Phoneme Embeddings on Predicting Fundamental Frequency Pattern for Electrolaryngeal Speech

Mohammad Eshghi*, Kazuhiro Kobayashi†, Kou Tanaka‡, Hirokazu Kameoka‡ and Tomoki Toda†
* Graduate School of Information Science, Nagoya University, Nagoya, Japan
E-mail: mohammad.eshghi@g.sp.m.is.nagoya-u.ac.jp
† Information Technology Center, Nagoya University, Nagoya, Japan
E-mail: kobayashi.kazuhiro@g.sp.m.is.nagoya-u.ac.jp, tomoki@icts.nagoya-u.ac.jp
‡ NTT Communication Science Laboratories, NTT Corporation, Japan
E-mail: {tanaka.ko, kameoka.hirokazu}@lab.ntt.co.jp

***Abstract*—Electrolaryngeal (EL) speech has robotic quality owing to constant fundamental frequency (F_0) patterns. In existing F_0 pattern prediction frameworks, acoustic models are trained on spectral features of a large corpus of healthy speech. However, EL speech does not embed any useful information about F_0 into spectrogram. Moreover, creating datasets with reasonably large number of EL utterances for training neural networks is very time-consuming. Hence, F_0 prediction based on other features with sharing capability between EL and normal speech must be investigated. In this study, we investigate F_0 prediction based on clustering of the phoneme embeddings. For a dataset consisting of utterances of both speech types, phoneme labels are extracted. These phoneme labels are then used to learn phoneme embeddings in a common 2-D space. Through clustering of the learned phoneme embeddings, new onehot features are created for F_0 prediction. Experimental results show that when considering training sets consisting mixed utterances of EL and normal speech, by using new features, improvements in F_0 prediction accuracy can be achieved. Moreover, accurate F_0 patterns can be predicted even based on lower-dimensional features corresponding to small values for the number of clusters. This could simplify the structure of the recognition system required to extract phoneme labels from EL speech.**

I. Introduction

Removal of the vocal folds from an anatomically functional larynx is called total laryngectomy (TL). This surgical procedure is usually performed when patients are diagnosed with larynx cancer. Given that the phonetic system of most languages are notably consisted of voiced consonants and vowels, the absence of vocal folds vibrations leads to marked voice abnormalities and speech with decreased intelligibility.

Over the past decades, many voice restoration methods have been proposed to fill the gap of vibrating apparatus and reproduce speech with enhanced intelligibility and naturalness. In [1], three main methods of voice restoration have been addressed as electrolaryngeal (EL) speech, esophageal (ES) speech, and tracheoesophageal speech through a tracheoesophageal puncture (TEP), with TEP speech as the current gold standard. In [2], an alternative method has been introduced in which nonaudio bio-signals can be directly converted to speech. Amongst these different available methods, EL speech has been considered as the most widely used method by laryngectomees. In this method, a battery operated vibrator, called an electrolarynx, is placed against the neck and single-tone excitation signals are mechanically generated from outside. These tone signals are articulated by the tongue, lips and teeth, and form a relatively intelligible speech.

Even though patients' oral cavity and articulatory abilities are preserved while producing EL speech, the produced F_0 patterns of EL speech do not resemble natural patterns. In fact, monotonicity of the excitation signals results in constant F_0 patterns free from any paralinguistic information such as intonation. Traditionally, statistical F_0 modeling and prediction [3], [4], [5], [6] and, recently, F_0 prediction based on deep neural networks (DNNs) [7], [8], [9] have been used to predict natural F_0 patterns for EL speech. Though by using these techniques the overall perceived naturalness has been improved, the predicted F_0 patterns are still suffering from fairly limited accuracies and linguistically incorrect intonations. In this regard, two particular reasons can be highlighted: (i) extracting valuable information from EL speech spectral features for predicting F_0 patterns is relatively hard; and (ii) since creating EL speech datasets with large number of training utterances is very costly, modeling of F_0 prediction with small datasets could be inaccurate. Therefore, in order to achieve F_0 patterns with satisfactory accuracies, we need to introduce some sort of shared features between normal and EL speech. By having shared features, we would be able to benefit from easily available normal speech datasets in favor of developing a system for precise F_0 pattern prediction for EL speech.

