International Journal of Electrical, Electronic and Communication Sciences
ISSN: 2517-9438
Vol:3, No:12, 2009

Transformation of Vocal Characteristics: A Review of Literature

Dong-Yan Huang, Ee Ping Ong, Susanto Rahardja, Minghui Dong and Haizhou Li

***Abstract*—**The transformation of vocal characteristics aims at modifying voice such that the intelligibility of aphonic voice is increased or the voice characteristics of a speaker (source speaker) to be perceived as if another speaker (target speaker) had uttered it. In this paper, the current state-of-the-art voice characteristics transformation methodology is reviewed. Special emphasis is placed on voice transformation methodology and issues for improving the transformed speech quality in intelligibility and naturalness are discussed. In particular, it is suggested to use the modulation theory of speech as a base for research on high quality voice transformation. This approach allows one to separate linguistic, expressive, organic and perspective information of speech, based on an analysis of how they are fused when speech is produced. Therefore, this theory provides the fundamentals not only for manipulating non-linguistic, extra-/paralinguistic and intra-linguistic variables for voice transformation, but also for paving the way for easily transposing the existing voice transformation methods to emotion-related voice quality transformation and speaking style transformation. From the perspectives of human speech production and perception, the popular voice transformation techniques are described and classified them based on the underlying principles either from the speech production or perception mechanisms or from both. In addition, the advantages and limitations of voice transformation techniques and the experimental manipulation of vocal cues are discussed through examples from past and present research. Finally, a conclusion and road map are pointed out for more natural voice transformation algorithms in the future.

***Keywords*—**Voice transformation, Voice Quality, Emotion, Individuality, Speaking Style, Speech Production, Speech Perception.

I. INTRODUCTION

TRASFORMATION of vocal characteristics can change a voice either in the intelligibility of aphonic speech or into a voice so that it sounds like the voice of another speaker. The former application of transformation of vocal characteristics can be found in high-end hearing aids, vocal pathology and voice restoration. The latter situation is usually referred to as voice modification, when not a specific target is provided. To the contrary, if a target is specific, a voice modification is referred to as voice conversion or voice mimic. Voice morphing, which we often encounter in computer games or speech software show, is another type of voice modification. However, it is different from previous two cases.

D.-Y. Huang, E.P. Ong, S. Rahardja, M. Dong and H. Li are with the Institute for Infocomm Research, Agency for Science, Technology and Research, 1 Fusionopolis Way, #21-01 Connexis, South Tower, Singapore 138632 (corresponding author to provide phone: 65-6408-2639; fax: 65-6776-1378; e-mail: {huang, epong, rsusanto, mhdong, hli@ i2r.a-star.edu.sg}).

It creates a third voice for a specific sentence from two speakers (source speakers), who utter the same sentence. This third voice can be continuously changed from one speaker to another speaker. Since 1990's, transformation of vocal characteristics has been a hot, novel and fast growing topic.

Especially, with the recent developments in high quality TTS system by the employment of large databases and unit techniques, there is an increasing demand for high quality voice transformation methods not only for creating target or virtual voices as voice conversion requires less training data, but also for increasing the intelligibility of speech in noisy environment, synthesize emotions and speaking style, to make more natural and emotional dialog systems which use speech synthesis etc. Moreover, voice transformation can be applied to areas like language tutor, singing voice transformation, dubbing movies, and music industry, toys, chat rooms and games, security and speaker individuality for interpreting telephone. Integrated with 3-D facial animation techniques, voice conversion can be applied in creating newsreader, virtual story-teller, virtual assistant for e-commerce, e-learning, or e-care. As high quality voice transformation is so high demand for these applications, this paper attempts to review literature on voice transformation methodology and to make a road map for some of promising approaches for interdisciplinary research in this area.

The purpose of voice transformation is to control non-linguistic information of speech signals such as voice quality, voice expressivity and voice individuality. However, how to control this information? What are the phenomena relevant to voice quality, voice expressivity and voice individuality during sound production? What acoustic cues are relevant to different types of information and can be modified? How to separate the linguistic from the expressive and individual information? It is obvious that high quality voice transformation system design depends on better understanding of the processes and mechanisms of human speech production and perception, as well as their relationship at brain processing level. The realization of these processes and mechanisms by machine should be relevant to variety of speech research areas such as acoustical modeling, perception, recognition, synthesis, coding, linguistics etc. All these fields share many common methods and approaches. Voice transformation can serve as a connection point between these major fields. However, voice transformation goes beyond these technologies. For example, the mechanisms of human speech production and perception are efficiently used in speech coding systems and synthesis systems (e.g., LPC coding system and STRAIGHT synthesis system). Voice

transformation systems require the same process. Furthermore, the parameters of speech model should be changed with respect to the production of speech while keeping the naturalness of speech signal in perception level. The speech coding system aims at "reproducing" the speech, while voice transformation is itself modification. The challenges for modifying above process are how to make the process to be understood and how to make the modifications of parameters of speech model in a way that the modified/transformed speech sounds intelligible and natural. The relationship between technologies used in speaker verification/recognition and voice transformation is that there are the common research interests: identification, representation and detection of acoustical cues relevant to voice individuality, however, voice transformation needs further to modify these acoustical cues such that the modified/transformed speech sounds natural. Therefore, for voice transformation, the selection of speech models should be based on whether it is able to efficiently represent these cues, modify and synthesize them to generate natural speech signal.

Research and design of high quality of voice conversion algorithms should be based on a theory. In this paper, the modulation theory of speech is suggested to be a guideline for modeling and modifying speaking style of linguistic, expressive and individual information of speech under one framework. The modulation theory concerns various types of information of speech including phonetic quality, affective, personal, and transmittal. It is different from previous theories of speech perception, which mainly concern with problems relating to phonetic quality alone.

