3rd International Conference on Advanced Technologies
for Signal and Image Processing - ATSIP'2017,
May 22-24, 2017, Fez, Morroco.

A Comparative Study of Voice Conversion Techniques: A review

Kadria Ezzine[1]
ATISP, ENET'COM, Sfax University, Tunisia
[1]ENICARTHAGE, Carthage University, Tunisia
kadria.ezzine@gmail.com

Mondher Frikha
ATISP, ENET'COM, Sfax University, Tunisia
mondher_frikha05@yahoo.fr

***Abstract*—Speaker identity, the sound of a person's voice, is one of the most important characteristics in human communication. Voice conversion (VC) is an emergent problem in voice and speech processing that deals with the process of modifying a speaker's identity. More particularly, the speech signal spoken by the source speaker is modified t o s ound a s i f i t h ad been pronounced by another speaker, referred to as the target speaker. A variety of VC techniques has been proposed since the first appearance of the voice conversion problem. The choice among those techniques represents a compromise between the similarity of the converted voice to the target voice and the quality of the output speech signal,both rated by the used technique. In this paper, we review a comprehensive state-of-the-art of voice conversion techniques while pointing out their advantages and disadvantages. These techniques will be applied in significant and most versatile areas of speech technology; applications that are far beyond speech synthesis.**

Keywords*—Voice Conversion, Speech synthesis, Voice quality, Similarity.

I. Introduction

In daily communication, voice individuality is among the main characteristics of human speech. It is especially important for identifying other person such as in a telephone conversation. Although the advancement of information technology have made speech communication possible in different conditions but we can not deny that there are some instances in which we face with some difficulties p articularly i n remote communication [1]. For instances, we may lose our voice during surgery intervention to take off speech organs such as larynx because of laryngeal cancer or other diseases. Thus, disturbance of voice appears and affects the quality of the voice produced and specifically t he t ransmitted v oice signal. This degrades its intelligibility and makes it more difficult to communicate, exchange ideas and interact in real time. Actually, many problems remain in speech communications. Therefore, the evolution of technology to overcome these inherent problems, has an important role to give individuality to synthesized speech. In fact, in order to enhance the quality of the speech signal, it is necessary to properly transform the acoustic features without changing the linguistic contents.

In the area of speech processing, acoustic features transformation techniques can be used in various applications, such as single-channel enhancement [2], emotional conversion [3], and band width extension of narrowband speech [4], but perhaps the most obvious application is the voice conversion (VC). For the last two decades, a considerable amount of effort has been devoted to the voice conversion problem [5][6][7][8]. These approaches have helped to enhance the human-machine interface to a high degree of authentication. A voice conversion system should be able to identify individual caracteristics of speech spoken by one speaker (source speaker) and substitute them with those of an another specific speaker (taget speaker) without loss of information or modification in the transferred message. Thus, the main objective of voice conversion is to change the voice of a source speaker to get an output (transformed signal) that resembles the voice of a specific target speaker. In order to modify the speech signal, it is necessary to extract some features which characterizes the speech and speaker's identity such as spectral envelope, fundamental frequencies and the closing phase of glottal wave. Some of the much talked about VC applications are the custamization of talking devices, vocal pathology by designing vocal tools to help people who have voice disorders, the dubbing of films and the translation into different languages [9][10]. However the most obvious use case for VC is the synthesis of text-to-speech (TTS) where voice conversion systems can be used to create novel, natural and intelligible voices in a cost-effective way [11][12]. Many voice conversion techniques have been proposed during the last few years , trace out from Codebooks mapping and Vector Quantization [6][5], to Dynamic Frequency Warping [13], towards Hidden Markov Models, Gaussian Mixture Models [14][7][15] until eventually evolved to Artificial Neural Networks [16] and some combinations of them [8]. The goal of this paper is to provide comprehensive state-of-the-art about these different techniques to generate voice conversion. A comparative study between those techniques is conducted while pointing out advantages and disadvantages of each technique.

The paper is organized as follows. Section II, gives a brief overview about the overall architecture of voice conversion system. Section III, describes the evaluation of various voice conversion techniques. Section IV is devoted to a comparative study between the existing VC approaches. Finally, we summarize this paper in the Section V.