Recent advancements in text-to-speech (TTS) [10], [11], [12] and nonparallel voice conversion systems based on phonetic posteriorgrams (PPGs) [13] have shown the feasibility of producing human-like speech using text-related features. Motivated by these works, in our previous study [14], we investigated F_0 pattern prediction based on phoneme labels. Considering a full set of phoneme labels, we found that onehot encoded phoneme labels could be used to predict F_0 patterns with relatively high correlation coefficients. More interestingly, we observed that even a reduced set of phoneme labels (*e.g.* when only using the vowels set) could result in comparable prediction accuracies to those one can achieve when using EL

speech spectral features. Being able to reduce the dimensionality of phoneme set could be very beneficial for practical applications because it simplifies the structure of the required recognition system. In other words, for precisely extracting the set of all phoneme labels, a highly accurate recognition system that has sensitivity to all vowel and consonant labels must be developed, whereas for a subset of these labels we no longer need to use a complex recognition system.

In this work, we aim to explicitly investigate how we can extract lower-dimensional input features, yet based on phoneme labels, for predicting natural F_0 patterns. To this end, we consider learning phoneme embedding in a continuous vector space from nominal phoneme labels. For a dataset consisting of mixed utterances of normal and EL speech, forced-aligned phoneme labels are extracted frame-by-frame in advance. By considering the dictionary of all phoneme labels, these labels are onehot encoded and used as input linguistic features. We then use an embedding network on the front end of F_0 prediction network and train the resulting structure so that it learns how to map input features into desired target F_0 patterns through transition from the embedding network. Upon training, the embedding network is used to extract phoneme embeddings for existing utterances in our dataset. These phoneme embeddings are then clustered into predefined number of classes. Depending on how many classes have been defined, the clustered embeddings are onehot encoded to generate new categorical input features for F_0 prediction. Ultimately, the embedding network is detached and distinct prediction networks are trained from scratch.

The advantages of using this structure are twofold: firstly, by clustering phoneme embeddings in a common 2-D space, we can define a unique onehot encoding scheme for phoneme embeddings of both speech types; secondly, by comparing prediction accuracies versus the number of classes, a suitable value as the dimension for the reduced set of phoneme labels can be determined.

II. Related Works

Commercial electrolarynx devices lack a unit for pitch controlling and voice onset and offset timing. Considering the source-filter model of speech production, after total laryngectomy, the excitation source is removed, while the oral cavity and the articulatory abilities are preserved. Hence, dedicated research on EL speech naturalness enhancement focuses on how to compensate this deficit on excitation source in order to add pitch prediction and controllability to EL speech.

A. F_0 prediction based on spectral features

Statistical F_0 prediction and modeling based on Gaussian mixture models (GMMs) [4], [5], [6], [15], and F_0 prediction based on DNNs [7], [8] are two common approaches on EL speech naturalness enhancement using spectral features. In statistical F0 prediction, a parallel dataset consisting of utterance pairs of EL and normal speech is developed in advance and a two-step training-prediction procedure is performed to predict natural F0 patterns from segmental spectral features. In the training step, the joint probability density function for acoustic features of EL and normal speech is modeled with a GMM. In the prediction step, segmental spectral features of EL speech are mapped into the most likely natural F_0 pattern based on the maximum likelihood parameter generation (MLPG) technique.

Constructed upon similar principles, DNNs can be considered as powerful tools for F_0 prediction due to their distinct capability to learn higher-level features from provided input features. With feature learning, more subtle links between predictor and response variables can be learned by the network. However, the performance of neural network is tightly bound to the available amount of training data. The more training data we provide the network with, the more accurate F_0 patterns we can expect to be predicted by the network. Unfortunately, creating EL speech dataset is considered as time-consuming and expensive. Therefore, there are only a handful of freely available EL speech datasets with limited number of utterances. This turns the F_0 prediction into a challenging problem and could result in learning less accurate mappings between EL speech spectral features and target F_0 patterns by the prediction networks.

