Figure 2: *Mismatch in short pauses between EL speech and normal speech. From top to bottom: Extracted F_0 contour for normal speech, waveform for normal speech, and the corresponding waveform for EL speech.*

the corresponding position to SP index, so that we are able to make current phoneme labels similar again. Every time, once warping for the current pair is done, we stack warping paths to gradually form our final warping functions. This process is repeated until the last pair of phoneme labels is warped. Final warping functions are then applied to target F_0 contours to extend their time span. However, due to zero-padding, in the warped F_0 contours several flat patterns will be generated that make these contours invalid as natural F_0 contours. To resolve this issue and furthermore to change these discontinuous contours into continuous ones, spline interpolation is utilized. Finally, continuous target F_0 contours are low-pass filtered to filter out rapid ripples, known as microprosody [13]. Preparation of target F_0 contours has been illustrated in Figure 3.

4. Experimental Evaluation

4.1. Experimental conditions

Dataset and feature extractor: The ATR speech dataset [14] comprising of 503 Japanese sentences uttered with and without an electrolarynx by a Japanese male speaker was used in our experiments. The utterances in this dataset have been grouped in 10 sets each with 50 utterances, except for the 10th set that contains 53 utterances. Forced-aligned phoneme labels and required acoustic features were extracted using the open-source Julius speech recognition system [15] and the STRAIGHT vocoder [16], respectively. The first 25 mel-cepstral coefficients extracted for both speech types were used as spectral features for time warping.

Network architecture: Two stacked bi-directional Long Short-Term Memory (BiLSTM) layers followed by a single time-distributed dense layer formed the architecture of our F_0 prediction network. For recurrent layers, the hyperbolic tangent ($\tanh$) activation function was used, and the number of hidden units was set to 128. In the last layer, the linear activation function was utilized and by defining loss function as the root mean square error (RMSE) between predicted F_0 contours and target ones, network parameters were optimized using the Adam optimizer [17] for utterance batches of size 32. The learning rate α, β_1 and β_2 were set to 0.0004, 0.9 and 0.999, respectively.

Experiments: For every speech type, predicting F_0 contours based on conventional spectral features for the existing utterances in set A of the ATR dataset was considered as the baseline method. We then conducted tow different sets of experiments to investigate: 1. how predefined scenarios for one-hot encoding of the phoneme labels would affect the accuracy of the predicted F_0 contours, and 2. whether or not these features could be shared between EL speech and normal speech (i. e., can we use easily available datasets for normal speech to increase the F_0 prediction accuracy for the EL speech). Experiments addressing the first goal of our investigations are denoted as G1, and those related to the second goal as G2. For G1 experiments, same utterances as for the baseline method were used, namely 50 utterances in set A. For G2 experiments, we had to form training sets with utterances of both speech types, but with different contents (data augmentation). To achieve this, 32 EL utterances from set A were always fixed as training set and additional 32, 64, 128 and 256 normal utterances from different sets, other than set A for normal speech, were augmented to this training set. We further performed G2 experiments for only normal speech where instead of fixing 32 EL utterances from set A as constant members of the training sets, 32 normal utterances from set A were substituted. Investigating the impact of having no mismatches in short pauses was the main motivation for performing G2 experiments for only normal speech. In all experiments, target F_0 contours were standardized to zero-mean and unit variance using the statistics of the training sets, and 4-fold cross validation test was conducted for the evaluation set (i. e., 10 EL utterances from set A) to report the final results.

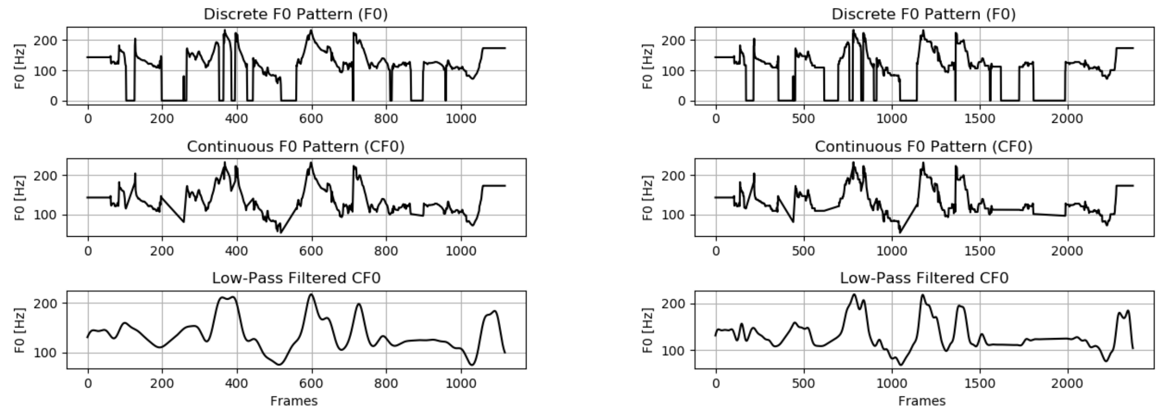
4.2. Experimental evaluations

Predicted F_0 contours for the evaluation set were objectively evaluated for only voiced frames using Pearson's product-moment correlation measure r given by [18]:

$$r_{xy} = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}\sqrt{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}}, \quad (1)$$

where x_i and y_i are the individual F_0 values from the predicted and target F_0 contours, respectively. Further, $\bar{x}$ and $\bar{y}$ are the mean values and n is the length of F_0 contours for only voiced frames.