II. Overview of the voice conversion system

Voice conversion system is generally performed in two common phases: During the training phase, the system collects voice informations from source and target speakers and automatically formulates voice conversion measures. The conversion phase uses these measures to modify the source voice so as to correspond to the characteristics of the target voice.

978-1-5386-0551-6/17/$31.00 ©2017 IEEE

A typical VC system model is displayed in Figure 1. The training phase ussually includes three steps, acoustic analysis, alignment, and model training. In the analysis step, some specific parameters are extracted from the speech waveform which describe the characteristics of source and target voices (e.g. Mel cepstrum, pitch frequency, glottal wave shape, power values, etc). The following step, alignment, is essential to determine corresponding units in the source and the target voices. Dynamic Time Warping (DTW) [17] and Hidden Markov Models (HMM) [18] are the desirable automatic alignment techniques to use, since the manual alignment takes a lot of time. The final training step, a learning process is applied so as to construct the conversion model which serves to mapping the acoustic space of source speakers to that of the target speakers. During the conversion phase, once the VC model is trained, this latter is tested and deployed taking into account the speech of the source speakers as input. The acoustic parameters are extracted, then the conversion model is applied, finally the converted speech is generated.

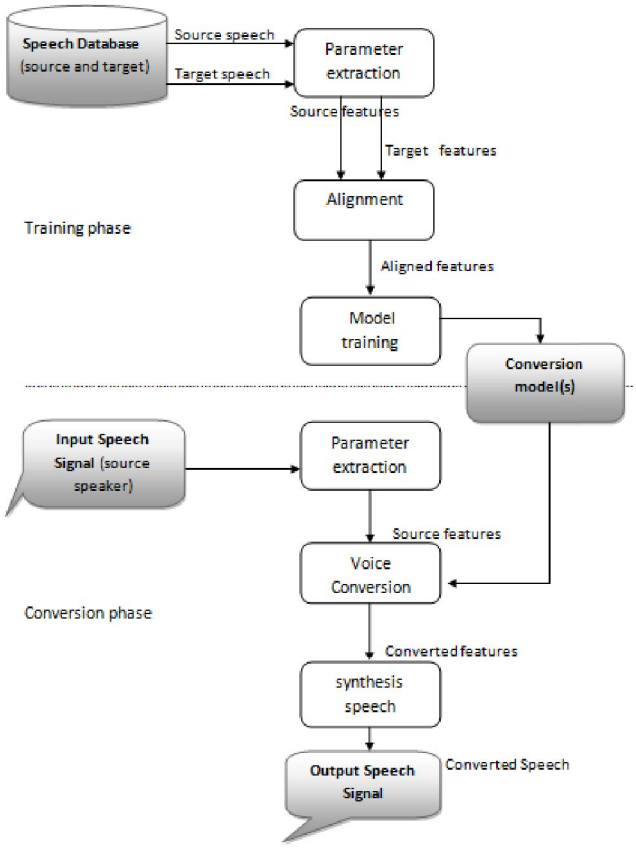

Fig. 1: Global Architecture of Voice Conversion System

III. Voice Conversion Techniques

A. Codebook Mapping and Vector Quantization (VQ)

Since the late 1980s, many approaches in the area of voice conversion have been released [19]. The Codebook Mapping and the Vector Quantisation are the basic VC techniques. Codebook mapping model define automatically the matching acoustic features of source and target speaker and assemble them in the codebooks. Vector quantization method is used for the compression of data that is built from codebooks. A vast number of works are using these techniques, the most recognized ones are presented as follows:

Abe et al [6] propose a codebook mapping technique based on vector quantization and spectrum mapping. The basic idea of their work was to determine the converted feature vector through the quantification of the source characteristic vectors to the closest centroid vectors of source codebooks and replace them with matching centroid vectors of the mapping codebooks. These methods are composed of tow stages: a learning stage and a conversion stage. The first stage involves the production of histograms that show the correspondence between the source and the target vectors which are previously aligned. Then, these histograms are utilized as weighting functions to produce a mapping codebook based on a linear combination. The conversion synthesis stage synthesizes speech and the output speech is achieved using mapping codebooks. Auditory tests prove that the obtained results are near to those of the target speakers while the mapping directory (codebooks) remains fair a linear combination of vectors of a specific speakers. Furthermore, another limitation of the codebook model is the discreet representation of acoustic spaces as a restricted set of spectral envelopes which cause a discontinuity of spectra and degenerate the quality of the converted voice.

