VOICE TRANSFORMATION: A SURVEY

Yannis Stylianou

Institute of Computer Science, FORTH, and Multimedia Informatics Lab, CSD, UoC, Greece
email: yannis@csd.uoc.gr

ABSTRACT

Voice transformation refers to the various modifications one may apply to the sound produced by a person, speaking or singing. Voice Transformation is usually seen as an add-on or an external system in speech synthesis systems since it may create virtual voices in a simple and flexible way. In this paper we review the state-of-the-art Voice Transformation methodology showing its limitations in producing good speech quality and its current challenges. Addressing quality issues of current voice transformation algorithms in conjunction with properties of the speech production and speech perception systems we try to pave the way for more natural Voice Transformation algorithms in the future. Facing the challenges, will allow Voice Transformation systems to be applied in important and versatile areas of speech technology; applications that are far beyond speech synthesis.

Index Terms— Voice Transformation, Speech production, speech perception, voice quality, speaking style.

1. INTRODUCTION

Voice Transformation aims at the control of non-linguistic information of speech signals such as voice quality and voice individuality. To efficiently control this information, Voice Transformation systems need to understand the process and mechanism at a production and a perception level. Of course there are other speech research areas where the production and perception mechanisms were efficiently used. This is for example the speech coding systems. Note that in this case and assuming an ideal communication channel the main problem is to efficiently model the perceptual relevant properties of the speech signal. In Voice Transformation the above process is also required. However, in Voice Transformation the parameters of the speech model should be modified in a way that respects the production of speech while perceptually is acceptable. In a few words, we may say that the main task of speech coding is "reproduction" of speech, while Voice Transformation is its "modification". It is however well known that for modifying a process, the process should be understood. Let us provide another example to further support this. While there are common research interests between Voice Transformation and speaker dependent technologies (e.g., speaker recognition/verification) Voice Transformation goes beyond these technologies. For Voice Transformation is not sufficient to just identify, represent and detect the cues that are relevant to voice individuality; the representation of these cues needs to be modified in a way that the modified/transformed speech signal sounds natural. Therefore, speech models suggested for Voice Transformation should be able to efficiently represent these cues and modify them.

Building high-quality Voice Transformation systems require to take into account phenomena that are usually ignored or overlooked in other speech research areas. This includes the nonlinear nature of speech, and the interaction of vocal tract and source characteristics. Modulation phenomena observed in speech signals are tightly connected with the production process. Ignoring these phenomena may lead to less-natural speech transformations and to a quality which is not be possible to have been produced by humans. Therefore, such phenomena should be respected during transformation and should be manipulated accordingly if natural transformed voice is the target and not a cartoon character (which however, could be indeed a target!).

Most of the main speech signal technologies, i.e., speech recognition, speech enhancement, speech synthesis, etc. work at a frame level. For Voice Transformation, a fusion of prosodic features at levels higher than that of the usual speech frame should be performed in order to define the speaking style of a speaker, to recognize characteristic patterns and then suggest techniques to map one speaking style to another or just conduct convinced transformations on a style. Although Voice Transformation goes beyond the goal of the aforementioned speech research areas, obviously it is heavily based on these areas. We may mention as examples the speech representations from speech coding, the cues of voice individuality from speaker recognition, transformations/adaptations from speech recognition and definitions of speaking style from natural language processing.

It is widely accepted that some Voice Transformation methods, like time-scaling, are quite successful, while quality of pitch modified speech still needs improvement. Methods for mapping a spectral envelope of a source speaker to that of a target speaker, a process that is required for Voice Conversion, are effective in transforming speaker identity but they produce very low quality of speech. Furthermore, current Voice Conversion systems mainly work at a sentence level and their extension to long speech signals (discourse, lectures) should consider speaking style information. This means that even if Voice Conversion is considered to be successful in transforming speaker identity at a sentence level, this is not true for longer speech (i.e., a short lecture). Other limitations of current Voice Transformation systems can be even identified in the main definitions of basic transformations such as pitch and time-scale and energy modifications.

The remainder of the paper is organized as follows. First, the various definitions used in the context of voice transformation are outlined in Section 2, followed by a review of the current state-of-the-art methods for voice transformation. In Section 4 limitations of the current systems is discussed and we provide a list of the challenges that should be addressed in order the voice quality of transformed speech signals to be improved. In Section 5 the versatile areas of application of voice transformation systems are reviewed and overall conclusions are provided in Section 6.

