INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Parallel vs. Non-parallel Voice Conversion for Esophageal Speech

Luis Serrano, Sneha Raman, David Tavarez, Eva Navas, Inma Hernaez

University of the Basque Country (UPV/EHU)
{lserrano, sneha, david, eva, inma}@aholab.ehu.eus

Abstract

State of the art systems for voice conversion have been shown to generate highly natural sounding converted speech. Voice conversion techniques have also been applied to alaryngeal speech, with the aim of improving its quality or its intelligibility. In this paper, we present an attempt to apply a voice conversion strategy based on phonetic posteriorgrams (PPGs), which produces very high quality converted speech, to improve the characteristics of esophageal speech. The main advantage of this PPG based architecture lies in the fact that it is able to convert speech from any source, without the need to previously train the system with a parallel corpus. However, our results show that the PPG approach degrades the intelligibility of the converted speech considerably, especially when the input speech is already poorly intelligible. In this paper two systems are compared, an LSTM based one-to-one conversion system, which is referred to as the baseline, and the new system using phonetic posteriorgrams. Both spectral parameters and f_0 are converted using DNN (Deep Neural Network) based architectures. Results from both objective and subjective evaluations are presented, showing that although ASR (Automated Speech Recognition) errors are reduced, original esophageal speech is still preferred by subjects.

Index Terms: voice conversion, speech and voice disorders, alaryngeal voices, speech intelligibility

1. Introduction

Laryngectomees are people whose larynx has been surgically removed. The larynx is a fundamental organ in speech production, since the vocal folds are located inside it. Even in the absence of the larynx and the vocal folds, it is still possible to utter intelligible speech using alternative vibrating elements. The acquired new voice is called the alaryngeal voice. There are three main ways to produce the alaryngeal voice: esophageal (ES), electrolaryngeal (EL) and tracheoesophageal (TES) speech. The experiments described in this paper use only esophageal speech.

Unlike EL and TES, the production of ES does not require any device. It is learned with the help of a speech therapist. With this method, the pharyngo-esophageal segment is used as a substitutive vibrating element for the vocal folds. Due to the nature of the anatomical alteration, the air used to create vibrations in the esophagus cannot come from the lungs and the trachea as happens during normal speech production. Instead, the air is swallowed from the mouth and introduced in the esophagus, and then expelled in a controlled way which produces the vibration.

These huge differences in the production mechanisms lead to a diminution of naturalness and intelligibility [1, 2, 3]. As a consequence, the communication with other human beings is hindered. Moreover, these less intelligible voices are an added problem for the automatic speech recognition algorithms that are becoming ubiquitous in human computer interaction technologies. The work presented in this paper aims to improve the quality and intelligibility of ES, with the final goal of contributing to a better life for the laryngectomee.

There have been different approaches to enhance the quality and intelligibility of alaryngeal voices. One possible approach is to make use of a voice conversion (VC) system. In a VC system, the objective is to transform utterances from a given source speaker into a specific target speaker, i.e. apply some techniques to perceive the sentences as uttered by the specific target speaker. For the purpose of enhancing alaryngeal voices, a healthy speaker is chosen as the target speaker. Different examples of techniques based on statistical voice conversion can be found in [4], [5] or [6], where the characteristics of the target speaker can be tuned to obtain a more personalized converted voice.

The VC process can be divided in two stages: Firstly, a training phase is needed in order to learn the correspondences between source and target acoustic features. These learned relationships are then stored in the form of a conversion function. The second step is the conversion itself. The conversion function is applied to transform new input utterances from the source speaker. Although the identity of the speaker is also contained in the suprasegmental (prosody) and even linguistic features, VC research has been focused mostly on mapping spectral features [7, 8, 9].

VC is a field that has been researched for a long time. An exhaustive recent review of the field can be found in [10]. A wide variety of approaches have been proposed to obtain the conversion function; from codebooks [11, 12] and hidden Markov models [13, 14, 15] to Gaussian Mixture Models (GMMs) [16, 7, 17, 18] or Gaussian processes [19]. In recent years, special focus has been given to shallow/deep neural network (S/DNN) solutions [20, 21, 9, 22, 23, 24].

