IberSPEECH 2018
21-23 November 2018, Barcelona, Spain

LSTM based voice conversion for laryngectomees

Luis Serrano, David Tavarez, Xabier Sarasola,
Sneha Raman, Ibon Saratxaga, Eva Navas, Inma Hernáez

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

Abstract

This paper describes a voice conversion system designed with the aim of improving the intelligibility and pleasantness of oesophageal voices. Two different systems have been built, one to transform the spectral magnitude and another one for the fundamental frequency, both based on DNNs. Ahocoder has been used to extract the spectral information (mel cepstral coefficients) and a specific pitch extractor has been developed to calculate the fundamental frequency of the oesophageal voices. The cepstral coefficients are converted by means of an LSTM network. The conversion of the intonation curve is implemented through two different LSTM networks, one dedicated to the voiced unvoiced detection and another one for the prediction of F0 from the converted cepstral coefficients. The experiments described here involve conversion from one oesophageal speaker to a specific healthy voice. The intelligibility of the signals has been measured with a Kaldi based ASR system. A preference test has been implemented to evaluate the subjective preference of the obtained converted voices comparing them with the original oesophageal voice. The results show that spectral conversion improves ASR while restoring the intonation is preferred by human listeners.

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

1. Introduction

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

Unlike the other two methods, the production of ES does not require any device. This kind of speech is learned with the help of a speech therapist. In this method the pharyngo-oesophageal segment is used as a substitutive vibrating element for the vocal folds. Due to the nature of the intervention, the air used to create the vibration of the oesophagus can not 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 oesophagus, being then expelled in a controlled way while producing 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 others is hindered. Moreover, these less intelligible voices are an added problem for the automatic speech recognition algorithms that are becoming ubiquitous in the human computer interaction technologies. The work presented in this paper aims at improving the quality and intelligibility of ES, with the final purpose of contributing to a better life of the laryngectomee.

There have been different approaches to enhance the quality and the intelligibility of alaryngeal voices. Some research works use the source-filter analysis of the pathological signal and focus on the modifications of one or both source and filter. An example of this approach can be found in [4] where an adaptive gain equalizer algorithm is used to modify the source of ES; or in [5] where a reconstruction of normal sounding speech for laryngectomy patients is attempted through a modified CELP codec. Another approach is to work with the prosodic elements. In [6], the pitch information extracted from an electroglottograph (EGG) is used to create a synthetic glottal signal, reducing jitter and shimmer. Additionally, spectral smoothing and tilt correction were applied. These modifications reduced the harshness and breathiness of the TE speech. The same authors relate in [7] a repairing method of the durations of the pathological phonemes. In the same line, [8] presents a system where concatenation of randomly chosen healthy reference patterns replaces the pathological excitation, adjusting the short, medium and long-term variability of the pitch.

A different approach to the problem is to make use of a voice conversion (VC) system. In a VC system the pursued goal 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 target speaker. Different examples of techniques based on statistical voice conversion can be found in [9], [10] or [11], 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 laying 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 [12, 13, 14].

The VC is a field that has been researched for a long time, an exhaustive recent review of the field can be found in [15]. A wide variety of approaches have been proposed to obtain the conversion function: from codebooks [16, 17] and hidden Markov models [18, 19, 20] to Gaussian Mixture Models (GMMs) [21, 12, 22, 23] or Gaussian processes [24]. In the last years, special focus has been given to shallow/deep neural networks (S/DNNs) solutions [25, 26, 14, 27, 28].

10.21437/IberSPEECH.2018-26

In this paper we propose to take advantage of the recent advances in machine learning techniques to train a deep learning system to convert from the audio of an oesophageal speaker to a healthy speaker. Though the literature focuses specially on spectral conversion, the importance of prosody in oesophageal speech can not be left out. A good intonation is important for the utterances to be perceived as more natural and pleasant. Therefore, the proposed system will not only convert the spectral features but will also estimate f_0. The spectral and f_0 conversion contributions have been then evaluated separately by means of word error rate (WER) and a perceptual test.

2. Voice conversion system

The proposed architecture uses two parallel DNN based systems to convert separately the spectral envelope and the fundamental frequency curve f_0. They will be described in detail in the following subsections.

