Techniques for Crosslingual Voice Conversion

Anderson Fraiha Machado
University of São Paulo, Brazil
Email: dandycgms@gmail.com

Marcelo Queiroz
University of São Paulo, Brazil
Email: mqz@ime.usp.br

***Abstract*—The crosslingual voice conversion problem refers to the replacement of a speaker's timbre or vocal identity in a recorded sentence, assuming that the source speaker and target speaker use different languages. This problem differs from typical voice conversion in the sense that the mapping of acoustical features cannot depend on time-aligned recordings of source and target speakers uttering the same sentences. This paper presents an overview of a general crosslingual voice conversion system and discusses the most important techniques used in each step of the conversion process.**

***Keywords*-Crosslingual Voice Conversion; Timbre transformation; Vocal Identity;**

I. Introduction

Through speech, humans express not only facts and ideas but also emotions and sensations, through emphasis, melodic contour and rhythmic variation of speech, while at the same time communicating other attributes of their individuality, such as regional accents. All these aspects together contribute to what is known as speech identity, which consists of a set of time-varying acoustic parameters (such as prosody or spectral and dynamic envelope) that characterizes the speaker and allows its differentiation among other speakers. This set of acoustic parameters is related to the musical concept of **timbre**, which will be used here to refer to a speaker's sound identity.

The development of speech processing digital systems has an increasing interest in many areas of application, and several projects involved in speech and natural language processing have been highlighted in recent years, such as the personalization of automatic voice translators or Speech-to-Speech Translation (SST) systems. The TC-Star project [9], for instance, brought together dozens of researchers from around the world to aid in a multi-lingual SST system. Recently[1], Google executives disclosed that Android cell phones will soon have an integrated Google Translate system to perform real-time voice translation.

The process of **voice conversion** is defined as the transformation of the sound identity of a speech signal in such a way that the sentences appear to be spoken by a different speaker, while the message contents remains unmodified. Manipulating the timbre or speaker identity of a voice signal is a very challenging task, because the modification of musical parameters should not interfere with the phonetic structures that guarantee intelligibility of the sentences.

[1]The Times, February 7 2010.

A. Definitions

In order to define the timbre and prosody transformations, Voice-Conversion Systems (VCS) usually rely on a training phase, which can be text-dependent or text-independent. In the first case, both source and target speakers record the same sentences, which are then time-aligned to allow synchronous feature-extraction and mapping of acoustic parameters. On the other hand, the Crosslingual Voice Conversion (CVC) problem [1] assumes that source and target speakers use different languages, and the training phase is usually text-independent (although text-dependent training with bilingual individuals is sometimes possible).

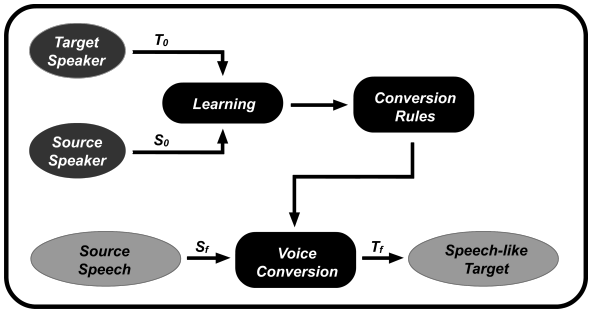

Figure 1. A typical voice conversion system.

The focus of this work lies on solutions for text-independent CVC systems using spectral transformations, including fundamental frequency f_0 shifting, formant frequency mapping and spectral tilting, which are some of most important descriptors of voice timbre [12].

B. State of the Art in Crosslingual Voice Conversion

Since the original formulation of the voice conversion problem by Childers et al. [5], the analysis/re-synthesis structure remains basically unchanged. The main differences between proposed systems lie in the set of techniques used for representation and transformation of the acoustic parameters. Five categories of techniques are proposed in [16]: statistical techniques (e.g. GMM, HMM, PCA, K-means), cognitive techniques (e.g. ANN), linear algebra techniques (e.g. SVD) and signal processing techniques (e.g. VQ, FW).

Figure 2 shows how often such techniques appear in the literature of the field. This is based on 76 references listed in [16], covering the period from 1986 to 2010.

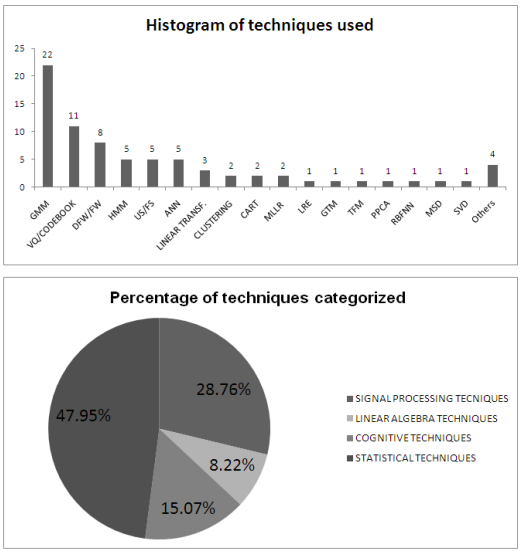

Figure 2. Techniques used in VC systems.