In this paper we propose taking advantage of the recent advances in machine learning techniques to train a deep learning system to convert the speech of an esophageal speaker to a healthy speaker. In a previous paper [25], we proposed a long short-term memory (LSTM) network to perform this conversion. Here, we have implemented a Phonetic Posteriorgram (PPG) based conversion following the approach related in [26] for healthy speakers. A PPG based conversion is advantageous in that it does not need data from the source speaker for its training. To obtain the PPGs, it is necessary to use a speaker-independent automatic speech recognition (ASR) system. We have built such an ASR system for esophageal speech. The final goal is to compare the performance of the PPG based (non parallel) and LSTM based (parallel) systems.

The paper begins with the description of the VC systems, followed by the experimental setup. Then the results of both objective and subjective evaluation of the proposed conversion systems are presented and discussed. Finally, we conclude with a brief discussion of the results.

http://dx.doi.org/10.21437/Interspeech.2019-2194

2. The voice conversion system

In this section we compare two voice conversion systems. The first system uses an LSTM neural network to convert cepstrum coefficients of an esophageal source speaker to those of a healthy target speaker. The system is described in detail in [25], although a different intonation module is used for the work documented in this paper. The second system is based on PPGs. For both systems, the same method has been used to obtain $\log f_0$ curves. This method is described in subsection 2.3.

2.1. LSTM-based spectral conversion using parallel data

The architecture of this system is described in detail in [25]. We will briefly describe it here.

The LSTM network is trained with an aligned set of parallel source-target vectors. These vectors contain Mel-cepstral coefficients (MCEP) and their first order derivatives, normalized in mean and variance. The MCEP coefficients will be used by the vocoder to synthesize the spectrum. The first order derivatives are needed to perform Maximum Likelihood Parameter Generation (MLPG) [7] on the output of the network to mitigate over-smoothing. In addition, to improve naturalness, we use the global variance calculated from the target healthy speaker.

2.2. PPGs based spectral conversion

As mentioned in the introduction, the system implemented here follows the architecture used in [26]. The two main components are an ASR system, which is used to obtain the PPGs of the input source signal, and a network (a DBLSTM in the case of the cited reference) trained to predict the corresponding acoustic parameters of the target speaker from those PPG vectors. The ASR system is speaker independent (SI-ASR) so that it is assumed that it is able to produce speaker independent PPGs (SI-PPGs). Trained with the SI-PPGs produced by the SI-ASR for the target speaker, and the acoustic vectors of the same target speaker, the DNN is able to convert from any unknown source speaker to the specific target speaker. It is then assumed that the SI-ASR system will produce similar PPGs for different input source speakers (i.e. SI-PPGs), which is reasonable if the ASR performs adequately. However, a standard SI-ASR will perform differently for a non-standard input signal. If that is the case, the PPGs obtained will be very different from those used to train the DNN and the resulting output signal might show a low degree of intelligibility, although the quality of the conversion, in terms of keeping the identity of the target speaker, might still be good. For example, if there is an incorrect or special pronunciation of a certain phone by the source speaker, the PPGs obtained will differ significantly from those used in the training of the DNN and consequently the network will not produce the expected output (i.e. the special pronunciation of that phone with the identity of the target speaker).

Therefore, in the case of esophageal signals, whose acoustic parameters differ greatly from those of laryngeal speech, a standard SI-ASR system can not be used. Instead, we used a specifically designed ASR system, as shown in Figure 1. The first step was to develop an esophageal speaker recognizer, using only esophageal signals to train it and limiting the number of senones to 150. Next, using MAP (Maximum a Posteriori) adaptation, the acoustic GMM models are adapted to the specific target healthy speaker, with the aim of reducing the differences between both acoustic models. These adapted models are used to train the p-norm neural network that will produce the PPGs.