The acoustic analysis is performed using Ahocoder [29]. This vocoder does the harmonic analysis of the speech audio files every 5 ms, extracting the Mel-cepstral (MCEP) representation of the spectral envelope, $\log f_0$ and maximum voiced frequency (MVF). The MVF is related to the harmonicity of the signal and will not be converted (the MVF value obtained from the source signal is used). To extract f_0 from the alaryngeal signals, the default method of Ahocoder has been replaced by a more specific method, designed to cope with the irregularities present in these signals. This strategy mainly consists on using the autocorrelation method for pitch extraction over the residue signal of the PSIAIF analysis [30]. In this work 25 MCEP coefficients in combination with their first order derivatives are used for spectral conversion. In the experiment described here the 0^{th} coefficient (related to the energy), is not converted but directly copied from the source to the target. Similarly, the source MVF is not modified and it is directly copied to the target. Other alternatives such us using a constant MVF were also experimented but showed no significant differences in informal tests. For pitch prediction 40 MCEP coefficients are used. This number was chosen due to quality reasons observed during the training stages.

For the networks architecture, the chosen approach was to use a Long Short-Term Memory (LSTM) architecture. This type of neural network allows the net to retain (and gradually forget) information about previous time instants taking advantage of the strong temporal dependency that exists between consecutive frames in the speech signals, specially for the f_0 curve. The neural networks are implemented in Python using the Keras library.

2.1. Spectral conversion

Before training the LSTM network, it is necessary to align the source and target utterances. Due to the characteristics of the oesophageal speech, there is a very important mismatch between both healthy and oesophageal signals which causes the inadequacy of using a dynamic time warping (DTW) algorithm directly [31]. This is why both signals were labelled at phone level, and then the iterative alignment procedure described in [32] and [33] was applied for each pair of oesophageal and healthy phones.

With the alignment results, the inputs and outputs of the network are built. The corresponding first order derivatives are appended to the 24 source aligned cepstra (c_0 and c_0' excluded). All vectors from all sentences are appended one after the other, resulting in two matrices for the source-target speaker pair. Then, mean and variance normalization is applied to each dimension of the cepstrum vector and its derivative. In addition, a voiced/unvoiced decision vector in the form of 0 or 1 obtained from the $\log f_0$ values is appended, so the dimension of the input is finally 49. The resulting matrix is divided in sequences of 50 frames before entering the net.

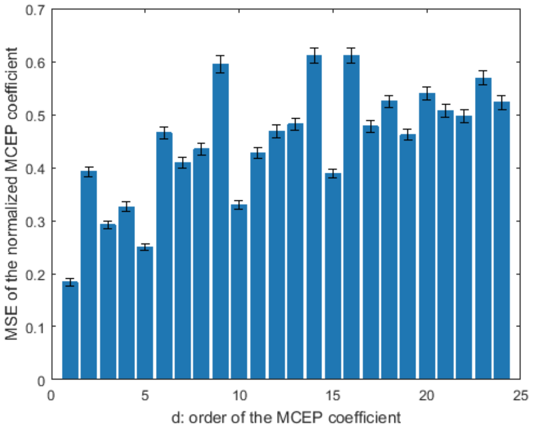

Figure 1: *MSE of each normalized MCEP coefficient with the 95% confidence interval for one fold - 10 sentences (test data)*

The net will predict the 48 MCEP coefficients of the target speaker and the voiced/unvoiced decision vector. The metric we search to minimize is the MSE.

The number of cells of the LSTM layer is 100. As stated before, the number of frames that composes each batch sequence is 50, and the input dimension is equal to 49. The output is obtained from a fully connected layer and gives us a sequence of 50 frames and dimension 49 for each sequence in the input.

We used the RMSProp optimizer (divide the gradient by a running average of its recent magnitude) with a learning rate of 0.0001. To avoid overfitting, a dropout ratio of 0.5 was used. We trained the network for 60 epochs.

The performance of the network has been analysed comparing the predicted result with the target data. As the order of the coefficients goes up, the similarity between the predicted MCEP and the target one goes down. This can be seen by calculating the MSE for each normalized cepstral coefficient between the predicted and target MCEPs (Figure 1).

2.1.1. Maximum Likelihood Parameter Generation (MLPG)

In order to solve the problem of the smoothing present in the conversion process, we apply the maximum likelihood parameter generation algorithm [12]. We consider the MCEP obtained from the LSTM spectral conversion network as the mean vectors of a Gaussian distribution. Same as in [34], the covariance matrix is obtained from the squared error between the original features and the predicted converted vectors.