Firstly, the modulation theory of speech can provide an adequate speech model for voice transformation and explanation for each above question [1]. This theory attempts to let us know the interplay between the linguistic and the non-linguistic aspects of human speech production and perception [1] [2]. This speech model allows us to separate the linguistic from the expressive and organic information, and take into account several factors, which are usually ignored in other speech research areas including the nonlinear and time-varying of speech, the interaction between vocal tract and source characteristics, and the modulation phenomena related closely to the speech production process. Less-natural transformed speech with a non-human producing quality (e.g., cartoon character) may produce if ignoring these factors in voice transformation. It shows also how to manipulate them for high quality voice transformation. The high quality voice transformation is desired to be capable of transforming accent, speech style of linguistic information, emotion and attitude information, and personal information.

Many studies have been carried out on voice transformation for the last two decades. Current voice transformation methods focus mainly on spectral transformations for voice conversion and prosody transformations including time-scaling (duration modification), pitch modification, energy modification. Among these methods, only time-scaling methods are quit successful for voice transformation, while pitch modification methods should be further improved in quality of transformed speech. The quality of current voice conversion algorithms to map the source and the target spectral envelope for transforming speaker identity is very poor. The current of speaking style technologies normally work at frame level, as in most of research areas, i.e., speech recognition, speech enhancement, speech synthesis, etc. It suggests converting the speaking style from one person to another at higher level. An excellent survey on voice conversion can be found in [3], which pointed out the limitations of the current voice conversion algorithms from our perceptual experience of speech.

The purpose of this paper tries to present the problems of voice conversion based on the modulation of theory and reveal what should be modified for accent, speaking style, emotions and personal transformation, and how to change by the state-of-the-art signal processing for improving voice quality of current voice transformation techniques, how to transpose the current techniques of voice conversion to the domain of emotion voice transformation. In order to improve the quality of the current voice conversion algorithms, we review the state-of-the-art voice conversion techniques and classify them based on the underlying scientific principles either from the speech production or perception mechanism or from both from perspectives of human speech production and perception. Furthermore, we discuss the advantages and disadvantages of algorithms, and several factors affecting the quality of voice conversion algorithms, as well as the training and modifications of acoustical features based on the modulation theory.

The paper is organized as follows. Section II presents modulation theory and explains what can be modified in the speech signal and how for voice transformation. Section III reviews the current state-of-the-art methods for voice transformation and discusses the advantages and disadvantages of the methods under the framework of modulation theory. Section IV discusses the issues for future research efforts to improve the voice quality of transformed speech signals.

II. The Modulation Theory of Speech

According to the Modulation Theory (MT) [1], speech is seen as a biological innovation founded on a facility of expressive communication and will continuously play an important role in human communication. It is founded on an analysis of how the different kinds of information are fused in speech production. We briefly outline this theoretical model: 1) Speech signals are regarded as the results of a process that a carrier signal (the speaker's voice), with properties given by organic (personal, individual) and expressive factors, has been modulated with conventional linguistic speech. 2) For each speech sound in a given context, the acoustic properties of speech signals are specific and deviate from those of an neutral carrier signal. 3) The listeners demodulate the signals to separate the different types of information. to tune in to the carrier (the voice) on the basis of an analysis of a stretch of speech and to evaluate its modulation.

The theory shows that a linguistically neutral carrier signal is defined as a "colorless' vowel, a primitive human vocalization. Its properties are determined by the size of

speaker's organ of speech (vocal fold mass and length, vocal tract length, etc.) and by its paralinguistic "setting". For the second point, the theory explains that it is important to measure each type of deviations with a proper rule. For example, the deviation of pitch from its base value is measured on the logarithm of frequency, the formant frequencies on a bark scale, the intensity differences on a dB-scale. Finally, the theory explains the third point that listener is able to separate modulation and carrier and determine each by its own. They have to discover how the carrier signal has been modulated in order to recognize the conventional linguistic information.

According to the theory, for obtaining the individual and expressive information from the speaking speech, the listeners only evaluate the deviations of the current properties of speech signal (F0, formant frequencies, lip shape, etc.) from those they expect of a linguistically neutral sound with the same voice quality based on their assumptions on both intrinsic and extrinsic properties. The intrinsic refers to the listeners experience, e.g., they are familiar with the speaker or already have heard a speaker say some words. Otherwise, certain intrinsic properties of sonorant, such as the frequency positions of F3 and the higher formants provide cues for an appropriate demodulation.

The evidence of demodulation speech in speech perception shows that the accent and speaking style information are characterized by prosodic patterns, emotion and attitude information are control by vocal effort, speech rate, …etc, gender and age information are control by Larynx size, vocal tract length,… etc [1]. Many investigations are conducted to find the vocal correlates of emotions, attitude, gender, age, accent and speaking styles to validate the model based on MT [4]-[9]. The speaking style is a key feature of individuality and correlates closely to the F0 and its contour, the first two formant frequencies and speech rate. There are more complicates for expressive variation because not only the carrier but also amplitude and rate of the modulations are affected. Therefore, in models of speech production and perception based on MT, this can be handled by automatic gain control and speech rate control when listeners acquire the speech. The MT has nothing to say co-articulation, but it accommodates the extra- and paralinguistic aspects of human speech as well as its intra-linguistic aspects. The combinatorial of properties of speech sounds give the effects of co-articulation.

The MT explains also the process how an infant says its first word and imitates something an old person has said, by recognizing how a speaker had modulated his voice, storing a representation of modulation in memory, and modulating its own voice in the same way, but not requiring perturbing the vocal tracts in exactly the same way. It is quit useful in answering the questions what we can transform and how in Introduction. In parallel, some studies on the professional impersonators allow us to know how good the impersonator can imitate the target [10]-[12]. Although such studies are limited and there are different claims for the capacity of the mimicked voice to change formant frequencies of the speech style during the voice imitation, in rhythm, the intonation, the prosody and stressed words and phrases as close as to a given the target because of different data set used in the studies, all these studies show that the paralinguistic part is very important from perception view point.

Although control of articulators provide ways to control the quality, it is reported in literature that compared with the production characteristics, the control of nonlinguistic or paralinguistic characteristics is equally, or in some cases even more, important from a perception point of view for a high quality transformed/converted voice result [13].