B. F_0 prediction based on nominal phoneme labels

Rather than predicting F_0 patterns by DNNs based on conventional features, prediction by considering phoneme labels has been investigated in [14]. Here, phoneme labels for individual speech frames are extracted and onehot encoded as input features. Even though spectral features of EL and normal speech are completely different, the set of phoneme labels used to generate various utterances could be similar between them. Hence, less discrepant input features for F_0 prediction could be made out of phoneme labels. However, in order to accurately extract the set of all phoneme labels, a high-quality recognition system must be developed that could be costly for generic applications. This could be considered as a drawback for predicting F_0 patterns based on the full set of phoneme labels.

III. Phoneme Embeddings on Creating Features for F_0 Pattern Prediction

A. Phoneme embeddings

Allocating onehot codes to nominal phoneme labels has three main downsides:

1) Onehot codes are unlearned representations obtained without any supervision. Hence, the relationships between similar phonemes classes are completely ignored.
2) For onehot codes no distance metrics can be defined. Hence, it is impossible to measure similarity between adjacent phoneme classes.
3) Naive allocation of onehot codes to all phoneme labels does not provide us with any information on how to reduce the dimensionality of phoneme set. Having lower-dimensional phoneme sets could decrease the costs on developing recognition systems for extracting phoneme labels.

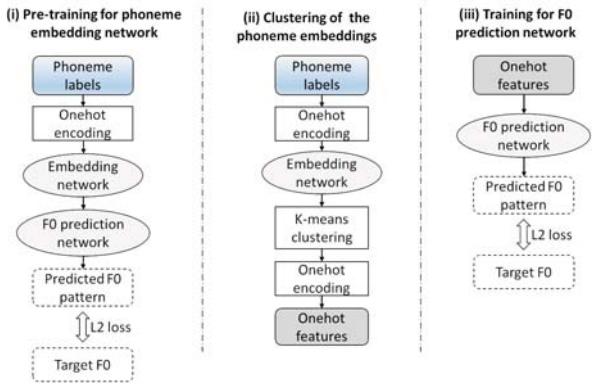

Fig. 1. Learning and clustering of phoneme embeddings for F_0 pattern prediction.

To address these issues, it is essential to convert onehot codes into real-valued vectors. By doing so, we would be able to project nominal phoneme labels into a continuous vector space, for which various distance metrics could be defined. Thus, we would be able to cluster data points with similar characteristics and compress down the dimensionality of phoneme sets.

B. Learning phoneme embeddings and F_0 pattern prediction

Fig. 1 illustrates the block diagram of the system in which F_0 patterns are predicted based on clustering of phoneme embeddings. This system comprises of three blocks. These are: (i) learning phoneme embeddings, (ii) clustering of the phoneme embeddings, and (iii) F_0 pattern prediction based on the clustered phoneme embeddings.

Initially, for a dataset consisting of mixed utterances of normal and EL speech, forced-aligned phoneme labels are extracted frame-by-frame. By considering the set of all phoneme labels, these labels are onehot encoded and used as input linguistic features. A recurrent neural network with an embedding network on the front end is then trained to map onehot encoded phoneme labels into natural F_0 patterns. Once training is done, the trained embedding network is used to transform phoneme labels into real-valued embedding vectors in a frame-by-frame manner. This is done for the entire utterances in the datasets, both for normal and EL speech. Next, considering both speech types, the obtained phoneme embeddings over all frames for individual utterances are concatenated together. At this stage, a k-means classifier is trained to cluster phoneme embeddings into specified number of classes. Upon clustering, embeddings in individual clusters are converted into onehot codes considering the overall number of clusters in the space. Lastly, the embedding network is excluded from the network architecture and the remaining prediction network is trained from scratch for new onehot features.

IV. Experimental Evaluation

A. Experimental conditions

Dataset: The ATR speech dataset [16] was used for our experiments. This dataset consists of 503 Japanese sentences uttered with and without an electrolarynx by a Japanese male speaker. The utterances in this dataset have been grouped in 10 sets each with 50 utterances, except for the 10^{th} set that contains 53 utterances. Forced-aligned phoneme labels and required acoustic features were extracted using the open-source Julius speech recognition system [17] and the STRAIGHT vocoder [18], respectively. Low-pass filtered continuous F_0 contours of normal speech were used as target F_0 patterns. These contours were standardized to zero mean and unit variance using the statistics of training sets.