The global variance (GV) is calculated from the target training data and is used in the MLPG to do the spectral conversion along the mean vectors and covariance matrices for each test sentence.

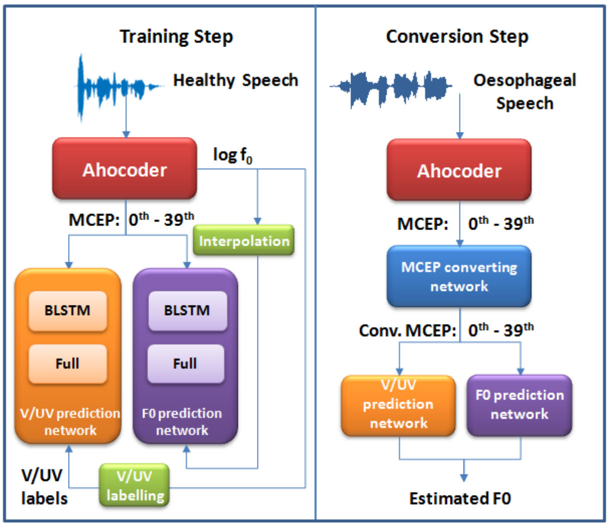

Figure 2: *Architecture of the f_0 estimation system for training and conversion stages.*

2.2. Fundamental frequency estimation

Two different networks have been used to perform the estimation of the fundamental frequency, one for the prediction of f_0 and another one for U/V decision estimation (see Figure 2). In both cases, the LSTM model is trained to map the relationship between the healthy target MCEPs and their corresponding f_0 sequence. Mean and variance normalization is applied to each dimension of the cepstrum vector. Log f_0 is linearly interpolated in unvoiced frames and mean and variance normalization is also applied in this case. Development experiments showed that the use of 40 MCEP coefficients provided better results for the U/V estimation than using only 25 coefficients, therefore, 40 MCEP (no derivatives) are used for the intonation modelling.

For the U/V decision we use one Bidirectional LSTM layer with 64 cells. A fully connected layer with sigmoid activation provides the final output. The network is optimized using the Adam algorithm with minibatches of size 10 and trained for 60 epochs. A dropout ratio of 0.2 was applied as regularization. We employ the binary cross-entropy as loss function.

The same configuration is used for the f_0 prediction: one Bidirectional LSTM layer with 64 cells followed by a fully connected layer, with linear activation in this case. Also the MSE is now the metric we search to minimize, as it occurs for the cepstrum network.

In the conversion stage, 40 converted MCEP coefficients are needed. Therefore another network has been trained, similar to the one used for spectral conversion, to obtain the necessary number of coefficients. Finally, the $\log f_0$ is estimated from the converted MCEPs by the two conversion networks.

2.3. Training and testing data

Due to the nature of the data, the amount of data available for training and testing the conversion system is scarce. The parallel data used are the recordings of 100 phonetically balanced sentences selected from a bigger corpus [35] made by one healthy speaker and one oesophageal speaker. Out of the 100 utterances, 90 are chosen for training and 10 for testing the system, using 10-fold cross-validation.

3. Evaluation

3.1. Objective evaluation: Kaldi ASR

To have an objective measure (WER) of the effect of the conversion we have prepared an automatic speech recognizer (ASR) for Spanish using the Kaldi toolkit [36]. It is implemented following the recipe s5 for the Wall Street Journal database. The acoustic features used are 13 Mel-Frequency Cepstral Coefficients (MFCCS) to which a process of mean and variance normalization (CMVN) is applied to mitigate the effects of the channel.

The training begins with a flat-start initialization of context-independent phonetic Hidden Markov Models (HMM), and then a series of accumulative trainings are done. For the final step of the recognizer, a neural network is trained. The input features to the neural network consist of a series of 40-dimensional features. The network sees a window of these features, with 4 frames on each side of the central frame. The features are derived by processing the conventional 13-dimensional MFCCs. The necessary steps are described in [37], and basically consist in applying a series of transformations to the normalized cepstra: first linear discriminant analysis (LDA), then maximum likelihood linear transform (MLLT) and global feature-space maximum likelihood linear regression (fMLLR). At the recognition stage, the same transformations are applied to the test data, handling them as a block.