Voice transformation algorithm does the similar process and imitates something a person has said. The difference is that the former is done by human, and the latter is done by machine. It can be designed based on MT. The speech model is still built upon analysis-by-synthesis system. The system demodulates the speech signal and separates the signal into a linguistically neutral carrier signal and different types of information (e.g., accent, speaking style, emotions, gender, age, etc.). This process is analysis part of system (analogous to the demodulation of listeners on MT, also to analysis part in speech encoding). The linguistically neutral carrier signal can be obtained through training database by machine learning methods, similar to the listeners experience or intrinsic properties of sonorant. The system modulates the voice with the neutral carrier signal by using the individual and expressive information (analogous to the modulation of speakers on MT, also to synthesis part in speech decoding). It is obvious that the speech model based on MT takes into account the nonlinear nature of speech, the interaction of vocal tract and source characteristics, and modulation phenomena observed in speech signals closely connected with speech production.

Based on MT, direct and indirect ways can be followed to design high quality voice conversion algorithm. Either we explore nonlinear speech models/speech modulation models for voice conversion and it is still on the way to go, or we integrate these effects of paralinguistic factors in the current voice conversion systems. The indirect way is simple and requires us to chart some of the promising approaches for integrating the effects of paralinguistic factors to improve the quality of transformed speech and to be transposed for emotion-related transformation.

III. Literature Review

This section will review available voice modeling and transformation techniques. As the database collection is very expensive, the available different speaker's gender, age, speaker's accent, emotional speaking styles database are sparse to be used to define speech synthesis voices. That leads to the idea to introduce some parametric modification capabilities at the level of the synthesis system to compensate for the high cost database collection . The modification of different attributes of speech can be conducted under the framework of above MT.

In order to guarantee the voice quality and naturalness of modified speech, speaker's individuality-related parametric and emotion-related parametric modifications will be focused on the domains of voice quality and prosody, as suggested by

several reviews addressing the vocal correlates of gender, age, speaking style, speak identity, accent, and emotions [14]-[18]. Firstly, a review starts with the techniques related to prosodic parameterization and modeling. Then the techniques of prosodic modifications will be reviewed from the perspectives of speech production, speech perception, or from both. Then some techniques related to voice quality modeling and modification are reviewed, as well as the techniques related to glottal flow modeling. Next, the existing methods of cross-speaker voice transformations will be reviewed in view of a transposition to the domain of general cross-emotion voice transformation. This topic will be exposed from the perspective of the parametric spectral modeling of speech and then from the perspective of available spectral transformation techniques, which will be classified in the way as the classification of prosody modification methods.

A. Parameterization of Prosody

In linguistics, prosody is the rhythm, stress, and intonation of connected speech. Prosody may reflect various acoustic features of the speaker such as pitch, duration, loudness, formants, speech rate, rhythm, ...etc. Pitch and duration are the most important acoustic features. based on MT, all the information such as the linguistic, expressive and personal information is mainly characterized by pitch, duration and loudness. The formant frequencies can be modified through spectral transformation techniques via different signal models.

1) Modeling of Duration and F0

The modeling of prosody is a whole area of research on speech synthesis and voice conversion. In the framework of unit-selection based speech synthesis, with the help of ToBI-type features, the duration and prosody values rely on the target features as accurate-enough predictors of the prosody and durations [19]. The IBM developed so-called “trainable” unit-based synthesis systems [20] [21], a set of context-clustering CART-trees is used to predict explicitly the prosody and duration values, but independently of the unit sequence. Alternately, in the framework of continuous and semi-continuous probabilistic approaches, the pitch and duration are predicted from an external Gaussian model based on Gaussian Mixture models (GMM) [22], the jointly source and target modeling by GMM [23] [24], and approaches combining GMM and Dynamic Frequency Warping (DFW) [25] [26]. Other probabilistic approaches include the work described in papers [27]-[30]. In the framework of HMM-based synthesis, the prosody is predicted as an explicit value, but in correlation with the more detailed acoustic context, on the basis of the GMM underlying the HMMs [57]; the durations are predicted from an external Gaussian model. In the framework of discrete deterministic approaches, no model is available, the prosodic contours can be used as a template, with the help of DTW, a practice referred to as prosodic transplant, which can be realized by using the VQ, the speaker interpolation approaches and the use of correction filters [31]-[35].

Following the modeling stage, the prosodic modifications can be carried out by using non-parametric and parametric approaches to correct some audible discontinuities, and/or to change a different prosodic contour.

2) Modification of Duration and F0

a) Non-parametric Approaches

For non-parametric approaches, they consist in shifting and/or duplicating some speech frames localized by some corresponding pitch marks. We can mention the Pitch-Synchronous Overlap-Add (PSOLA) algorithm [63], where the speech frames are original signal waveforms; however, the Overlap-Add paradigm can be extended to any other form of coding of the speech frames, including spectral frames in the frequency domain, or envelope-based LPC coefficients [64].

b) Parametric Approaches

For the parametric models, the speech signal is decomposed into parameters which characterize the speech signal. Then these parameters are re-synthesized into speech signal.

From the perspectives of human speech production and perception, these speech analysis-by-synthesis techniques (Vocoders) can be classified based on the underlying scientific principles either from the speech production or perception mechanism or from both.

Methods Based on Production Mechanisms - The ARX-LF based source/filter models for speech signal can be classified into this category [36] [37]. They are developed based on the assumptions that the speech production system is well approximated by an autogressive (AR) filter excited by an LF (Liljencrent-Fant) glottal waveform. A drawback of the ARX-LF model is that, although most of the energy of the AR residual is captured by the LF waveform, a significant portion of the excitation signal is neither captured nor modeled.

In [36], the residual signal is described by a Harmonic-plus-Noise Model (HNM) similar to [22]. The HNM representation seems to be suitable for high-quality time/pitch scaling modifications but it is not evident how to modify the LF residua when the LF source is also modified.