2. TRANSFORMING VOICES; WHAT CAN WE TRANSFORM AND HOW?

During speech production two mechanisms maybe distinguished; one is referred to as "Software" and the other is referred to as "Hardware" [1]. Software is mostly related to the speaking style, emotions, mood, and social status of the speaker, while Hardware is mostly related to the articulators that take part during voicing. Although control of articulators provide ways to control the quality of voice, it has been reported in the literature that compared with the Hardware characteristics, the control of Software characteristics is equally, or in some cases even more, important from a perception point of view for a successful (or, convinced) transformed/converted voice result. Actually a mime tries to mimic the Software part of the target speaker. Studies contacted with professional impersonators are quite useful in answering the question of what can we transform and how. However such studies are limited. Most of the studies show that the professional impersonator captures the speech style in the voice imitation, particularly the rhythm, the intonation and stressed words and phrases [2] while it is difficult to accurately modify source formant frequencies towards a given target [3]. A recently published paper confirms the prosody and speaking style imitation [4] however states that there are changes in the impersonator's formant frequencies towards the frequencies of the target voice. There is however a very important difference between the data set used in these studies. In the former mimicry studies, the speech material was from about 30s long excerpts of *uninterrupted* speech, while the latter study was only based on two short sentences. Nevertheless, all these important studies show that the "Software" part is very important. Unfortunately, the Software part of speech is not easily measurable and only recently there are some efforts to define the speaking style of a speaker. For example, most of the research for a specific case of Voice Transformation, that of voice conversion, deals with the control of the Hardware part and it has been shown that for isolated sentences the converted speech from a source speaker may be perceived as if the speech has been uttered by a target speaker. However, transforming longer speech such a short lecture, it is not be a convinced conversion by just concatenating transformed short sentences. There, the high-level prosody/style information will be missing and the absence of the target speaking style model will be easily noticeable.

Focusing therefore on the articulators, and assuming that the source-filter is a valid one for the production of speech, the following speech modifications/transformations have been defined for the source and the filter part.

2.1. Source Modifications

Source modifications mainly include three types of modification: *time-scale modification*, *pitch modification*, and *energy modification* and they are usually referred to as *prosodic modifications*. The goal of time-scale modification is to change the apparent rate of articulation without affecting the perceptual quality of the original speech. This means that the formant structure is changed at a slower or faster rate than the rate of the input speech, but otherwise the structure is not modified. The goal of pitch modification is to alter the fundamental frequency in order to compress or to expand the spacing between the harmonic components in the spectrum while preserving the short-time spectral envelope (the locations and bandwidths of formants) as well as the time evolution. The goal of energy modification is to modify the perceived loudness of the input speech. It is considered to be the simplest modification among the prosodic modifications since the signal is just multiplied by a scale factor which corresponds in amplifying or attenuating all the frequencies by the same factor.

2.2. Filter Modifications

By filter modification we mean the modification of the magnitude spectrum of the frequency response of the vocal tract system. It is widely accepted that magnitude spectrum carries information of speaker individuality. Therefore, by modifying the magnitude spectrum of the vocal tract, speaker identity may be controlled. We may distinguish two types of filter modification: (1) Without a specific target: In this case the magnitude spectrum is modified in a general way without having a specific target speaker in mind. For example, modifying a female voice so that it sounds more like a child voice. This type of modification is usually referred to as *Filter Modification*. (2) With a specific target: In this case the filter of a speaker (source speaker) is modified in a way that the modified filter approximates in the mean squared sense the characteristics of the filter of another speaker (target speaker). Usually this modification is referred to as *Filter Mapping*.

It is worth noting that filter modifications are usually defined only for the magnitude spectrum while it is well known that phase spectrum carries information about speech and speaker characteristics [5].

2.3. Combining Source and Filter Modifications

Usually, Voice Transformation systems combine source and filter modifications. For example, if we want to modify the voice of a speaker so that it sounds like the voice of another speaker, prosody and vocal tract modifications should be combined. If a specific target speaker is provided then this is referred to as *Voice Conversion*. To the contrary, when not a specific target is provided, this is usually referred to as *Voice Modification*. *Voice Morphing* is another type of combined source and/or filter modifications. In this case, however, the same sentence is uttered by two speakers (source speakers) and then, a third speaker (a virtual one) is created for that only specific sentence.