To reduce these problems, several methods have been proposed such as trellis structured vector quantization [20], hierarchical codebook mapping [21], local linear transformation [22] and weighted linear combination of codewords [23].

Trellis structured vector quantization proposed in [20], deals with the issue of spectral discontinuity that is the limit of various approaches based on the codebook mapping model. The main objective of this method is to acquire dynamic information and to utilize a trellis structure and a dynamic programing for optimizing the codebook path from the obtained dynamic information. Learning speech quantified in the form of codebook sequences is aligned and subsequently the source and the target codebook are trained. This model works with successive frame blocks. Therefore, the previous codebooks in the source and target sequences are consolidated with each pair of successive codebooks training blocks that makes the speech dynamic [23]. This approach represents an accurate way to solve the problem of spectral discontinuity by the use of dynamic information and simultaneously preserves the benefits of a good conservation of spectral information. However , the calculation of time during process was fairly broad and should be considered a disadvantage.

B. Dynamic Frequency Warping (DFW)

Dynamic frequency warping technique, is considered again as one of the basic method for voice conversion. The purpose of the DFW model is to create a warping function among the source and target spectrum to synthesize the output speech. Proposals founded on Dynamic frequency warping technique are introduced as follows.

Shuang et al. in [24] proposed a novel approach to generating frequency warping function through mapping formants of the source and target speakers of each aligned frame. Dynamic time warping algorithm is used in the alignement process. This

technique directly modifies the frequency scale of the source spectrum in accordance with a mapping function previously formed for its matching acoustic class by maintaining a normal output speech. The benefit of this approach is that it only needs a small amount of training data and provides good quality of output speech. Although, the spectral form was not been totally altered due to the formants that are absolutely modified to another frequency without changing their related intensity. Therefore, the similarity scores between converted voice and target voice are nevertheless disadvantageous.

Weighted Frequency Warping technique [25], is proposed for voice conversion to improve the speaker independent recognition. This method corresponds to a temporal frequency warping function that maps the frequencies of source speakers to those of the target speakers. In the learning stage, the common acoustic space of the speakers is divided into overlapping classes by the use of gaussien mixture model, then a warping function is formed for both the source and the target classes from spectral envelopes. In the conversion stage, a warping function resulting from the weighted combination of the qualified functions is applied to the original spectrum. Then, GMM-based conventional statistical methods (consisting of a weighted combination of linear transformations) are applied to correct the energy of the warped spectrum. This method achieves a good speech quality scores which allows it to be exploited in real applications, while the conversion scores remained virtually unchanged.

C. Hidden Markovian Models (HMM)

Hidden Markov models (HMMs) represent in recent years a new voice conversion branch which is used for speech synthesis. This model is considered as a Markovian process with many hidden states that are built by a collection of rules. The target spectrum is obtained by designing these states between source and target speech. Existing works about this technique are discussed below:

Kim et al. in [26] proposed a HMM as a voice conversion model using dynamic characterstics of speaker. The main purpose of this approach is the use of state transition probability as dynamic features of speaker where each state has a conversion rule . Reasonably, each state has its own conversion rule being the HMM is a set of acoustically similar feature parameters. For an intelligible conversion process, this method performs two codebook which have a one-to-one mapping links as recognition-codebook and synthesis-codebook. The HMM structure is a class of conversion rules with stochastic transitions. This last defined as "conv-HMM" which works to find a relationship between the parameters of two speakers and generates conversion rules on each state. This technique provides a performance that is higher than a classical codebook model based on VQ.

The work of [27] describes an inversion system based on HMM which collects the articulatory motions from acoustic speech . So it becomes valuable for the synthesis of speech. This method uses a trajectory HMMs a s a common generative model for articulatory acoustic data which is usable for the synthesis and the recognition of speech signals. This technique denotes that the acoustico-articulatory models trained jointly are more precise than those trained separately. Although there are important techniques based on trajectory HMM training, many advantages encourage to use of the latter as a unified model to synthesize a high quality of the speech signal.