Two versions of LF-vocoder and LF+HM are proposed in [37]: the LF-vocoder is a high quality vocoder that replaces the residual part with modulated noise. The second uses a sinusoidal harmonic representation of the residual signal. The latter does not degrade the signal during analysis/synthesis and provides higher quality for small modification factors, while the former has the advantage of being a compact.

As we mentioned above that modulation phenomena are closely connected with the production process, ARX-LF vocoder is an appreciate candidate of speech model for voice quality transformation.

Methods Based on Perception Mechanisms - The Sinusoidal Transform Coder, STC [38], and the Harmonic plus Noise Model, HNM [22], phase vocoders [39] [40] can be classified into methods based on perception mechanism as all of these methods are developed based on transform method. The advantages and limitations of these approaches are well studied in the literature. Generally, these state-of-the-art vocoders are able to robustly handle a wide range of speech and audio signals, but seem to be restricted in the following ways: 1) the naturalness of synthesized speech is poor for high modification factors, 2) there is reverberation when pitch is significantly lowered (i.e. during female to male conversion),

and 3) they face difficulties when providing sophisticated voice qualities like a relaxed, a harsh or a breathy voice, etc. The weak connection between the signal model and the production mechanism of speech might cause to a significant portion of these deficiencies.

Methods Based on Source, Joint Production and Perception Mechanisms - The Speech Transformation and Representation using Adaptive Interpolation of Weighted spectrum, STRAIGHT [41]. STRAIGHT is basically a channel Vocoder. However, its design objective greatly differs from its predecessors. STRAIGHT is developed based on human speech perception system, which decomposes input sounds in terms of excitation (source) and resonant (filter) characteristics and it performs a periodic/aperiodic decomposition of the source signal and the estimation of Pitch, as well as the spectrum. As the design is based on our auditory system and source/filter theory, STRAIGHT is classified into method based on source, joint production and perception Mechanisms. The quality of synthesized is high. However, the computational complexity is also very high and other parameters related to voice quality are not explicitly estimated.

B. Parameterization of Voice Quality

1) Techniques based on glottal flow models

The voice quality is characterized by the glottal source [65]. It relates closely to the speech production mechanism. Subsequently, it is natural to define the parameterization and modification of voice quality in relation to measurements of the glottal flow from speech waveform.

There are methods aiming at recovering a time-domain signal which describes the glottal flow [42]-[44]. However, in view of voice quality modification, it is necessary to build up a relationship between voice quality or speaker categories and a control model over the glottal wave shapes, in view of modification or classification.

There are considerable work on the analysis, modeling and modification source characteristics in voice quality research [45] – [49]. From the above works, glottal flow modeling and modifications are so complicate that its application to the characterization of voice quality seems to remain an open field of research. A standard practice or a “best model” does not appear, especially regarding the voice quality parameters that should be used to control the glottal flow model.

Globally, only the voiced speech and sustained vowels are analyzed in most of the glottal flow extraction studies, but the behavior of, e.g., the Open Quotient (OQ) or the Normalized Amplitude Quotient (NAQ) measurements in the unvoiced parts of natural speech is not well identified.

Therefore, in order to improve the voice quality of the current voice conversion algorithm, the system can integrate minimal modeling which the glottal phenomena are concerned instead of the use of explicit glottal waveform models.

2) Voice Transformation as Spectral Transformation

Most of the existing cross-speaker voice transformation techniques focus mainly on global spectral transformation. It is widely accepted that the spectral information carries information of speaker individuality. There are two main stages in voice conversion: training and transformation. The training stage consists of three steps in general: acoustic modeling, segmentation and alignment, and acoustic mapping. The transformation stage consists of also three steps: acoustic analysis, modifications/transformations of acoustic features, and transformed speech synthesis. Here mentions the notable works in this domain:

Spectral Transformation Based on Production Mechanisms - Inspired from the work done for speaker adaptation by Shikano *et al.* [50], Abe *et al*. [31] proposes to use vector quantization (VQ) codebook method to solve the mapping of the problem of LPC-based spectral transformation. The approach applies to a cluster of the spectral parameters of both the source and the target speakers. The limitation of this method is that the parameter space of the converted envelope is limited to a discrete set of envelopes, not continuous of envelopes. In practice, the restriction of the variability of the speech envelopes leads to a considerable degradation in the quality of the converted speech signal.

Valbret et al. [51] implement voice transformations as a mix of prosodic transformations based on the PSOLA technique and Linear Prediction Cepstral Coefficients (LPCCs) spectral transformations, inspired from the work of [31]. Vector-Quantisation is used to determine a partition of the speaker’s spectral spaces for a set of linear transforms between the source and the target speaker’s spectra. linear transforms. However, Alsteris and Paliwal [66] remarked that the interpolation of LPCCs can produce unstable Auto-Regressive filters in the equivalent LPC spectral domain, in an experiment produces for speech coding.

In [52], the Line Spectral Frequency is used for spectral transformation through the estimation of a single GMM for the two speakers. In a follow up work [67], in order to increase in the resulting synthetic speech quality, a transformation of the LPC residual is added in the form of a prediction from the spectral shapes. The subjective listening test showed that an overall degradation of the quality with respect to real speech is nevertheless noticed after the voice transformation.

Arslan [53] proposes a voice conversion algorithm based on a two-stage “Segmental Codebooks” mapping method in the Line Spectrum Frequency (LSF) domain. The idea is to quantize the source and target speaker’s subspaces by aligning HMMs to the speech data, and then considering the mean of each state as a codeword. The source speaker acoustic parameters are matches with the source speaker codebook on a frame-by-frame basis by preserving the weights vector while switching the codebook to that of the target speaker. the transformed utterance is obtained by applying a time-varying filter to match the target speaker's acoustic characteristics such as the bandwidth, the pitch-scale, the duration-scale and the energy-scale. The results are evaluated by listening assessment but also, notably, by application of some automatic speaker recognition techniques.