Figure 4 gives a comparison between average r values for different types of input features when only 32 utterances were used for training (G1 experiments without any data augmentation). It is evident that by using spectral features, accurate F_0 contours with very small standard errors could be predicted for normal speech. However, the prediction accuracy drops when EL speech spectral features were used. When vocal folds are vibrating, for instance at the generation time of voiced consonant or vowels, particular F_0 values are observed in the spectrogram of the normal speech. This useful information can help the network to learn the underlying patterns required for an accurate F_0 pattern prediction. However, F_0 values of the EL speech are mechanically generated independent of utterance content. Hence, the network performance is relatively poor when trained on a small amount of training data. If we consider one-hot features, we can see that for all scenarios comparable or higher r



(a) *Prepared for the normal speech*

(b) *Prepared for the EL speech*

Figure 3: *Preparation of target F_0 contours. Time warping was applied to the extracted F_0 contours for normal speech.*

values have been achieved. This indicates that one-hot encoding of the phoneme labels can indeed be used for F_0 prediction, even if a reduced set of phoneme labels is considered. This is beneficial, because the prediction is made not based on frequency patterns embedded in the spectral features, but based on unique codes that represent phoneme labels.

The impact of increasing the number of training utterances on the prediction accuracy (G2 experiments) has been presented in Figure 5. Considering the obtained results for the normal speech, we can see that by increasing the number of utterances in the training sets, the network prediction capability has been improved and higher correlation coefficients have been obtained. This indeed was expected, since providing a network with more data has a direct influence on its performance.

Furthermore, it is evident that through one-hot encoding of all phoneme labels, it is possible to obtain high average r values. This is because when we consider all labels, we can well encode possible combinations between vowels and consonants in an utterance. It is known that vowels have a distinct steady formant patterns when occurred in isolation. These patterns, however, are altered by the adjacent consonant which is known as formant transition, and have important information about the place and manner of articulation of the following or the preceding consonant. These important information are embedded in the spectral features and that is the reason why F_0 prediction based on spectral features of the normal speech results in high r values. Using the same analogy, if we one-hot encode all of the phoneme labels, we enforce the network to learn the possible vowel-consonant combinations, and hence we are able to achieve higher correlation coefficients.

Considering the augmentation of the training sets with additional normal utterances for the EL speech experiments, we can see that the obtained average r values for the case of using all phoneme labels are higher than those of the other two cases. However, they are not as high as the r values calculated for the experiments in which only normal speech was used. One possible reason for this difference could be the way we augmented training sets with additional utterances. We used 32 EL utterances of set A, and the additional utterances were selected from normal utterances of other sets. This might not have provided sufficient training patterns for the F_0 prediction network,

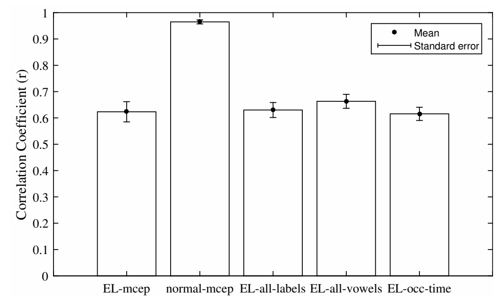


Figure 4: *Comparison of average r values obtained for G1 experiments (when only 32 utterances of set A were used for training of the prediction network).*

mainly due to existing mismatches in the count and position of the short pauses between EL speech and normal speech. Samples of the predicted F_0 contours can be found in Figure 6.

Lastly, it is worth exploring the impact of network architecture on the obtained prediction accuracies. In sequential prediction tasks, where samples of the input sequence for all time steps are available, we may prefer to use bi-directional recurrent networks. By using bi-directional recurrent networks, we benefit from context information in both forward and backward directions provided by the input features and their reverse copy. However, providing a system with the reversed copy of its input features violates the causality property stating that, for any time step t, outputs of the system should not depend on future inputs. Consequently, in order to realize a real-time prediction system, uni-directional recurrent networks should be used. It is also worth mentioning that in our experiments, phoneme labels were extracted based on forced alignment. That is, for any utterance in the dataset, the corresponding transcript was also available. When speech is produced in real-time by laryngectomees, no transcript can be considered. To address this issue, speech signal must be delayed for some frames corresponding to a specific time in [msec], and then phoneme labels must be extracted in a frame-by-frame manner using a phoneme recognizer. By doing so, we will be able to extract a fragment of the underlying transcript.