D. Gaussian Mixture Models (GMM)

The Gaussian Mixture Model (GMM) is the most popular voice conversion approach. In this method, the speaker's acoustic space is modeled through the GMM model without using VQ and the acoustic characteristics are transferred from the source speakers to the target speakers via a mapping function which is a weighted sum of the local regression functions. Some works of literature are presented as follows:

A GMMs based voice conversion model has been proposed by Stylianou et al. in [7] in order to estimate probability density of the source spectrum vectors. They assumes that the available data comprise two set of paired spectral vectors as $X_t = [x_1 x_2 ..... x_t]$ and $Y_t = [y_1 y_2 ..... y_t]$ which have the same length t which respectively describe the spectral envelopes of source and target speakers. This approach aims to find the function $F()$ so that the transmitted envelope $F(x_t)$ which more corresponds to the target envelopes of the training set. The conversion function is determined through the minimization of the sum of square conversion errors between the converted and target envelopes given as:

$$e = \sum_{i=1}^{T} \|y_t - F(x_t)\|^2 . \tag{1}$$

Where T represents the entire sum of paired learning vector (x_t, y_t).

The conversion function is usually supposed linear in each gaussian, it is given by :

$$F(x_t) = \sum_{n=1}^{M} w_{n,t}(\beta_n x_t + b_n). \tag{2}$$

Noting that β_n as linear transformation matrix for samples of class n, b_n the bias vector. Observation weights $w_{n,t}$ are posterior probabilities which have expressed as:

$$w_{n,t} = \frac{\alpha_n N(x_t; \mu_n, \sum_n)}{\sum_{m=1}^{M} \alpha_m N(x_t; \mu_m, \sum_m)}. \tag{3}$$

Where α_n is a prior probability of Gaussian $n = 1...M$ and $N(x_t; \mu_n, \sum_n)$ represents the normal distribution where μ_n is a mean vector and $\sum_n$ a covariance matrix .

This method is trained with the Expectation-Maximization (EM) algorithm by the use of dataset that are aligned and quantized automatically via dynamic time warping algorithm (DTW). This technique proves that it is capable to producing good results in the noisy environment, however, this implies a large amount of calculations within the spectral analysis as it is necessary to calculate and maximize the joint probability. In addition, this technique is unable to convert the nonlinear speech caracteristics.

In his work [28], Kain lightly consolidated this method. In this work, a VC approach is proposed to generate a detailed

description of the converted speech spectrum. In this method, a Gaussian mixture model is estimated on the conjoint density of source and target caracteristics, then a consecutive regression produces the ending conversion function. The modeling of the conjoint density instead of solely the source density may lead to a more wise distribution of the mixing componnents as well as avoid some digital problems during the invertion of large and possibly malformed matrices. This method of mapping is properly effective. However, the voice conversion performance is constantly insufficient.

A year later, Toda [8] proposes a new voice conversion model called STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum). This approach combines GMM model and Dynamic Frequency Warping (DFW) that allow to manipulate spectral, acoustic and rhythmic paramters. This method is yet fragile being the quality of the converted voice is weakened because of some factors, for example, spectral motion with unsuitable dynamic features produced per the frame-by-frame conversion mode and the excessive smoothing of the converted spectrum.

To address these problems, Toda et al. [29] proposed a spectral conversion based on the maximum likelihood estimation (MLE) of GMM-based spectral parameters trajectory in the same way of HMM-based speech synthesis [30]. This method performs the estimation of a relevant spectral sequence considering not only static but also dynamic features which are taken into account in mapping function. Moreover, the over-smoothing problem is avoided by using a feature known by the global variance (GV) of converted spectra. This technique has the ability to produce more smooth and more precise parameter trajectories compared to methods based on classical GMMs. Nevertheless, this latter still suffers from the limitations such that it does not employ all linguistic knowlledge on the acoustic data.

E. Artificial Neural Network (ANN)

Acctually, the models based on ANN are recently more used in VC system. These models are well known for modeling complex (nonlinear) relations between source and target speakers. Furthermore, components of the field of learning machines like Boltzman machine and neural network are used by these models for training and testing the (VC) model. Some of current works based on ANN techniques are described as follows:

Desai et al. in [31] proposed an approach based ANNs using a parallel data set supported by sources and targets speakers for regularly extracting the pertinent training data in order to map the spectral caracteristiques of source speaker's over the acoustical space of target speaker's.