In a follow-up work, several factors are identified by Turk and Arslan [33], which may degrade the performances of the algorithm, particularly when the spectral shapes to be matched between the source and the target codebook are too far from each other. Therefore some methods are proposed to solve

such ill-defined cases out of the codebooks by using correction filters.

But the published sound examples sound quite far from natural.

Spectral Transformation Based on Perceptual Mechanisms - Stylianou et al. [55] proposes Voice Transformation as a linear transformation, operated in a different parametric Mel-cepstral domain, which maps the two speaker's spectral subspaces in a more global way. The speaker's subspaces are modeled by global Gaussian Mixture Models (GMMs) instead of locally isolated partitions. In this work, the PSOLA-based prosodic modification is not used, in contrast to an anterior work by the same authors [22].

Spectral Transformation Based on Source, Joint Production and Perception Mechanisms - Toda et al. [25] proposed the GMM-based algorithm for the spectral transformation, extracted from STRAIGHT analysis method, with dynamic frequency warping to avoid the over-smoothing and an addition of the weighted residual spectrum, which is the difference between the GMM-based converted spectrum and the frequency-warped spectrum, to avoid the deterioration of conversion-accuracy on speaker individuality. However, the spectrum over-smoothing problem has not yet been solved. In a follow up work, Toda et al. [27] proposed the use of delta coefficients for addressing the problem.

A number of speaker adaptation techniques have been proposed and brought successful results in the framework of HMM-based speech recognition [56] and speech synthesis [57]-[61]. In order to compensate the sparsity of speaker-specific data, the idea behind these methods is to use a large amount of speaker-independent data to build a generic speech model, and then to deduce a speaker-adapted model from the generic one by using the available small amount of speaker-specific data. The state-of-the-art adaptation techniques can be divided into three broad classes: 1) Maximum A-Posteriori (MAP) adaptation; 2) Maximum-Likelihood Linear Regression (MLLR); 3) Model Interpolation and Eigenvoices: after HMM training, we obtain the generic model which corresponds to a set of speaker-specific models referred to as "anchor models". A new model for an un seen speaker is computed as a linear combination of the combination of the parameters of the anchor models according to a Maximum Likelihood paradigm with respect to a limited amount adaptation data. This process is interpreted as an interpolation of the anchor models. In order to reduce the dimensionality of the linear transform, Principal Components Analysis (PCA) can be applied to the parameters of the anchor models. This reduced set of anchor models are called Eigenvoices.

The above works have been applied to voice transformation between speakers, not between voice qualities within the same speaker. The work of [3] is an exception, where spectral shapes related to voice quality have been linearly interpolated in the LSF domain to produce intermediate voice quality levels, with good results in terms of subjective perception. This result supports the generalization of the cross-speaker voice transformation techniques to organic-related, speech style-related, emotion-related modifications of the voice quality if there are the organic-related, speech style-related, emotion-related acoustic features in systems. Based on MT, one can surmise that cross-speaker transformations encompass single-speaker/cross emotion transformations because they are designed to model the speaker variability across a whole range of speaking styles. Speaker transformation methods will be tested on (possibly emotional) speaking styles database both on the source and on the target speaker sides for the confirmation.

3) Discussion

The quality of voice transformation algorithms are usually evaluated in subjective tests. So far, from these test results of above methods, the overall impression is that time-scale modification is quite successful for moderate scale factors, while pitch modified signals by pitch scale factors over 1.2 and below 0.8, suffer from various artifacts and listeners classify the modification as not natural. The quality of voice conversion algorithms in transforming the identity of the source speaker to that of the target speaker is not so good as expected and the transformed sound is "deaden". The quality of the current voice conversion algorithms depends on multiples factors such as speech analysis-by-synthesis modeling, the choice of acoustic features, the mapping algorithms, the modification algorithms, and the implementation of algorithms.

Speech Models - As voice transformation belongs to the research of speech synthesis, the speech model should be a type of analysis-by-synthesis vocoder. Most of the current speech models are explored only part of human speech production and perception mechanisms. For example, LPC-based speech model is based on our understanding of the physical properties of the human speech production system. STRAIGHT is based on the human speech perception system. ARX-LF model is assuming that the speech production system is well approximated by an autogressive (AR) filter excited by an LF (Liljencrent-Fant) glottal waveform. The two first speech models have been applied for voice conversion [31] [51] [67] [53] [25] [68]. However, the voice qualities are not taken account into the models. ARX-LF is successfully applied for time, pitch and voice quality modifications [37]. As in Section II, it was mentioned that, based on MT, a good speech model should take into account the nonlinear nature of speech, the interaction of vocal tract and source characteristics, and modulation phenomena. It can model and is capable of controlling acoustic features relevant to voice quality, voice expressivity and voice individuality during sound production. Thus, more accurate speech model tools should be developed for voice transformation.

Choice of acoustic features - Once the speech model is selected, it is preferable to establish linear transformations between the acoustic spaces of two speakers by voice transformation algorithm. For example, LPCs, LSFs, LPCCs and MFCCs features are selected for modifications in voice conversion techniques [31] [51] [67] [53] [25] [68]. However, some of them are unstable during the modification. The selection of acoustic features is determined by theirs linearization, stability and interpolation properties. For speech recognition, MFCCs are good acoustic feature. However, for speech synthesis, it is impossible to recover the original FFT-based spectrum from the MFCCs because the filterbanks operate a non-invertible integration of the spectral samples..

Mapping Algorithms -In Section II, it was mentioned that different types information are characterized by different acoustic features. For example, the organic variation causes the change of the pitch (F0) and the positions of the formants (F1, F2, F3) in the relation to each other and to the F0 reference. However, the current voice conversion algorithms consists in only magnitude of spectral spectrum conversion, pitch and time. The fine structure of spectral details , the formants, and the phase spectrum are not converted. It is reported that human is sensitive to speech phase spectra [66]. Phase sensitivity is more evident considering long analysis window and the solution suggested in [29] is only frame based. A challenge is then to develop voice conversion strategies for these possibly more complicated but more accurate speech models.