The audio material used to train the ASR is described in [38]. However, although the acoustic models have been maintained, the lexicon has been created from the 100 sentences corpus used in the experiment (701 words). This has been done because using the original lexicon (with $37,632$ entries) as much as 23% of the words were out of vocabulary (OOV) words. This is due to the fact that the sentences are phonetically balanced and many sentences containing proper names and many unusual words were chosen to maximize the variability of the phonetic content. Together with this reduced lexicon, a unigram language model with equally probable words has been used. Although the final WER numbers obtained this way are not comparable to a realistic ASR situation, the procedure serves our purpose of evaluating the ability of the conversion system to improve (or not) the performing of any ASR system.

Three different recognition experiments have been carried out. The first one does the recognition of the 100 original sentences recorded by the laryngectomee speaker. The second one evaluates the resynthesized audio signals with Ahocoder using the original cepstrums and the estimated f_0 obtained from the network. The last one uses as input the 100 sentences resynthesized with the converted cepstrum and the estimated f_0.

The results are shown in Table 1. The WER of the original oesophageal signal is very high (56.93% vs 10.15% obtained for the target healthy speaker) showing the difficulty of the task. As expected, the system performs very similarly when only f_0 is changed. The small differences are probably due to the resynthesis process and recalculation of the parameters from the waveforms performed by Kaldi. However, when the recognized signals are those which are resynthesized using the converted cepstrum and f_0, the WER drops by a 15% in absolute terms showing that some problems present in the oesophageal spectrum are corrected by the conversion system. The WER of the converted signals is though far away from the performance for the healthy voice.

Table 1: *WER results for the different experiments*

Case	WER (%)
original healthy	10.15
original oesophageal	56.93
original spectrum + estimated f_0	57.91
converted spectrum + estimated f_0	41.48

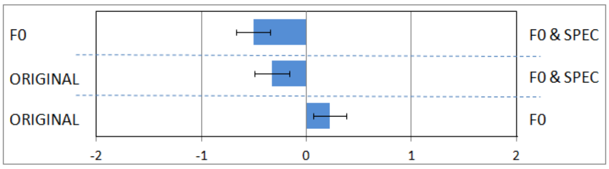

Figure 3: *Averaged preference results with confidence intervals (-2:I strongly prefer sentence 1, -1:I prefer sentence 1, 0:I don't have any preference, 1:I prefer sentence 2, 2: I strongly prefer sentence 2).*

3.2. Subjective evaluation: perceptual test

A perceptual listening test has been designed in order to evaluate the degree of preference over each processing technique (including no processing at all) by human evaluators. In this test, a total of 18 random sentences are selected for each listener, and for each of the sentences two of the three cases (original oesophageal (ORIGINAL), original spectrum + estimated f_0 (F0) and converted spectrum + estimated f_0 (F0 & SPEC)) are presented to the listener. He or she is asked to rate his or her preference over one of the two cases, by giving a score in a five point scale.

A total of 31 native Spanish speakers took part in the test. Figure 3 shows the averaged preference results of comparing the three systems in pairs. As it can be seen, the preferred system is the one with the original spectrum and modified intonation. The system that converts only f_0 is clearly preferred over the system with converted spectrum and f_0. Additionally, the original oesophageal signals are preferred over the signals with the converted spectrum. Thus, the signals with a converted spectrum are least preferred. This can be due to the loss of quality introduced by the conversion of the MCEP coefficients. Figure 4 shows the degree of preference for each pair of systems. The figure shows that the 'strong preference' option is the least chosen option in all three pairs. Also, it can be seen that the f_0-only modification is the most preferred method, followed by no modification (original signals).

A few examples of the converted utterances can be found in the following website: http://aholab.ehu.eus/users/lserrano/ib18/demoib18.html.

4. Conclusions

In this paper we have presented a conversion system from one oesophageal speaker to one healthy speaker using a deep learning approach. We have trained an LSTM network to convert the spectral features of the speaker, and two BLSTMs to estimate the intonation curve: one to learn the underlying relationship existing between the healthy MCEPs and their corresponding $\log f_0$, and another to predict from the same healthy cepstrum the V/U decision. With the system implemented, we have evaluated the effects that the conversion has in comparison with the original pathological signal in terms of intelligibility and pleasantness. The effect of the conversion of the prosody has been evaluated separately from the complete conversion (spectral and f_0 conversion together).