In his work [32], Desai exploited the mapping expertise of ANN and showed that this latter could be used for the spectral processing of a continuous speech signal by the use of Mel-cepstrum coefficients (MCEPs) as characteristics.

Takashima et al. [33], proposed a method of converting voice based on exemplars (dictionary) for noisy signals. This method uses parallel dictionaries composed by source and target instances having identical texts spoken by source and target speakers. The speech conversion process proceeded as follows: The input signal is defined by a dispersed representation of the source and noise exemplars obtained from input signals and their corresponding weights. Then, the construction of dictionary includes the use of the STRAIGHT model and a magnitude spectrum called STFT for the source speaker, since this latter exists in a noisy environment. The main objective of the use of STFT is to extract noise exemplars and to estimate the weights. Finally, using these weights, the converted speech is generated from the target exemplars . This technique shows a better performance in normal and noisy speech than a conventional GMM-based method but the constraint was that parallel exemplary measurement construction were boring.

In recent years, Deep Neural Networks models (DNNs) were used in voice conversion. The approaches using deep models have the capacity to learn deeper architecture and to prove performance in the areas of speech synthesis and speech recognition [34][35].

A new voice conversion technique using DNN in super-frame feature space is proposed in [36]. The main objective of the work is the mapping of the spectral envelopes of the source and target speakers, in which the neighbor frame's influence is considered. Then the powerful ability of this method which has a five-layers architecture made up three restricted Boltzmann machines (RBMs) was exploited to drive the spectral conversion function. Dynamic Time Warping (DTW) is used to extract linear prediction cepstrum coefficients (LPCC) from source and target voices. This technique adopted the DNN as a regression model to predict LPCC parameters of target speakers. Generally, the DNN models training consist of two steps: pre-training step where each pair of layers is traited like a restrictive Boltzman machines (RBM) when the parameters are formed layer by layer and supervised fine-tuning step which aim to minimising the mean squared error between DNN output and LPCC caracteristics of target speakers as defined below:

$$E = \frac{1}{N}\sum_{n=1}^{N}\left\|\hat{S}_n^t(S_{n\pm\tau}^s, W, b) - S_n^t\right\|^2 + \lambda\left\|W\right\|^2. \quad (4)$$

Where $\hat{S}_n^t$ and S_n^t are the n^{th} D-dimensional LPCC vectors of respectively output and target speaker. $(S_{n\pm\tau}^s)$ is a D$(2\tau+1)$-dimensional LPCC vectors of input features with neighbor right and left τ frames. The weight and the bias parameters noted respectively by W and b, λ represent the weighting coefficient of regularization in order to eliminate the overfitting. This technique based on DNN shows that the synthesized speech quality and speaker identity are better than the GMM-based technique [36].

In 2016, Nakashika et al. [37], proposed a voice conversion method that does not use any parallel data during the formation of the model, as no parallel data is needed in the training phase. This method considers that the speech signals are generated from a probabilistic model based on a Restrictive Boltzmann Machines whither phonological and speaker informations are explicitly determined. In the training phase, the parameters of speaker independent and speaker dependent are formed at the same time under adaptive training of speaker. In the conversion phase, a speech signal is devided into both the phonologic and

the speakers information. The source speakers informations are modified by those of the target speakers, then, from the mixture of the latter, the transmitted output speech is acquired. Experiments were proved that this model performs better than another non-parallel model [38][39], and give results close to these of the standard parallel-training approach.

IV. Comparative study

Table I provides a summary of the voice conversion techniques, while highlighting their advantages and disadvantages.