Network architecture: Two stacks of bi-directional long short-term memory (BiLSTM [19], [20]) layers followed by a single 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 64. For the dense layer, a linear activation function was utilized, and by defining the loss function as the root mean square error (RMSE) between predicted and target F_0 patterns, the network parameters were optimized using Adam optimizer [21] 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. As embedding network, an embedding layer on the front-end, right before the recurrent layers, with output dimension of 2 ($odim = 2$) was utilized.

Experiments: Accuracy of F_0 prediction with respect to training data was investigated in four scenarios. In the first scenario, for every speech type, distinct prediction networks were trained on 32 utterances selected from the corresponding set $\mathcal{A}$ to that speech type (baseline system). In the second scenario, a training set consisting of 64 utterances was made by unifying the two sets used in the first scenario. This represented the case where we had parallel dataset for training. In the third scenario, the usage of EL speech in training was ignored and only normal utterances were considered. Our goal was to specify the upper bounds on prediction accuracies for increasing number of training utterances. Created training sets for this scenario have been summarized in Table I. In the fourth scenario, the training sets in the third scenario were further augmented with 32 EL utterances from set $\mathcal{A}$. These represented training sets comprising of both speech types with fixed small portion of EL speech, but varying large portions of normal speech.

TABLE I
Training datasets used in the third scenario to specify upper bounds on prediction accuracies.

Speech type	Used sets	#Utterances
Normal	$\mathcal{A}$	32
	$\mathcal{A}, \mathcal{BCD}$	32 + 150 = 182
	$\mathcal{A}, \mathcal{BCDEFG}$	32 + 300 = 332
	$\mathcal{A}, \mathcal{BCDEFGHIJ}$	32 + 453 = 485

In all experiments, once the embedding network was trained, phoneme embeddings were extracted and clustered into 8, 15, 23, 30 and 38 clusters, with 38 corresponding to the total number of phoneme labels for our dataset. Except for the scenarios in which the network had to be trained on only normal utterances, 8 EL utterances from set $\mathcal{A}$ were always fixed as validation set for parameter tuning and best model selection. Finally, 4-fold cross validation test on 40 EL speech utterances from set $\mathcal{A}$ was conducted to evaluate the prediction accuracies.

B. Experimental evaluations

Predicted F_0 patterns were objectively evaluated for only voiced frames using the Pearson's product-moment correlation coefficient r given by [22]:

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

where x_i and y_i are the i^{th} values in predicted and target F_0 patterns, respectively. $\bar{x}$ and $\bar{y}$ are the respective mean values, and n is the total length when only voiced frames are considered.

Obtained correlation coefficients for the first and second scenarios have been presented in Fig. 2. This figure shows that F_0 patterns predicted for normal speech are more accurate that those predicted for EL speech. This could be explained by the fact that EL speech utterances are generally much longer than normal speech utterances. Unfortunately, it is hard for recurrent networks to keep track of long-term dependencies when input sequences are too long. Hence, the network misses some structural information on input features which could result in decreased accuracies. Now, if we consider the second scenario, we can observe that by learning phoneme embeddings from both speech types and clustering of these embeddings, the prediction accuracies have been improved. This indicates that learning real-valued phoneme embeddings from both speech types in a common 2-D space with subsequent clustering and onehot encoding has resulted in features that could be used for predicting more accurate F_0 patterns. In other words, the networks has seen more input features with consistent onehot encoding scheme and was able to improve its prediction performance.

The upper bounds on network performance in case of using only normal speech have been depicted in Fig. 3. It is evident that by increasing the number of training samples with homogeneous structures, the network has learned more subtle features and achieved higher correlation coefficients. Furthermore, we can see that for varying number of clusters, the network has successfully predicted accurate F_0 patterns. Since the number of clusters is equivalent to the dimension of input features, this can be interpreted as the possibility of using a reduced set of phoneme labels for F_0 pattern prediction. Hence, instead of developing a complex recognition system for extracting all of the phoneme labels, we could develop a much simpler one that outputs a reduced set of phoneme labels and obtain relatively similar correlation coefficients.