The mapping methods is now discussed, which related to the training process. This process is analogous to that of human obtaining the individual and expressive information from the speaking speech. Based on the modulation theory, the listeners only evaluate the deviations of the current properties of speech signal (F0, formant frequencies, lip shape, etc.) from those they expect of a linguistically neutral sound with the same voice quality based on their experience for the speaker or intrinsic properties of speech (e.g., female high pitch, male, low pitch, ...etc.). The training process aims at obtaining this linguistically neutral sound with the same voice quality of the speech. For voice conversion by machine, this process can be realized through machine learning methods.

A systematical comparison study on the mapping methods have been conducted by Baudoin and Stylianou [69], which are based on Vector-Quantization Codebooks, Gaussian-Mixture Models (GMM), Neural Networks and Linear Multivariate Regression. In all cases, the Cepstral Coefficients resulting from the regularized estimation of the spectral envelope [62] is used to perform the mapping . The similar study is also applied to the harmonic part of the HNM model [22]. In terms of normalized spectral distance from the source speaker (further is better) and the target speaker (closer is better), objective tests indicate that the methods performing a global mapping over the whole space, namely GMMs and VQ with weighted map, perform better than the methods which perform a class-dependent mapping or a one-to-one mapping in isolated parts of the vocal space. Subjective listening tests indicate that the GMM-based mapping gives the best results, although speech quality after transformation is judged very poor in all cases. Nevertheless, these results suggest that voice variability is better modeled as a transformation between global models that cover the whole acoustic feature space for each speaker, rather than in terms of a template lookup applied over separated zones of both speaker spaces.

In parallel, speaker adaptation methods in the HMM framework are successfully applied to both areas in speech recognition and in speech synthesis [57]-[61]. In fact, Gaussian Mixture Models (GMMs) and HMMs are closely related. GMMs can be interpreted as a "soft" and global partition of the acoustic space in terms of a set of Gaussian classes, deployed in a multi-dimensional feature space. In this model, a single Gaussian represents the repartition of a data class in terms of second order statistics, i.e., by specification of a mean and a variance. The maximum likelihood measures the degree of belonging of a data point to a Gaussian class. For a mixture of Gaussians, every data point belongs to every class within a certain proportion of likelihood. Models made of sets of class-dependent GMMs are a way to introduce some supervision in the training process, while HMMs are a way to introduce some constraints on the sequential ordering of a set of GMMs. To summary, GMMs and HMMs are based on a statistical modeling paradigm which amounts to a "soft" and global clustering of the acoustical space, and where the use of statistics allows every data point to be informative about every cluster. GMMs and HMMs are therefore particularly well suited to the definition of global transforms between acoustic feature subspaces.

The idea behind speaker adaptation techniques in the framework of HMM-based synthesis fits well the modulation theory. We can interpret that the model of a generic - speaker-independent acoustic space obtained by HMM-based speaker adaptation techniques corresponds to a linguistically neutral acoustic features of a set of speakers in MT. During the voice conversion, the mapping method aims to vary the voice characteristics by adapting HMM parameters to the target speaker. In fact, this adaptation process consists of adapting the generic speaker to the target speaker. That minimizes the different voice properties between the target speaker and the generic model is equivalent to the deviations of the current properties of speech signal (F0, formant frequencies, lip shape, etc.) from the neutral speech sound.

Modification Methods - In transformation part, the transformed features are need to be fused, corrected, smoothed in boundaries or discontinuity points for generating smooth natural speech. More elaborate acoustic feature modification methods should be developed and the appropriate matching parameters between source speaker and target speaker need time to be turned.

IV. Conclusion and Road Map

The present paper has started with presenting the Modulation Theory (MT) of speech founded based on speech perception mechanism, but the combinatorial properties of speech reflect effects of co-articulation. Based on MT, we showed how to improve the quality of voice transformation systems, more efforts should be made to take into account the nonlinear nature of speech, the interaction of vocal tract and source characteristics, and modulation phenomena and results from the natural language processing area.

As a matter of fact, we review the start-of-the-art techniques for prosody, voice quality and spectral spectrum transformations. We attempted also to give a reasonable explanation of different methods based on the modulation theory.

The evaluation of speaker modification techniques has support Gaussian-based acoustic modeling to be the primary the state of the art algorithm, because it is a flexible statistical model to abstract the acoustical realizations of the speech units by a limited number of parameters that are still able to account for some variability, and because it realizes speech

transformations in a class-dependent way (e.g., different transformations for different phonemes), across classes from the data.

In addition, in the area of speech and speaker recognition, a range of Gaussian-based speaker adaptation techniques has been extensively developed, and has been successfully applied to HMM-based voice transformations. Then sets of HMMs or Gaussian models adapted to distinct speech classes can be used to define a model space, where the modification methods such as model interpolation or eigenvoices can be applied to obtain a more explicit control of speaker-specific or speaking style-specific. However, the limitations of HMMs for modeling speech are the following:

- statistics do not vary within an HMM state;
- the probability of state output depends only on the current state;
- the probability of state duration decreases exponentially with time.

Real speech does not hold any one of them.

In the future, voice conversion techniques should take grand challenges in the following aspects:

- to develop more accurate speech analysis-by-synthesis methods by taking into account of the nonlinear nature of speech, the interaction of vocal tract and source characteristics, voice quality and modulation phenomena;
- to integrate human mimicking speech processing and the natural language processing;
- to develop more elaborate acoustic feature modification and matching methods for source speaker and target speaker;
- to develop more better acoustic modeling in the framework of advanced statistical machine learning;
- to enhance robustness of voice conversion systems.

REFERENCES

[1] H. Traunmüller. Evidence for demodulation in speech perception. *ICSLP, workshop on The Nature of Speech Perception*, 2000

[2] H. Traunmüller. Modulation and demodulation in production, perception, and imitation of speech and bodily gestures. in *FONETIK 98*, Dept. of Linguistics, Stockholm University, pp. 40 – 43. 1998.