TABLE I: Summary of different VC techniques

Voice Conversion Method	Advantages	Disadvantages
Codebook Mapping and Vector Quantization	Produces a focused speech founded only on mapping and mathematical calculations.	This method provide only a one-to-one mapping of codewords. Temporal discontinuities to the converted speech.
Dynamic Frequency Warping	This approach needs a very little quantity of training data.	Focus only on specific parts of spectrum. The formants of the source and the target spectrum were absolutely modified to novel frequency without changing their related intensity.
Hidden Markovian Models	Provided a flexible framework for voice synthesis, whither every characteristics of speech can be modeled at the same time into identical multi-stream HMM.	This method is limited by statistical over-smoothing and it is also rules-based although it added a few dynamic features.
Gaussian Mixture Models	Offers a good similarity between transformed and target voices without presenting discontinuities within the converted speech.	The well known disadvantages of this method are over-smoothing and over-fitting. As well, this method provides a degraded target speech when there is noise.
Neural Network Models	This method is able to work properly in noisy environments, provides the ability to model non-linearities and dynamics and gives results that are highly close to the source signal.	The good performance and the accuracy of the synthesis were high only when the architectures were deeper. So it is necessary to construct several layers through which the desired output can be obtained.

A wide variety of works in littrature, conduct comparative studies to evaluate how a VC method performs. The performance of VC systems is generally evaluated according to the quality of the output speech and the similarity between the converted voice and the target voice [25]. The results of voice conversion are evaluated by many perception tests (per performing listening tests). For instance, ABX test is used like a measure to evaluate in a subjective way the performance of the VC systems. This is a bidirectional test which is frequently used to compare the similarity between converted and target speech.

Table II describes a set of ABX results of some VC systems that showing how the ABX index changes in each of them as mentioned by the authors [45]. Among these, a method proposed by Ye et Yu [36] stands out with extremely high scores. This result makes it significantly apparent that ANN-based VC method, provides converted voice with more authenticated spectral similarity with the desired target voice compared to the other voice conversion methods. In addition, a comparative study between this method based on Deep Neural Netwok (DNN) and the conventional GMM is conducted. Experiments have shown that this technique provides better accuracy, since the speaker identification rate of conversion speech achieves 97.5% which is 0.8% higher than the performance of GMM method and the value of average cepstrum distortion is 0.87% which is 5.4% higher than the performance of GMM method.

TABLE II: Experimental Results for ABX indices in VC systems.

Year	Author	ABX Index	Technique
1998	Stylianou [7]	97%	GMM
2001	Toda [8]	77% 83%	GMM DFW
2002	Arslan [9]	91.1%	Codebooks Mapping
2005	Toda [40]	84%	MLE
2005	Zhang [24]	87.5%	GMM
2006	Ye et Young [41]	91.8%	GMM
2006	Shuang [24]	64.3%	DFW
2007	Toda [29]	95%	GMM+MLE
2008	Yue [42]	92.0%	GMM+HMM
2009	Zhang [33]	68%	VQ
2010	Desai [32]	95%	ANN
2010	Helander [43]	84%	GMM
2012	Laskar [44]	98% 99%	GMM ANN
2015	Ye et Yu [36]	100%	ANN

Furthermore, more recent approaches in [32]-[36] also carry out a comparative survey of VC techniques using an ANN model and the conventional GMM model. The results obtained from these studies confirm that an ANN-based VC system is more efficient than a GMM-based VC system. Moreover, the converted speech has a very intelligible quality and has the characteristics of a target speaker.

V. Conclusion

Voice conversion system plays an important role in improving the quality of the human voice. Over the last few decades, considerable research has been focused on VC and many substantial approaches have been proposed. This paper aims at presenting a survey including the best existing works in order to understand the development of the voice conversion system by many raised techniques as well as advantages and disadvantages of each technique. The perceptual tests shown that techniques based on neural network provide better output quality even in the presence of noise once the number of layers increases in the neural network. More precisely, to obtain greater synthesis accuracy we should essentially get a deeper non-linear architecture with lots of hidden layers. Although the fact that conventional VC techniques gives very profitable results, it is still necessary to process additional research advances in order to progress more on how to provide excellent speech quality and highly successful speech identity in the same time.

REFERENCES

[1] T. Toda, M. Nakagiri, K. Shikano, "Statistical voice conversion techniques for body-conducted unvoiced speech enhancement," IEEE Transactions on Audio, Speech, and Language Processing, 20(9), 2505-2517, 2012.

[2] A. Mouchtaris, J. Van der Spiegel, P. Mueller, and P. Tsakalides, "A spectral conversion approach to single-channel speech enhancement," IEEE Trans. Audio, Speech, Lang. Process., vol. 15, no. 4,pp.1180-1193, May 2007.