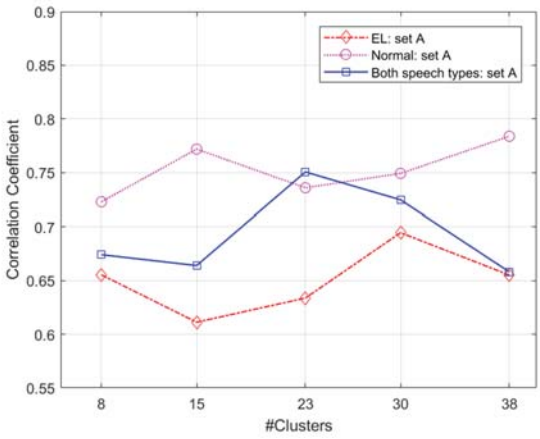

Fig. 2. Obtained correlation coefficients for EL speech in baseline system (dashed red line) vs. those obtained for parallel training set in the second scenario (solid blue line).

Obtained F_0 prediction accuracies for the fourth scenario have been also illustrated in Fig. 3. By considering the respective values, we can see that the overall network performance on predicting F_0 patterns for EL speech has been enhanced. Nonetheless, as opposed to the case for normal speech, two types of irregularities can be observed. Irregularities in terms of obtained accuracies versus the number of clusters, and irregularities in terms of obtained accuracies versus the augmented portion of normal utterances to the training sets. These could have happened due to two main reasons. (i) Discrepancies in the set of phoneme labels between normal and EL speech: in our dataset, EL speech had more phoneme labels than normal speech, and (ii) existing mismatches in the count and position of short pauses between normal and EL speech: EL speech has many short pauses that can strongly affect the shape of target F_0 patterns. To alleviate this issue, it is recommended to record normal utterances in the evaluation set, out of which ground truth F_0 contours are extracted, with as many short pauses as the ones exist in their corresponding EL utterances used for evaluation. Samples from predicted F_0 patterns in third and fourth scenarios for 8, 23 and 38 clusters have been presented in Fig. 4.

Finally, it is worth mentioning why in this work we exclusively considered phoneme labels and tried to learn 2-D embedding vectors for them. There are plenty of other linguistic features that are typically used in the state-of-the-art TTS systems for synthesizing natural and expressive speech waveforms. Phoneme duration, accent type, position of the rise and downfall of the F_0 contours within an accentual phrase, syllable position, etc., are examples of these linguistic features. No matter they are used directly with their original numerical values, or as the ones for which new embeddings should be learned, in order to incorporate these into the vector of input

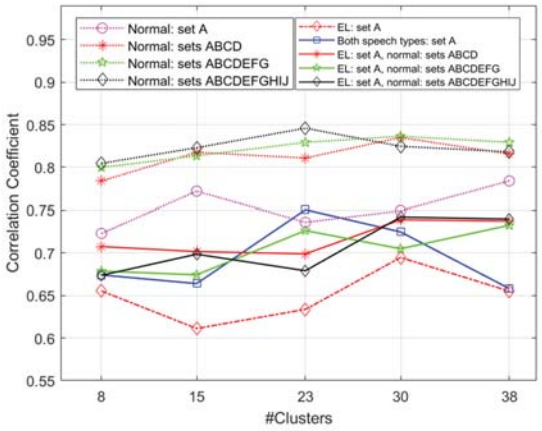

Fig. 3. Comparison between resulted correlation coefficients for normal and EL speech in third and fourth scenarios.

features, significant domain knowledge on how to estimate and process such features is required. Therefore, for the purpose of filtering out the possible estimation errors and reducing the system complexity, only phoneme labels were investigated.

As for output dimension of the embedding layer ($odim$) and hence the dimension of the resulting phoneme embeddings, any value greater than or equal to two could be used. However, higher-dimensional ($odim > 2$) embeddings are often difficult to interpret and visualize. Thus, it is essential to reduce the dimensionality of such features in order to increase interpretability, while at the same time minimizing the information loss. Consequently, in practice, principal component analysis (PCA) or other dimensionality reduction techniques could be applied before k-means clustering to eliminate low-variance dimensions. For the sake of simplicity, $odim = 2$ was used in this work.