Early works dealt with the voice conversion problem within a single language. The text-dependent CVC problem involving bilingual subjects first appeared in 1990 [1], and text-independent CVC appeared in 2003 when Kumar [15] proposed the creation of a corpus of acoustical information for each individual speaker (Voice Fonts). After this seminal contribution, some more recent contributions to the text-independent CVC problem are shown in Table I.

Year	Author	Technique
2003	Kumar [15]	GMM
2003	Sundermann [23]	DFW
2006	Duxans [9]	GMM & CART
2006	Sundermann [22]	US
2007	Erro [10]	WFW
2009	Uriz [25]	FS& FW
2009	Zhang [29]	VQ
2009	Desai [8]	ANN

Table I
AUTHORS AND TECHNIQUES IN TEXT-INDEPENDENT CVC

In the next section some of the most frequent techniques in voice conversion will be presented, as a conceptual framework for a crosslingual voice conversion system.

II. A FRAMEWORK FOR A CROSSLINGUAL VOICE CONVERSION SYSTEM

The process of converting the voice quality of a speaker has two somewhat distinct phases: the **training phase** and the **transformation phase**. The training phase is divided into two modules: the first deals with obtaining the corpus of source and target speakers, and the second creates the mapping between phonetic classes and acoustic parameters of both speakers.

A. Obtaining the Corpus

In the process of obtaining the so-called corpus of a speaker, both local and global acoustic features related to sound identity are stored. Figure 3 shows a detailed diagram of this process.

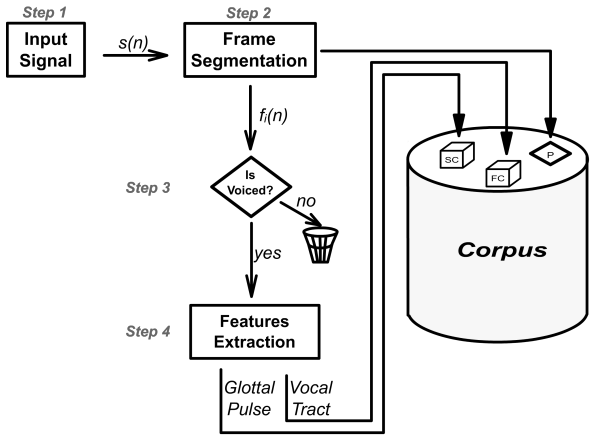

Figure 3. Process of obtaining a corpus of a speaker.

1) Obtaining the Input Signal: A VC system usually takes as input a set of recorded sentences from each speaker, that corresponds to a training set. Some care must be taken to ensure that the training set is adequate: for instance, sentences must explore the whole phonetic content of that particular language, to ensure that the feature space covers the phonetic spectrum that may be found in the next phase; recordings must be carefully registered, preferrably with sampling rates above 8 KHz and samples of 16 bits or more, and with as little background noise as possible.

2) Frame Segmentation: In order to adequately represent the voice quality of a speaker, one is supposed to take into account both sentence-based features that capture global aspects of the voice, such as pitch and energy profiles, and also frame-based features that capture instantaneous information, such as phonetic content.

The input signal may be divided into frames according to two different approaches: (1) fixed-size frames, usually on the order of 10 ms to 25 ms [1]; or (2) *pitch-synchronous frames* [26], where the choice of the size depends on an estimate of the fundamental frequency.

3) Voiced/Unvoiced Decision: The *voiced/unvoiced* (or V/UV) decision problem [13] is closely related to the fundamental frequency estimation problem and silence detection in noisy recordings. Some of the techniques for deciding whether a given signal is voiced or unvoiced use measures

such as frequency band ratios [19], zero crossing rates [4] and cepstrum coefficients [2].

One of the biggest problems in this phase is the lack of robustness of this decision module, which might result in treating a noisy information as an almost periodic signal, causing errors in a subsequent stage of conversion.

Unvoiced frames are usually discarded during the training phase, since such portions are not considered to significantly characterize the sound identity of an individual. However, some systems model these excerpts as white noise filtered by a band-pass filter, for purposes of voice representation [18].

4) Acoustic Feature Extraction: Given a frame of a voiced signal, vocal tract (also referred to as the **filter component**) and glottal pulse (**source component**) are separated, and stored in the form of acoustic feature vectors, as shown in Figure 4.

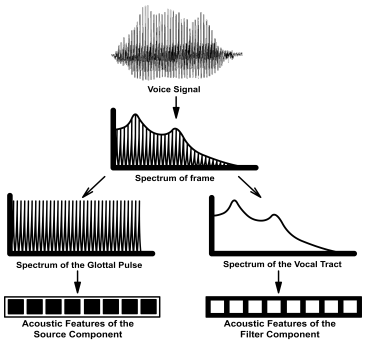

Figure 4. Process for obtaining acoustic feature vectors.