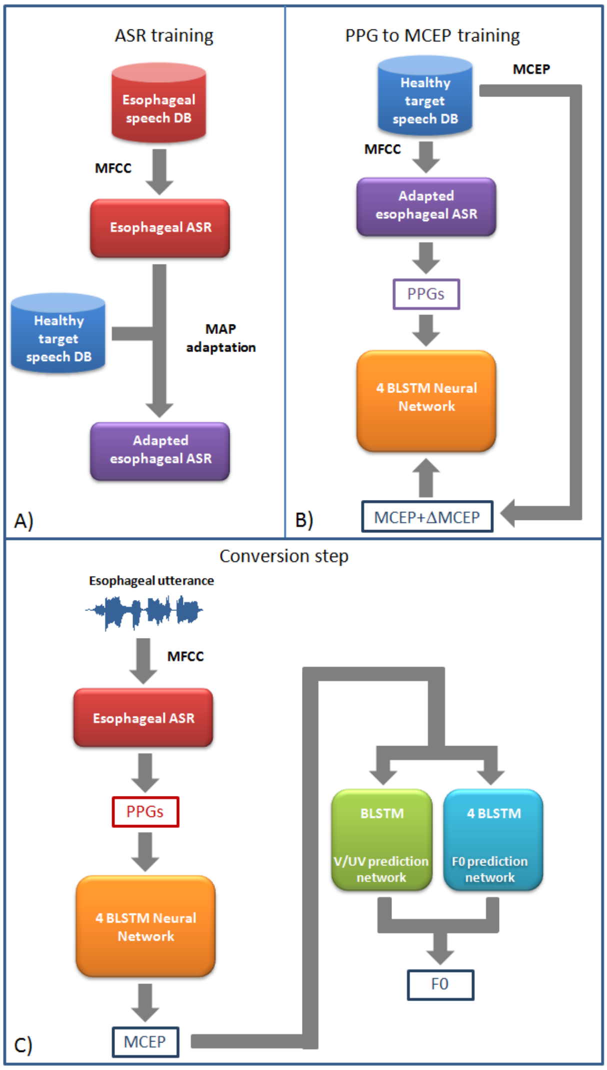

Figure 1: *Schematic diagram of the VC system using PPGs. A) Obtention of the ASR system adapted to esophageal voices which will be used to obtain the PPGs. B) Training of the DNN used for spectral conversion C) Prediction of MCEP coefficients and F0 values*

The network converting from the PPGs to the converted acoustic parameters is a 4 layer BLSTM (Bidirectional LSTM) followed by a fully connected layer. The converted parameters feed the vocoder, together with the estimated f_0 explained next.

2.3. Fundamental frequency estimation

Two different networks have been used for the estimation of the f_0 curve, one to obtain the $\log f_0$ value, and another one to estimate the voiced/unvoiced (V/UV) decision. Both $\log f_0$ and the V/UV decision are predicted from the spectral parameters, MCEP, obtained from the healthy target speaker. The DNN used for $\log f_0$ estimation is a 4 layer BLSTM followed by a fully connected layer with linear activation, while for the V/UV decision a network comprised of one BLSTM layer followed by a fully connected layer with sigmoid activation is used.

3. Experimental setup

3.1. Databases of healthy and esophageal speakers

Two speech datasets have been used, one for the esophageal voices and another one for the healthy voice selected as target speaker. The healthy voice used in this experiment as target speaker contains 3995 recordings made by one male speaker [27].

The esophageal speech dataset consists of recordings of 100 phonetically balanced sentences from 30 esophageal speakers (25 male, 4 female, and one tracheoesophageal male). One esophageal male speaker out of the 30 available was chosen as source speaker.

3.2. Esophageal ASR system

To obtain the PPGs, we use the neural network architecture of the Kaldi ASR toolkit [28].

This recognizer is trained using the recordings of 29 esophageal speakers (90 utterances for each speaker, resulting in 2610 sentences). The training process follows the Kaldi WSJ recipe, but in the last iteration of the GMM phase training, we changed the number of leaves to 150 to limit the number of phonetic classes (senones in this case). This model is the starting point to train the p-norm neural network that gave us the PPGs for the esophageal speaker. In an effort to make the healthy target speaker's and the esophageal speaker's models closer, we apply a MAP adaptation to the GMM models with all the healthy target speaker training material (3895 sentences). This adapted model is the one used to obtain the PPGs for the healthy target speaker utterances (part A of Figure 1) and will be used to train the PPG to MCEP conversion network.

Using 10-fold cross-validation, 10 different systems have been obtained in order to have the parameters of 100 utterances. In the conversion step, unknown sentences from the esophageal source speaker will go through the esophageal ASR to obtain the PPGs (see part C of Figure 1).

3.3. Spectral conversion

To train the subsystem that will convert from PPGs to MCEPs, we implemented a neural network with 4 BLSTM layers, each layer composed of 64 cells. The input to this network are the PPGs obtained with the adapted ASR using 3895 sentences of the healthy target speaker (Figure 1 B).