[3] Y. Stylianou. Voice Conversion: Survey. icassp, pp.3585-3588, 2009.

[4] H. Traunmüller. Perceptual dimension of openness in vowels. *J. Acoust. Soc. Am.* ***69***: 1465 -1475, especially Exp.2 - 4, pp. 1469 - 1472, 1981.

[5] H. Traunmüller. The context sensitivity of the perceptual interaction between F_0 and F_1. *Actes du XIIème Congres international des Science Phonetiques*, Aix-en-Provence, vol. 5, pp. 62 - 65, 1991.

[6] H. Traunmüller. Conventional, biological and environmental factors in speech communication: A modulation theory. *Phonetica* ***51***: 170 - 183, 1994.

[7] H. Traunmüller. Articulatory and perceptual factors controlling the age- and sex-conditioned variability in formant frequencies of vowels, *Speech Comm.* ***3***: 49 - 61, 1984.

[8] R.P. Fahey, and R.L. Diehl. The missing fundamental in vowel height perception. *Perc. & Psychophys.* ***58***: 725 - 733, 1996.

[9] A. Klinkert and D. Maurer. Fourier spectra and formant patterns of German vowels produced at F_0 of 70 - 850 Hz *J. Acoust. Soc. Am.* ***101***: 3112 (A)., 1997.

[10] E. Zetterholm. Same speaker different voices: A study of one impersonator and some of his different imitations. *Proc. Int. Conf. Speech Sci. & Tech.*, pages 70–75, 2006.

[11] A. Eriksson and P. Wretling. How flexible is the human voice?-A case study of mimicry. *Proc. Eurospeech*, pages 1043–1046, 1997.

[12] T. Kitamura. Acoustic analysis of imitated voice produced by a professional impersonator. *Proc. Interspeech*, pages 813–816, 2008.

[13] H. Kuwabara and Y. Sagisaka. Acoustic characteristics of speaker individuality: Control and conversion. *Speech Communication*,16(2):165–173, 1995.

[14] S. Furui. *Digital Speech Processing, Synthesis, and Recognition*. Marcel Dekker, 1989.

[15] L. Rabiner, and B.-H. Juang. *Fundamental of Speech recognition* Prentice-Hall, Upper Saddle River, NJ, 1993.

[16] M. Schröder. Emotional speech synthesis: A review. In Proc. Eurospeech'01,Scandinavia, 2001.

[17] M. Schröder. Speech and Emotion Research. An Overview of Research Frameworks and a Dimensional Approach to Emotional Speech Synthesis. PhD thesis, Institut für Phonetik , Universität des Saarlandes. Phonus no.7, 2004.

[18] S. Roehling, B. MacDonald, and C. Watson. Towards expressive speech synthesis in English on a robotic platform. In Proc. 11th Australasian International Conference on Speech Science and Technology, Auckland, New Zealand. Univ. of Auckland, 2006.

[19] K. Silverman, M. Beckman, M. Pierrehumbert, J. Ostendorf, M. Wightman, C. Price, P. and Hirschberg, J. Tobi. A standard scheme for labeling prosody. In Proc. ICSLP'92, Banff., 1992.

[20] R. Donovan, and E. Eide. The IBM trainable speech synthesis system. *In Proc. ICSLP'98*, Sydney, Australia, 1998.

[21] J. Pitrelli, R. Bakis, E. Eide, R. Fernandez, W. Hamza, and M. Picheny. The IBM expressive text-to-speech synthesis system for american english. *IEEE Transactions on Audio, Speech and Language Processing*, 14(4):1099–1108, 2006.

[22] Y. Stylianou, J. Laroche, and E. Moulines. High-Quality Speech Modification based on a Harmonic + Noise Model. *Proc. EUROSPEECH*, 1995.

[23] A. Kain. *High resolution voice transformation*. PhD thesis, OGI School of Science and Eng., Portland, Oregeon, USA.

[24] A. Mouchtaris, J. Van derSpiegel, and P.Mueller. Non parallel training for voice conversion based on a parameter adaptation. *IEEE Trans. Audio, Speech, and Language Processing*, 14(3):952–963, 2006.

[25] 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.

[26] D. Erro, T. Polyakova, and A. Moreno. On combining statistical methods and frequency warping for high-quality voice conversion. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2008.

[27] T. Toda, A.W. Black, and K. Tokuda. Spectral Conversion Based on Maximum Likelihood Estimation considering Global Variance of Converted Parameter. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 9–12, Philadelphia, USA, 2005.

[28] L. Meshabi, V. Barreaud, and O. Boeffard. GMM-based Speech Transformation Systems under Data Reduction. *6th ISCA Workshop on Speech Synthesis*, pages 119–124, August 22-24, 2007.

[29] H. Ye and S. Young. Quality-enhanced voice morphing using maximum likelihood transformations. *IEEE Trans. Audio, Speech, and Language Processing*, 14(4):1301–1312, July 2006.

[30] H. Duxans, A. Bonafonte, A. Kain, and J. van Santen. Including dynamic and phonetic information in voice conversion systems. *Proc. ICSLP*, pages 5–8, 2004.

[31] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara. Voice conversion through vector quantization. In *Proc. ICASSP88*, pages 655–658, 1988.

[32] N. Iwahashi and Y. Sagisaka. Speech spectrum transformation based on speaker interpolation. In *Proc. ICASSP94*, 1994.

[33] O. Turk and L. M. Arslan. Robust processing techniques for voice conversion. *Computer Speech and Language*, 20:441–467, 2006.

[34] W. Verhelst and M. Roelands. An overlap-add technique based on waveform similarity (wsola) for high quality time-scale modification of speech. In *Proc. ICASSP93*, pages 554–557, 1993.

[35] J. van Santen, A. Kain, E. Klabbers, and T. Mishra. Synthesis of prosody using multi-level unit sequences. *Speech Communication*, 46:365–375, 2005.