V. Conclusions

Predicting natural F_0 patterns for EL speech was addressed in this study. EL speech spectral features do not contain useful information on F_0 or paralinguistic features. Moreover, creating EL speech datasets with large number of training utterances is costly and very time-consuming. These two main factors make it hard to predict accurate F_0 patterns for EL speech based on conventional features used in existing frameworks. In order to find a work around for these issues, we investigated the accuracy of F_0 prediction based on clustering of the phoneme embeddings. For training sets consisting of both EL and normal utterances (with larger portions for normal speech), phoneme embeddings were learned and clustered in a common 2-D space as input features for F_0 prediction. Obtained results revealed that F_0 patterns predicted based on these features could achieve higher correlation coefficients. Moreover, by considering a small number of clusters and creating lower-dimensional features, we were still able to predict

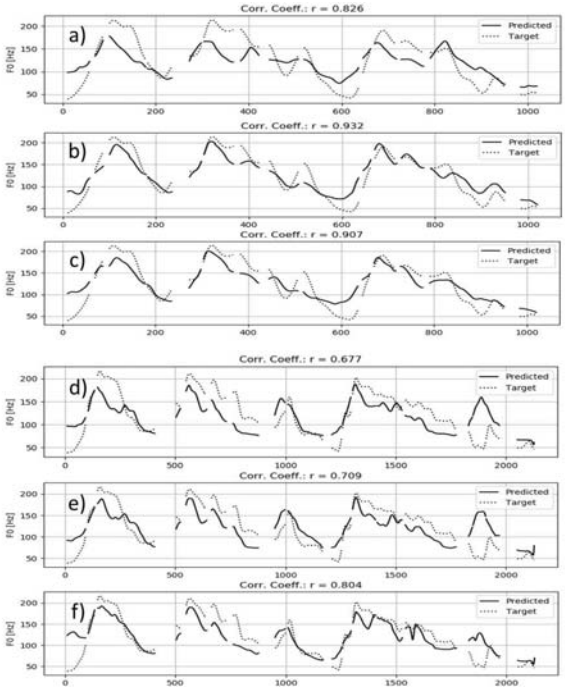

Fig. 4. Samples from predicted F_0 patterns. a) $\sim$ c) represent obtained patterns for normal speech in the third scenario for 8, 23 and 38 clusters, respectively. d) $\sim$ f) the corresponding ones for EL speech in the fourth scenario.

F_0 patterns for EL speech, though with smaller correlation coefficients. This could provide us with a trade-off between resulting accuracies for F_0 patterns versus the price we need to pay for developing a recognition system to extract phoneme labels for practical applications. Evaluating the accuracy of ASR system while extracting phoneme labels from EL speech, and its impact on the overall performance will be investigated in our future work.

Acknowledgment

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

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] L. Diener, T. Umesh, , and T. Schultz, "Improving Fundamental Frequency Generation in EMG-to-Speeech Conversion using a Quantization Approach," in *Automatic Speech Recognition and Understanding*, 2019.

[3] 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.

[4] 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.

[5] 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.

[6] A. Fuchs, M. Hagmüller, and G. Kubin, "The new bionic electro-larynx speech system," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, pp. 952–961, 08 2016.
[7] 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.
[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 2018*, 2018, pp. 122–126. [Online]. Available: http://dx.doi.org/10.21437/IberSPEECH.2018-26
[9] L. hui Li, T. Toda, K. Morikawa, K. Kobayashi, and S. Makino, "Improving singing aid system for laryngectomees with statistical voice conversion and vae-space," in *ISMIR*, 2019.
[10] Z. Ling, S. Kang, H. Zen, A. Senior, M. Schuster, X. Qian, H. M. Meng, and L. Deng, "Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 35–52, 2015.
[11] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. 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," 08 2017, pp. 4006–4010.
[12] 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," 04 2018, pp. 4779–4783.
[13] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," 07 2016, pp. 1–6.
[14] 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 Speech Synthesis Workshop*, 2019, pp. 251–256. [Online]. Available: http://dx.doi.org/10.21437/SSW.2019-45
[15] 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.
[16] 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.
[17] 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.
[18] 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.
[19] 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
[20] 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
[21] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 12 2014.
[22] D. Hermes, "Measuring the perceptual similarity of pitch contours," *Journal of speech, language, and hearing research: JSLHR*, vol. 41, pp. 73–82, 03 1998.