3. STATE-OF-THE-ART

For prosodic modifications there are many non-parametric and parametric approaches. For the non-parametric approaches we can mention the Time Domain - Pitch Synchronous Overlap and Add, TD-PSOLA [7] and the Waveform Similarity Overlap and Add, WSOLA [8]. For the parametric models we can mention the Sinusoidal Transform Coder, STC [9], the Harmonic plus Noise Model, HNM [10] and the Speech Transformation and Representation using Adaptive Interpolation of Weighted spectrum, STRAIGHT [11].
For filter modifications (mostly for mapping) we may mention discrete deterministic and continuous or semi-continuous probabilistic approaches. The latter approaches are most popular. As examples of discrete and deterministic approaches we can mention the VQ-mapping [12], the Speaker Interpolation approach [13] and the use of correction filters [14]. For the probabilistic approaches we may mention the Continuous Probabilistic Approach based on GMM [15], the Jointly Source and Target modeling by GMM [16] [17], and approaches combining GMM and Dynamic Frequency Warping (DFW) [18] [19]. Other probabilistic approaches include the work described in [20],[21],[6], and [22].
For modifications of short-time spectra for Voice Transformation

there are not currently too many works. We can only mention the work presented in [6].
Also there is a limited number of works towards the modeling and modification (or mapping) of the speaking style of a speaker. This high-level information is one important missing part in the current Voice Conversion systems. We can mention the work referred to as *Voice Fonts* [23] and the probabilistic approach using HMM [24].

4. LIMITATIONS AND CHALLENGES

Voice transformations are usually evaluated in subjective tests. The overall impression from the results obtained from these tests 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 making listeners to classify the modifications as not natural. Voice conversion reaches high score in transforming the identity of the source speaker to that of the target speaker. However, there are serious quality problems mostly referred to as *muffled effects*.
To improve the quality of speech produced by the various proposed in the literature Voice Transformation algorithms a better understanding of speech production and perception mechanisms is necessary. This may lead us to question the definitions provided earlier in Section 2. For example, when we want to increase the loudness of our voice while sitting in a cafeteria, we add stress to a *part* of our speech signal, like consonants, and *not* to all the speech events we produce. One hypothesis for this is that consonants carry most of the information load which is connecting with the intelligibility of the message we would like to transmit. According, therefore, to this hypothesis we only increase the stress to specific sounds by an amount that is sufficient to mask the cafeteria noise. Increasing the stress does not mean that the amplitudes of all the frequencies for this sound are equally increased. Stress means an increase of the subglottal pressure which will result in an abrupt glottal closure by accentuating the Bernoulli effect on airflow through the glottis [25]. This corresponds to more energy *mostly* at high frequencies. From this example, it is obvious that even the simple intensity modification is not as simple as we considered in Section 2. Continuing the above example, the increase of the subglottal pressure will increase the tension in the vocal folds resulting in an increase of the pitch. This shows that modifying one parameter may require the modification of another as well. Consonants are shorter in duration than vowels (which carry more prosodic information). Our perceptual system requires some time to process the perceived sounds. When we want to speak faster, we somehow *protect* the consonants. Pickett [26] has done extensive studies on the degree of change in vowels and consonants in speaking at a faster or slower rate. In [26] it was reported that when going from normal to the faster rate, the vowels were compressed by 50% while the consonants were compressed by 26%. However, going from the slowest to the fastest rate, both vowels and consonants were compressed by about 33% [25]. This shows that time-scale modifications should take into account the phonetic information. Speaking at faster or slower rate introduces modifications in pitch values since there are fluctuations in the subglottal pressure. This means that time-scale modifications should be performed jointly with pitch modifications. A real challenge therefore will be the modeling of these interdependent processes. For this, more accurate speech analysis tools should be developed.
In the source-filter theory for speech, it is assumed that glottal airflow source is not influenced by the vocal tract. In reality, there exists a nonlinear coupling between the source and the filter. Results from studies on the fine structure of the glottal airflow derivative waveform show that an increase in first-formant bandwidth and modulation of the first-formant frequency occurs during the glottal open phase [27]. Obviously, when pitch modification is applied, these interactions should be respected. Attempts have been made to incorporate some of these observations into the modification algorithms [28] [29] however, further work on speech analysis and modeling is required.
Many researchers report that voice converted speech has a muffling effect possibly because of the broadening of spectral peaks in the converted speech and of the lost of spectral details in the converted spectra. To cope with this effect, statistical approaches using the global variance of the converted spectra in each utterance have been proposed [20] and estimation of the lost spectral details [16]. These solutions are frame based without taking into account the evolution of spectral information over time which is perceptually more important than the local spectral information. A straightforward approach to address this problem is the use of delta coefficients [20]. To efficiently address the spectrum oversmoothing in frequency and in time, it requires the use of non-stationary time-varying speech representations which will be able to model long segments of speech. A challenge is then to develop conversion strategies for these possibly more complicated but more accurate speech models.
It becomes more and more evident that indeed we are sensitive to speech phase spectra [5]. When only spectral magnitudes are modified while the original phase spectrum is preserved a harsh quality is perceived which is quite annoying when one work with studio-quality speech recordings (in contrast to speech enhancement where the original recordings are very noise and where the original (noisy) phase is preserved and only the magnitude spectrum is modified). Phase sensitivity is more evident considering long analysis window and the solution suggested in [6] is only frame based. Another challenge is therefore, how to efficiently represent phase over time and frequency and then manipulate this information for voice conversion.
Last but indeed not least, the biggest challenge over all, at this stage of voice transformation/conversion algorithms, is the control (modeling, mapping, modification) of the speaking style of a speaker.