A fully connected layer provides the output. This output consists of a vector of dimension 48, with 1^{st} to 24^{th} cepstral coefficients and their first order derivatives obtained using Ahocoder [29]. The c_0 is taken from the source speaker.

For the training process, we have chosen a batch size of 50. The loss function to minimize is the Mean Square Error (MSE). The training was carried out over 25 epochs, with a dropout ratio of 0.2 and using the $Adam$ optimizer.

3.4. Intonation generation network

To obtain the intonation curve, two networks were used. One to obtain a V/UV decision and another to obtain the intonation curve. Informal experimental evaluation showed that this configuration works better than using only one network. Both networks are trained with 25 MCEP coefficients normalized in mean and variance obtained from the healthy speaker dataset (3895 sentences are used in the training process and 100 for validation).

For the V/UV decision network, the BLSTM layer contains 64 cells. The V/UV output vector used in the training is obtained directly from the logf_0 curve. The network is optimized using the $Adam$ algorithm with a batch size of 50, and was trained over 100 epochs, using the binary cross-entropy as loss function. A dropout ratio of 0.2 was applied.

Each BLSTM layer of the network used for the logf_0 prediction is made of 64 cells. During training the logf_0 curve is linearly interpolated in unvoiced frames. Then the delta is calculated and appended, and mean and variance normalization is applied. The training configuration is the same as the one used in the V/UV network, with but in this case the metric we search to minimize is the MSE (Mean Square Error).

In the conversion stage, the 25 MCEP coefficients obtained from the spectral conversion network are used to predict the values of the logf_0 and the V/UV vector (see part C in Figure 1).

4. Evaluation

The converted esophageal speech has been evaluated objectively by means of Word Error Rate (WER) from an ASR system and Mel Cepstral Distortion (MCD) measures. Subjective evaluation was performed using a preference test[1].

4.1. Objective evaluation

An ASR system for Spanish, built with Kaldi, was used to calculate the WER of the converted sentences. From an acoustic point of view, the ASR is a general purpose system, trained with healthy speech (see [30] for details). However, a limited lexicon was created from the corpus of 100 sentences used in the experiment (701 different words). This was done because with the general purpose dictionary there was a 23% of out-of-vocabulary (OOV) words, due to the low-frequency of some words included in the sentences to ensure good phonetic coverage. The language model used is a simple unigram model with equal probability for all the words. This ASR system is not comparable to state-of-the-art ASR systems in terms of vocabulary and language model used, but it is more suitable for our purpose of comparing different speech conversion strategies.

Table 1 shows the recognition results for the 4 sets of 100 sentences: original esophageal source speaker, healthy target speaker and the converted signals from the systems under evaluation. The WER of the original esophageal source speaker is 56.93%, well above the 11.88% obtained for the healthy target speaker, indicating the difficulty of recognizing alaryngeal speech for systems trained with healthy voices.

Table 1: *WER results for the different voice versions*

Case	WER (%)
Source (esophageal)	56.93
Target (healthy)	11.88
LSTM converted cepstrum + estimated f_0	40.58
PPGs converted cepstrum + estimated f_0	57.91

The parallel LSTM based conversion system has managed to fix some of the problems of esophageal cepstrum and gets them closer to those of the healthy voice, improving the recognition results by 16 percentage points. However, the non-parallel conversion system built with PPGs obtains a slightly worse value than the one given by the original esophageal sentences.

[1] At https://aholab.ehu.eus/users/lserrano/inter19/demointer19.html some examples can be found

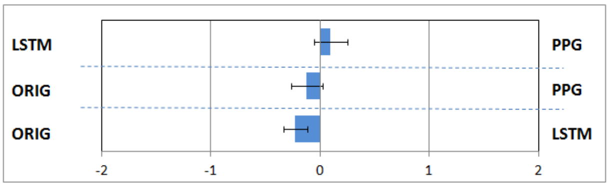

Figure 2: *Averaged preference results with confidence intervals*

This can be explained because although the recognizer used to extract the PPGs was trained with esophageal speech, the acoustic scores of the esophageal speaker are too far from those of healthy speech. Therefore, an acoustic class error is present in the obtained PPGs which propagates to the conversion phase. This error did not have the same relevance when the conversion described in [26] was tested using a pair of healthy source-target speakers because their acoustic spaces are closer than the acoustic spaces of healthy and esophageal speech.