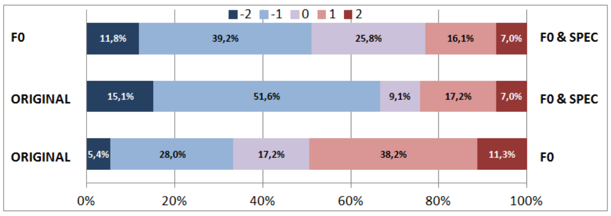

Figure 4: *Detailed results of the preference test. (-2:I strongly prefer sentence 1, -1:I prefer sentence 1, 0:I don't have any preference, 1:I prefer sentence 2, 2: I strongly prefer sentence 2).*

The intelligibility has been evaluated by means of ASR, using the WER. The results show that using a healthy target speaker the recognition rate can be greatly improved (as much as an absolute gain of 15% in our experiment). The WER rate is still far away from that obtained for the healthy voices, but this demonstrates that the conversion of the spectral parameters is helping the oesophageal signal to be more similar to the healthy voices used to train the ASR system.

As expected, f_0-only modification has little or no effect in the recognition rate. However, this modification was the most preferred by the listeners in the perceptual test, even over the original oesophageal voice with no modifications at all. This is an important result because it corroborates the relevance of having a restored source without modifying the speaker characteristics.

Once the deep learning conversion approach has been proved as valid, in the future we will better adjust the LSTM parameters, for example, the sequence size, to capture the prosody of a speaker in a more adequate way. We also plan to experiment different network architectures.

5. 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 European Unions H2020 research and innovation programme under the Marie Curie European Training Network ENRICH (675324).

6. 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] R. Ishaq and B. G. Zapirain, "Esophageal speech enhancement using modified voicing source," in *Signal Processing and Informa-*

tion Technology (ISSPIT), 2013 IEEE International Symposium on. IEEE, 2013, pp. 000 210–000 214.

[5] H. R. Sharifzadeh, I. V. McLoughlin, and F. Ahmadi, "Reconstruction of normal sounding speech for laryngectomy patients through a modified celp codec," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 10, pp. 2448–2458, 2010.

[6] A. del Pozo and S. Young, "Continuous tracheoesophageal speech repair," in *Eusipco*, 2006, pp. 1–5.

[7] ——, "Repairing tracheoesophageal speech duration," in *Speech Prosody*, 2008, pp. 187–190.

[8] O. Schleusing, R. Vetter, P. Renevey, J.-M. Vesin, and V. Schweizer, "Prosodic speech restoration device: Glottal excitation restoration using a multi-resolution approach," in *International Joint Conference on Biomedical Engineering Systems and Technologies*. Springer, 2010, pp. 177–188.

[9] 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.

[10] ——, "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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

[17] L. M. Arslan, "Speaker transformation algorithm using segmental codebooks (stasc) 1," *Speech Communication*, vol. 28, no. 3, pp. 211–226, 1999.

[18] 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.

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

[20] 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.

[21] 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.

[22] 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.

[23] H. Benisty and D. Malah, "Voice conversion using gmm with enhanced global variance," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[24] 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.

[25] 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.

[26] 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.

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

[28] 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.

[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] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech communication*, vol. 11, no. 2-3, pp. 109–118, 1992.

[31] M. Kishimoto, T. Toda, H. Doi, S. Sakti, and S. Nakamura, "Model training using parallel data with mismatched pause positions in statistical esophageal speech enhancement," in *Signal Processing (ICSP), 2012 IEEE 11th International Conference on*, vol. 1. IEEE, 2012, pp. 590–594.

[32] D. Erro, E. Navas, and I. Hernáez, "Iterative MMSE estimation of vocal tract length normalization factors for voice transformation," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[33] D. Erro, A. Alonso, L. Serrano, D. Tavarez, I. Odriozola, X. Sarasola, E. del Blanco, J. Sánchez, I. Saratxaga, E. Navas *et al.*, "Ml parameter generation with a reformulated mge training criterion-participation in the voice conversion challenge 2016." in *Interspeech*, 2016.

[34] J. A. Gonzalez, L. A. Cheah, A. M. Gomez, P. D. Green, J. M. Gilbert, S. R. Ell, R. K. Moore, and E. Holdsworth, "Direct speech reconstruction from articulatory sensor data by machine learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2362–2374, 2017.

[35] 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

[36] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[37] S. P. Rath, D. Povey, K. Veselỳ, and J. Cernockỳ, "Improved feature processing for deep neural networks." in *Interspeech*, 2013, pp. 109–113.

[38] 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.