5. APPLICATIONS

Voice Transformation was considered as a hot, novel and fast-growing topic in 1990s having as potential application the concatenated speech synthesis systems where new (virtual or target) voices could be created without requiring to pass through the extremely expensive process of developing new voices. By that time, it was widely accepted that Voice Transformation systems were far from providing the required performance. With the recent developments in speech synthesis this need is more pronounced. There is an increasing demand for high quality Voice Transformation methods not only for creating target or virtual voices, but also to model various effects (e.g., Lombard effect), synthesize emotions, to make more natural the dialog systems which use speech synthesis etc. Besides speech synthesis, however, Voice Transformation has other potential applications in areas like entertainment, film, and music industry, toys, chat rooms and games, dialog systems, security and speaker individuality for interpreting telephony, high-end hearing aids, vocal pathology and voice restoration.
Because of space limitations we can not refer further to these important applications of Voice Transformation. However it is evident that Voice Transformation needs to address the challenges listed in the previous section and improve then the transformed voice quality. Once this is achieved Voice Transformation will be used in

numerous and versatile applications.

6. CONCLUSIONS

Voice Transformation covers a wide area of research from speech production modeling and understanding to perception of speech, from natural language processing, modeling and control of speaking style, to pattern recognition and statistical signal processing. To improve further the quality of Voice Transformation systems, more efforts should be made to take into account the nonlinear phenomena during the speech production process and results from the natural language processing area. More accurate, flexible, and meaningful for speech, models should be developed for modeling longer speech segments than current (stationary) models do. However, it is evident that Voice Transformation requires more than just modeling the speech signal; it requires *understanding* the speech process in terms of production, perception, and natural language processing.

7. REFERENCES

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

[2] 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.

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

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

[5] 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.

[6] 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.

[7] E. Moulines and J. Laroche. Techniques for pitch-scale and time-scale transformation of speech. part I. non parametric methods. *Speech Communication*, 16(2), February 1995.

[8] 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.

[9] 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.

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

[11] H. Kuwahara. 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.

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

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

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

[15] Y. Stylianou, O. Cappé, and E. Moulines. Statistical methods for voice quality transformation. *Proc. EUROSPEECH*, 1995.

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

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] A. Verma and A. Kumar. Voice Fonts for Individuality Representation and Transfirmation. *ACM Trans. on Speech and Language Processing*, 2(1):1–19, 2005.

[24] C.H. Wu, C.C. Hsia, T.H. Liu, and J.F. Wang. Voice conversion using duration-embedded Bi-HMMs for expressive speech synthesis. *IEEE Trans. Audio, Speech, and Language Processing*, 14(4):1109–1116, 2006.

[25] T. F. Quatieri. *Discrete-Time Speech Signal Processing*. Prentice Hall, Engewood Cliffs, NJ, 2002.

[26] J.M. Pickett. *The Sounds of Speech Communication*. Pro-Ed, Inc., Austin, TX, USA, 1980.

[27] C.R. Jankowski. *Fine Structure Features for Speaker Identification*. PhD thesis, Massachusetts Institute of Technology, Dept. of EE and CS, June 1996.

[28] D. Kapilow, Y. Stylianou, and J.Schroeter. Detection of non-stationarity in speech signals and its application to time-scaling. *Proc. EUROSPEECH*, 1999.

[29] A. Kain and Y. Stylianou. Stochastic modeling of spectral adjustment for high quality pitch modification. In *Proc. ICASSP2000*, Istanbul, 2000.