[3] C. Hsia, C. H. Wu, and J. Q. Wu, "Conversion function clustering and selection using linguistic and spectral information for emotional voice conversion," IEEE Trans. Computers, vol. 56, no. 9, pp. 1245-1254, Sep 2007.

[4] K. Y. Park and H. S. Kim, "Narrowband to wideband conversion of speech using GMM-based transformation," in Proc. ICASSP, vol.3, pp. 1843-1846, vol.3, 2000.

[5] E. Helander and J. Nurminen, "A novel method for prosody prediction in voice conversion," in ICASSP, vol. 4, pp. 509-512, 2007.

[6] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in ICASSP, pp. 655-658, 1988.

[7] Y. Stylianou, "Continuous probabilistic transform for voice conversion," IEEE TSAP, no. 6, pp. 131-142, 1998.

[8] T. Toda, H. Saruwatari, and K. Shikano, "Voice Conversion Algorithm based on Gaussian Mixture Model with Dynamic Frequency Warping of STRAIGHT spectrum," In Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pages 841-844, Salt Lake City, USA, 2001.

[9] O. Trk and L. M. Arslan, "Subband based voice conversion," in Proc. ICSLP, Denver, CO, vol. 1, pp. 289-292, Sep 2002.

[10] O. Trk, O. Byk, A. Haznedaroglu, and L. M. Arslan, "Application of voice conversion for cross-language rap singing transformation," in Proc. IEEE ICASSP, Taipei, Taiwan, Apr 2009.

[11] A. Kain, M. Macon, W. , "Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction." In: Proceedings of IEEE International Conference of Acoustics, Speech, Signal Processing, vol. 2. pp. 813-816, 2001.

[12] D. Sundermann, "Voice conversion: State-of-the-art and future work," FORTSCHRITTE DER AKUSTIK, vol. 31, p. 735, 2005.

[13] H. Mizuno, and M. Abe, "Voice conversion algorithm based on piecewise linear conversion rules of formant frequency and spectrum tilt", Speech Communication, vol. 16, pp. 153-164, 1995.

[14] A. Kain and M. Macon, "Spectral voice conversion for Text-to-Speech synthesis", in Proc. of ICASSP, pp. 285-299, 1998.

[15] M. Mashimo, T. Toda, H. Kawanami, H. Kashioka, K. Shikano, and N. Campbell, "Evaluation of cross-language voice conversion based on GMM and STRAIGHT", in Proc. of Eurospeech, pp. 361-364, 2001.

[16] M. Narendranath, H. Murthy, S. Rajendran, and B. Yegnanarayana, "Transformation of formants for voice conversion using artificial neural networks", Speech Communication, vol. 16, pp. 207-216, 1995.

[17] F. Itakura, "Minimum prediction residual principle applied to speech recognition". IEEE Trans. Acoust. Speech Signal Process. ASSP-23 (1), 67-72, 1975b.

[18] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition". Proc. IEEE 77 (2), 257-286, 1989.

[19] D. G. Childers, B. Yegnanarayana, and K. Wu, "Voice conversion: Factors responsible for quality," ICASSP, pp. 748-751, 1985.

[20] M. Eslami, H. Sheikhzadeh, A. Sayadiyan, "Quality improvement of voice conversion systems based on trellis structured vector quantization," Proc. of Interspeech, pp. 665-668, 2011.

[21] Y. Wang, Z. Ling, R. Wang, "Emotional speech synthesis based on improved codebook mapping voice conversion," Proc. of ACII, pp. 374-381, 2005.

[22] V. Popa, J. Nurminen, G. Moncef, "A study of bilinear models in voice conversion," Journal of Signal and Information Processing 2(2): 125-139, 2011.

[23] E. Helander, M. Gabbouj, Jani Nurminen1, Hanna Siln2, Victor Popa2, "Speech enhancement, modeling and recognition algorithms and applications", 69, 2012.

[24] Z. W. Shuang, R. Bakis, S. Shechtman, D. Chazan, and Y. Qin, "Frequency warping based on mapping formant parameters," in Proc. Int.Conf. Spoken Lang. Process., 2006.