[36] D. Vincent and O. Rosec. A new method for speech synthesis and transformation based on a ARX-LF source-filter decomposition and HNM modeling. in *ICASSP*, 2007.

[37] Y. Agiomyrgiannakis, O. Rosec. ARX-LF-based source-filter methods for voice modification and transformation. icassp, pp.3589-3592, 2009.
[38] R. J. McAulay and T. F. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-34(4):744–754, Aug 1986.
[39] P. Depalle and G. Poirrot. SVP: A modular system for analysis, processing and synthesis of sound signals. in *Proceedings of the International Computer Music Conference*, 1991.
[40] J. Laroche and M. Dolson. Improved phase vocoder timescale modification of audio. *IEEE Transactions on Audio and Speech Processing*, vol. 7, no. 3, 1999.
[41] H. Kawahara. Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 1303–1306, Munich, Germany, 1997.
[42] J. Liu, G. Beaudoin, and G. Chollet. Studies of glottal excitation and vocal tract parameters using inverse filtering and a parameterized input model. *In Proc. ICSLP'92*, pages 1051–1054, Banff, Alberta, Canada, 1992.
[43] P. Alku. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech Communication*, 11:109–118, 1992.
[44] O. O. Akande, and P. J. Murphy. Estimation of the vocal tract tranfer function with application to glottal wave analysis. *Speech Communication*, 46:15–36, 2005.
[45] D. G. Childers. Glottal source modeling for voice conversion. *Speech Communication*, 16:127–138, 1995.
[46] G. Fant, J. Liljentcrats, and Q. Lin. A four parameter model of glottal flow. In Quarterly Progress and Status Report, number 4 *in STL-QPSR*, pages 1–13. KTH, Stockholm, Sweden, 1985.
[47] C. d'Alessandro, and B. Doval. Experiments in voice quality modification of natural speech signals: the spectral approach. *In Proc. 3rd ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*, Jenolan Caves House, Blue Mountains, NSW, Australia, 1998.
[48] P. Mokhtari, H. R. Pfitzinger, and C. T. Ishi. Principal components of glottal waveforms: towards parameterisation and manipulation of laryngeal voice quality. *In Proc. VOQUAL'03*, Geneva, 2003.
[49] M. Lugger, B. Yang, and W. Wokurek. Robust estimation of voice quality parameters under real world disturbances. *In Proc. ICASSP'06*, pages 1097–1100, 2006.
[50] K. Shikano, K. Lee, and R. Reddy, "Speaker adaptation through vector quantization," *in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 1986, pp. 2643–2646.
[51] H. Valbret, E. Moulines, and J. Tubach. Voice transformation using PSOLA technique. *Speech Communication*, 11:175–187, 1992.
[52] A. Kain, and M. W. Macon. Spectral voice conversion for text-to-speech synthesis. *In Proc. ICASSP'98*, volume 1, pages 285–288, 1998.
[53] L. M. Arslan. Speaker transformation algorithm using segmental codebooks (STASC). *Speech Communication*, 28:211–226, 1999.
[54] O. Turk, and L. M. Arslan. Robust processing techniques for voice conversion. *Computer Speech and Language*, 20:441–467, 2006.
[55] Y. Stylianou, O. Cappé, and E. Moulines, E. Continuous probabilistic transform for voice conversion. *IEEE Trans. on Speech and Audio Processing*, 6(2):131–142, 1998.
[56] P. Woodland. Speaker adaptation for continuous density hmms: a review. *In Proc. ITRW on Adaptation Methods for Speech Recognition*, pages 11–19, Sophia Antipolis, 2001.
[57] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In Proc. Eurospeech'99, volume 5, pages 2347–2350, Budapest, Hungary, 1999.
[58] T.Masuko, T., Tokuda, K., Kobayashi, T., and Imai, S. (1997). Voice characteristics conversion for HMM-based speech synthesis. *In Proc. ICASSP'97*, pages 1611–1614.
[59] T. Yoshimura, T. Masuko, K. Tokuda, T. Kobayashi, and T. Kitamura. Speaker interpolation in HMM-based speech synthesis system. *In Proc. Eurospeech'97*, Rhodos, Greece, 1997.
[60] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi. Speaker adaptation for HMM-based speech synthesis usingMLLR. *In Proc. 3rd ESCA/COCOSDAWorkshop (ETRW) on Speech Synthesis*, Blue Mountains, Australia, 1998.
[61] K. Shichiri, A. Sawabe, T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. Eigenvoices for HMM-based speech synthesis. *In Proc. ICSLP'02*, Denver, Colorado, 2002.
[62] O. Cappé, J. Laroche, and E. Moulines. Regularized estimation of cepstrum envelope from discrete frequency points. *In Proc. IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, Mohonk, 1995.
[63] E. Moulines, and F. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5):453–467, 1990.
[64] E. Moulines, and W. Verhelst. *Time-domain and frequency-domain techniques for prosodic modification of speech*. In Kleijn, W. and Paliwal, K., editors, Speech Coding and Synthesis, chapter 15, pages 519–555. Elsevier Science B.V., 1995.
[65] Laver, J. (1980). *The Phonetic Description of Voice Quality*. Cambridge University Press.
[66] L.D. Alsteris and K.K. Paliwal. Short-time phase spectrum in speech processing: A review and some experimental results. *Digital Signal Processing*, 17:578–616, 2007.
[67] A. Kain, and M. W. Macon. Design and evaluation of a voice conversion algorithm based on spectral envelop mapping and residual prediction. *In Proc. ICASSP'01*, 2001.
[68] J. Yamagishi, H. Zen, Y.-J. Wu, T. Toda, and K. Tokuda. The HTS-2008 system: Yet another evaluation of the speaker-adaptive HMM-based speech synthesis system in the 2008 Blizzard Challenge. In *Proc. Blizzard Challenge 2008*, Brisbane, Australia, September 2008.
[69] G. Baudoin, and Y. Stylianou. On the transformation of the speech spectrum for voice conversion. *In Proc. ICSLP'96*, Philadelphia, PA, USA, 1996.