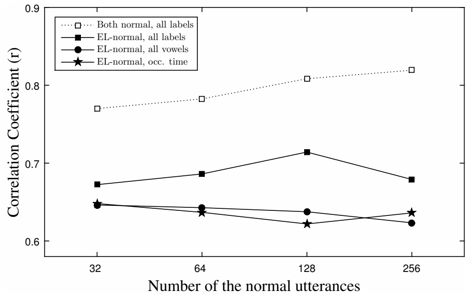


Figure 5: *Impact of augmenting training sets with various number of normal utterances on the F_0 prediction accuracies (G2 experiments).*

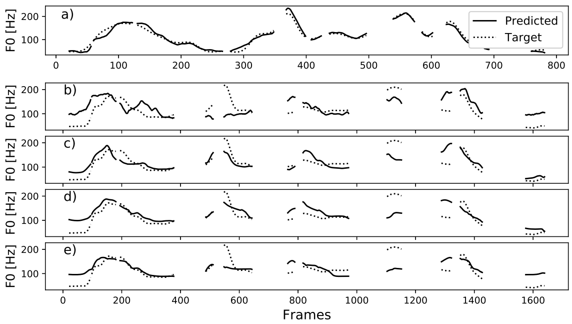


Figure 6: *Samples of the predicted F_0 contours for a) normal speech using spectral features, b) EL speech using spectral features, c) $\sim$ e) EL speech using one-hot encoding for all phoneme labels, all vowel labels and occurrence times for phoneme combinations, respectively. For one-hot cases, the training sets were augmented with 128 normal utterances (G2 experiments).*

5. Conclusion

Enhancing naturalness of the EL speech was addressed in this work. To circumvent the lack of sufficient EL speech data for training models that map EL speech spectral features into natural F_0 contours, we investigated F_0 pattern prediction based on other features that can be shared between EL speech and normal speech. As such features, we considered various scenarios for one-hot encoding of the phoneme labels. These features were generated both for EL speech and normal speech, and used in the training of a recurrent network that was designed to learn the mapping between phoneme labels and target F_0 contours. The findings revealed that by one-hot encoding of both vowels and consonants labels, we are able to achieve F_0 contours with higher correlation coefficients. Furthermore, by using a reduced set of the phoneme labels, we are still able to predict F_0 contours with comparable accuracies to those obtained based on the spectral features.

6. Acknowledgement

This work was supported in part by JST, PRESTO Grant Number JPMJPR1657, and JSPS KAKENHI Grant Number 17H01763.

7. References

[1] R. Kaye, C. G. Tang and C. F. Sinclair, "The electrolarynx: voice restoration after total laryngectomy," in *Medical devices*, vol. 10, 2017, pp. 133–140.

[2] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, March 1998.

[3] T. Toda, A. W. Black, and K. Tokuda, "Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory," *IEEE Trans. on ASLP*, vol. 15, no. 8, pp. 2222–2235, 2007.

[4] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech," *Speech Communication*, vol. 54, no. 1, pp. 134–146, 2012.

[5] K. Tanaka, T. Toda, G. Neubig, and S. Nakamura, "Real-time vibration control of an electrolarynx based on statistical F0 contour prediction," *2016 24th European Signal Processing Conference (EUSIPCO)*, pp. 1333–1337, Aug 2016.

[6] K. Kobayashi and T. Toda, "Electrolaryngeal Speech Enhancement with Statistical Voice Conversion based on CLDNN," *2018 26th European Signal Processing Conference (EUSIPCO)*, pp. 2115–2119, 2018.

[7] 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. A. Saurous, "Tacotron: Towards End-to-End Speech Synthesis," pp. 4006–4010, 08 2017.

[8] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions," pp. 4779–4783, 04 2018.

[9] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," pp. 1–6, 07 2016.

[10] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997. [Online]. Available: http://dx.doi.org/10.1162/neco.1997.9.8.1735

[11] M. Schuster and K. K. Paliwal, "Bidirectional Recurrent Neural Networks," *IEEE Trans. Signal Processing*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997. [Online]. Available: http://dx.doi.org/10.1109/78.650093

[12] T. Toda, M. Nakagiri, and K. Shikano, "Statistical Voice Conversion Techniques for Body-Conducted Unvoiced Speech Enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 9, pp. 2505–2517, Nov 2012.

[13] A. Sakurai and K. Hirose, "Detection of phrase boundaries in japanese by low-pass filtering of fundamental frequency contours," *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, vol. 2, pp. 817–820, Oct. 1996.

[14] 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 Communication*, vol. 9, pp. 357–363, 08 1990.

[15] A. Lee, T. Kawahara, and K. Shikano, "Julius—an open source real-time large vocabulary recognition engine," *Proceedings of European Conference on Speech Communication and Technology*, vol. 3, pp. 1691–1694, 01 2001.

[16] 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*, 2001.

[17] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 12 2014.

[18] D. Hermes, "Measuring the perceptual similarity of pitch contours," *Journal of speech, language, and hearing research: JSLHR*, vol. 41, pp. 73–82, 03 1998.