[25] D. Erro, A. Moreno, A. Bonafonte, " Voice conversion based on weighted frequency warping" IEEE Transactions on Audio, Speech, and Language Processing, 18(5), 922-931, 2010.

[26] E. Kim, S. Lee, and Y. Oh, "Hidden markov model based voice conversion using dynamic characteristics of speaker," in 5th ECSCT, 1997.

[27] L. Zhang, S. Renals, "Acoustic-articulatory modelling with the trajectory HMM," IEEE Signal Process. Lett. 15, 245-248, 2008.

[28] A. Kain, "High resolution voice transformation," Ph.D. dissertation, OGI School of Sci. and Eng., Beaverton, OR, 2001.

[29] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," IEEE Trans. Audio, Speech, Lang. Process., vol. 15, no. 8, pp. 2222-2235, Nov 2007.

[30] K. Tokuda, H. Zen, and A. W. Black, "An HMM-based speech synthesis system applied to English," in Proc. IEEE Workshop Speech Synth., Santa Monica, CA, pp. 227-230, Sep 2002.

[31] S. Desai, E. Raghavendra, B.Yegnanarayana, A. Black, K. Prahallad, "Voice conversion using artificial neural networks," In Acoustics, Speech and Signal Processing, ICASSP 2009. IEEE International Conference on (pp. 3893-3896), April 2009.

[32] S. Desai, B.Yegnanarayana, A. Black, K. Prahallad, "Spectral mapping using artificial neural networks for voice conversion," Audio, Speech, and Language Processing, IEEE Transactions on, 18(5), 954-964, 2010.

[33] R. Takashima, T. Takiguchi, and Y. Ariki, "Exemplar-based voice conversion in noisy environment," in Proc. IEEE Spoken Lang. Technol. Workshop (SLT), pp. 313-317, 2012.

[34] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, et al, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," Signal Processing Magazine, IEEE, 29(6):82-97, 2012.

[35] H. Lu, S. King, and O. Watts, "Combining a vector space representation of linguistic context with a deep neural network for text-to-speech synthesis," In 8th ISCA Workshop on Speech Syn- thesis, pages 281-285, Barcelona, Spain, August 2013.

[36] W. Ye, Y. Yu, "Voice conversion using deep neural network in superframe feature space," In Intelligent Control and Information Processing (ICICIP), 2015 Sixth International Conference on (pp. 465-468) IEEE, November 2015.

[37] T. Nakashika, T. Takiguchi, Y. Minami, "Non-Parallel Training in Voice Conversion Using an Adaptive Restricted Boltzmann Machine". IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24(11), 2032-2045, 2016.

[38] H. Sile'n, J. Nurminen, E. Helander, and M. Gabbouj, "Voice conversion for non-parallel datasets using dynamic kernel partial least squares regression," in Proc. Interspeech, 2013.

[39] D. Erro, A. Moreno, and A. Bonafonte, "INCA algorithm for training voice conversion systems from nonparallel corpora," IEEE Trans. Audio, Speech, Lang. Process., vol. 18, no. 5, pp. 944-953, Jul 2010.

[40] T. Toda, A. Black, and K. Tokuda, "Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter," in ICASSP, pp. 9-12, 2005.

[41] H. Ye and S. Young, "Quality-enhanced voice morphing using maximum likelihood transformations," IEEE TASLP, vol. 14, no. 4, pp. 1301-1312, 2006.

[42] Z. Yue, X. Zou, Y. Jia, and H. Wang, "Voice conversion using HMM combined with GMM," in CISP'08, vol. 5, 2008.

[43] E. Helander, T. Virtanen, J. Nurminen, M. Gabbouj, "Voice conversion using partial least squares regression," IEEE Transactions on Audio, Speech, and Language Processing, 18(5), 912-921, 2010.

[44] R. H. Laskar, D. Chakrabarty, F. A. Talukdar, K. S. Rao, K. Banerjee, "Comparing ANN and GMM in a voice conversion framework," Applied Soft Computing, 12(11), 3332-3342, 2012.

[45] A. F. Machado, M. Queiroz, "Voice conversion: A critical survey," Proc. Sound and Music Computing (SMC), p. 1-8, 2010.