In addition, we have calculated MCD [31] between target speech and both converted versions of esophageal speech to have an idea of the spectral distance between converted voices and target voice. The results are shown in Table 2. Both VC strategies bring the source signal closer to healthy speech and there is no significant difference between the two conversion methods.

Table 2: *MCD measures*

Case	MCD (dB)
Target & Source	7.840 ± 0.376
Target & LSTM converted	5.025 ± 0.314
Target & PPG converted	5.021 ± 0.313

4.2. Subjective evaluation

A perceptual test was carried out to determine which version is preferred by the listeners: the original source speech or the converted speech produced by either the parallel or the non-parallel VC methods. Out of the 100 available sentences, we used a set of the 30 most intelligible sentences for the preference test. The 30 sentences chosen were rated most intelligible by the above mentioned ASR system across all three speech types (original ES, LSTM converted and PPG converted).

From these 30 sentences, each listener had to evaluate 24 randomly chosen sentence pairs, 8 comparing original speech with speech converted by the LSTM system, 8 comparing original speech with speech converted by the PPG system, and 8 comparing speech converted by each of the VC systems. Listeners could only listen to each stimulus once and then they had to express their preference on a five point scale: I strongly prefer sentence 1 (-2), I prefer sentence 1 (-1), I cannot decide on either of the two sentences (0), I prefer sentence 2 (1), I strongly prefer sentence 2 (2).

Thirty-five native Spanish speakers took part in the test. The results can be seen in Figure 2, where the averaged preference rates are shown together with 95% confidence intervals. The original sentences are preferred over the ones converted by the LSTM system by a very small but statistically significant margin. This result corroborates those previously obtained described in [25]. When comparing the original sentences and those coming from the PPG conversion, the listeners do not show a significant preference. When comparing the two converted versions the listeners show a slight preference for the PPG system, but with no statistical significance. Figure 3 shows with more detail the degree of preference for each pair of systems. We can observe than when comparing the PPG vs. LSTM most people cannot decide between the two of them (39.6%). In contrast, in the case of PPG vs. original speech only 13.6% consider them equivalent, 48.9% prefer the original version and 37.5% the PPG converted version. For the LSTM vs. orginal case, 15% are undecided, 50.7% prefer the original version and 34.3% prefer the LSTM converted.

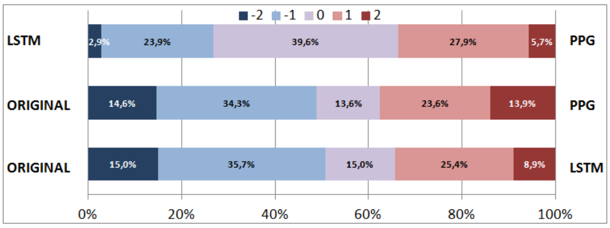

Figure 3: *Detailed results of the preference test*

The results of the subjective evaluation contrast with the ones obtained in the objective evaluation. The system with best WER (LSTM) is not the one preferred by the listeners, probably because even if it is more intelligible, these signals sound less natural than the other two versions.

5. Conclusions

In the work presented in this paper we evaluate the performance of a voice conversion architecture using non-parallel data with PPGs, adapted to the problem of improving the quality and/or the intelligibility of esophageal speech. This system is then compared with a more classical approach, that is an LSTM network and a set of parallel source-target sentences.

The intelligibility has been measured by means of WER using ASR. The results show that while the LSTM approach improves the recognition rate, the PPGs based system does not, with slightly worse results than those obtained by the original esophageal sentences. In terms of MCD, both methods reduce the spectral distance to the target equally.

The results of the subjective evaluation show that there is no a clear preference in favor of the converted sentences. The improved quality that was expected with the non-parallel strategy was blurred by the loss of intelligibility of the converted sentences. We think that the small amount of data available to train the different systems is the main cause of the poor final performance.

6. Acknowledgements

This work has been partially funded by the Spanish Ministry of Economy and Competitiveness with FEDER support (RESTORE project, TEC2015-67163-C2-1-R), the Basque Government (BerbaOla project, KK-2018/00014) and from the EUs H2020 research and innovation programme under the MSCA GA 67532*4 (the ENRICH network: www.enrich-etn.eu).

7. References

[1] B. Weinberg, "Acoustical properties of esophageal and tracheoesophageal speech," *Laryngectomee rehabilitation*, pp. 113–127, 1986.

[2] T. Most, Y. Tobin, and R. C. Mimran, "Acoustic and perceptual characteristics of esophageal and tracheoesophageal speech production," *Journal of communication disorders*, vol. 33, no. 2, pp. 165–181, 2000.

[3] T. Drugman, M. Rijckaert, C. Janssens, and M. Remacle, "Tracheoesophageal speech: A dedicated objective acoustic assessment," *Computer Speech & Language*, vol. 30, no. 1, pp. 16–31, 2015.

[4] H. Doi, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Statistical approach to enhancing esophageal speech based on gaussian mixture models," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 4250–4253.

[5] ——, "Esophageal speech enhancement based on statistical voice conversion with gaussian mixture models," *IEICE TRANSACTIONS on Information and Systems*, vol. 93, no. 9, pp. 2472–2482, 2010.

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

[7] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[8] D. Erro, E. Navas, and I. Hernaez, "Parametric voice conversion based on bilinear frequency warping plus amplitude scaling," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 3, pp. 556 – 566, 2013.

[9] L.-H. Chen, Z.-H. Ling, L.-J. Liu, and L.-R. Dai, "Voice conversion using deep neural networks with layer-wise generative training," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 12, pp. 1859–1872, 2014.

[10] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Communication*, vol. 88, pp. 65–82, 2017.

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

[12] L. M. Arslan, "Speaker transformation algorithm using segmental codebooks (STASC)," *Speech Communication*, vol. 28, pp. 211–226, 1999.

[13] A. Bonafonte, A. Kain, J. v. Santen, and H. Duxans, "Including dynamic and phonetic information in voice conversion systems," in *Eighth International Conference on Spoken Language Processing*, 2004.

[14] C. Lee, C. Wu, and J. Guo, "Pronunciation variation generation for spontaneous speech synthesis using state-based voice transformation," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, March 2010, pp. 4826–4829.

[15] H. Zen, Y. Nankaku, and K. Tokuda, "Continuous stochastic feature mapping based on trajectory HMMs," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 417–430, 2011.

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

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

[18] H. Benisty and D. Malah, "Voice conversion using GMM with enhanced Global Variance," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, no. August, pp. 669–672, 2011.

[19] N. Xu, Y. Tang, J. Bao, A. Jiang, X. Liu, and Z. Yang, "Voice conversion based on gaussian processes by coherent and asymmetric training with limited training data," *Speech Communication*, vol. 58, pp. 124–138, 2014.

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

[21] S. Desai, A. W. Black, B. Yegnanarayana, and K. Prahallad, "Spectral mapping using artificial neural networks for voice conversion," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 954–964, 2010.

[22] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 4869–4873.

[23] Z.-H. Ling, S.-Y. Kang, H. Zen, A. Senior, M. Schuster, X.-J. 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.

[24] I. B. Othmane, J. Di Martino, and K. Ouni, "Enhancement of esophageal speech obtained by a voice conversion technique using time dilated fourier cepstra," *International Journal of Speech Technology*, vol. 22, no. 1, pp. 99–110, 2019.

[25] 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

[26] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *2016 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2016, pp. 1–6.

[27] I. Sainz, D. Erro, E. Navas, I. Hernáez, J. Sanchez, I. Saratxaga, and I. Odriozola, "Versatile Speech Databases for High Quality Synthesis for Basque," in *8th international conference on Language Resources and Evaluation (LREC)*, 2012, pp. 3308–3312. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2012/pdf/126_Paper.pdf

[28] X. Zhang, J. Trmal, D. Povey, and S. Khudanpur, "Improving deep neural network acoustic models using generalized maxout networks," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 215–219.

[29] D. Erro, I. Sainz, E. Navas, and I. Hernaez, "Harmonics plus noise model based vocoder for statistical parametric speech synthesis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 184–194, 2014.

[30] L. Serrano, D. Tavarez, I. Odriozola, I. Hernaez, and I. Saratxaga, "Aholab system for albayzin 2016 search-on-speech evaluation," in *IberSPEECH*, 2016, pp. 33–42.

[31] M. Mashimo, T. Toda, K. Shikano, and N. Campbell, "Evaluation of Cross-Language Voice Conversion," *Proc. Eurospeech*, pp. 361–364, 2001. [Online]. Available: http://isca-speech.org/archive/eurospeech_2001/e01_0361.html