One of the critical decisions in this step corresponds to the choice of models for representing glottal pulse and vocal tract. For example, the vocal tract is frequently modeled by LPC, LSF, MFCC or IPSE [24] coefficients, where the choice of the number of coefficients is also an important aspect to be taken into consideration. Moreover, there is no universally accepted model for the vocal tract that also takes into account individual harmonic variations, besides those responsible for the phonetic content of the frame (formantic regions).

5) Storage of Feature Vectors: Finally, the acoustic parameters of the speaker are stored in a special database called *corpus* (see Figure 3). These are organized as Filter (or vocal tract) Components and Source (or glottal pulse) Components, which are labeled **FC** and **SC**, as well as global control variables related to prosody (**P**).

The corpus of the source and target speakers may be progressively fed by the system during training, or at runtime in the case of an adaptive training. The acoustic parameters of vocal tract and glottal pulse are normally stored in separate dictionaries called *codebooks*, which are properly organized in a map of acoustic characteristics. These maps also store the order in which these parameters entered the dictionaries, providing important information related to the temporal continuity of frames, which may be used in the resynthesis process.

B. Mapping between Phonetic Classes

The mapping between phonetic classes can be defined both in the training phase (static mapping), as in the techniques of dynamic selection of frames with selection units, and in the transformation phase (dynamic mapping), as in the case of neural networks and other decision structures. At this stage a mapping between the spaces of acoustic characteristics of source and target speakers is defined, which takes into account artificial phonetic classes, vocal tract characteristics and glottal pulse characteristics.

In dynamic mapping, some algorithms use matching of feature vectors from the source and target corpus by maximum similarity, not taking into account the phonetic classes. In some cases, grouping of phonetic classes is unnecessary, since the selection method treats the whole feature space as a single class, for instance as in *Frame Selection* [25] or in *Unit Selection* [9] methods.

1) Grouping in Artificial Phonetic Classes: An artificial phonetic class corresponds to a cluster of acoustic vectors around a central point in the space of acoustic parameters. Since artificial phonetic classes are defined by distance measures, such metrics should be chosen in order to minimize the perceptual difference between samples of the same phonetic class, such as the Itakura distance [28] or Bark Spectral Distortion [27]. There are several approaches to cluster points in the feature space, including supervised clustering techniques, such as K-means and K-histograms, and unsupervised techniques, such as hierarchical clustering.

2) Definition of the Selector Switches: Once phonetic classes are defined, the system must be able to access them from selector switches that are unique to each class. One of the most common alternative is to use the centroids of these clusters. Together with the centroids, other information can also contribute to specification of the selector switches, such as geometric information (e.g., the radius and orientation of the axes of the ellipsoid surrounding the cluster), statistical information (for example, measures of dispersion, skewness and kurtosis), among others.

After the set of parameters that make up the selector switch of the phonetic class is specified, this switch can be properly quantized (VQ) to be used as the primary key of its corresponding class. Using the VQ/codebook technique, Arslan [3] proposes a widely used algorithm for voice conversion, known as *Speaker Transformation Algorithm using Segmental Codebooks (STASC)*. This technique is a fusion of VQ/codebook with other mapping techniques, such as spectral, glottal pulse and prosody mappings. The general idea is to generate a codebook for each speaker from the acoustic parameters that represent the vocal tract.

An alternative to this process of quantization is to use decision-making structures associated with each phonetic

class. Examples of decision-making structures are Neural Networks [18], Topological Feature Mapping [21] and decision trees [9].

A neural network $\mathcal{R}_i(s)$, for example, receives a set of input values (acoustic features) associated with a particular phonetic class $\mathcal{C}_i$ of the target's corpus. During the training phase, the weights of the network are updated and adjusted, and during the transformation phase the system receives a parameter vector s of the source, and must find the corresponding phonetic class $\mathcal{C}_i$ which fits s best. In this case, we find the index k such that

$$k = arg\min_i(\mathcal{R}_i(s)).$$

3) Definition of the Mapping between Phonetic Classes: Let S and T be two artificial phonetic classes, belonging to the space of characteristics of source and target speakers, respectively, corresponding to a common phoneme. The objective of this phase of the voice conversion is to find a mapping function $\Psi : \mathcal{C}_S \rightarrow (\mathcal{C}_T)$ that maps the centroid $\mathcal{C}_S$ of S to the centroid $\mathcal{C}_T$ of T which minimizes the distance in both feature spaces. Normally, this mapping function Ψ is represented by a mapping structure $\mathcal{T}$ (e.g. an associative table), which is used in a later stage of the processing (see Figure 5).

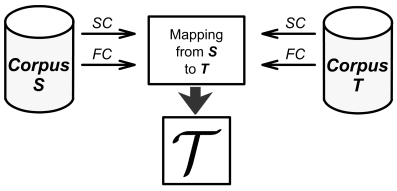

Figure 5. Mapping from the source corpus to the target corpus.

The mapping of classes $S \Rightarrow T$ can be accomplished in accordance with the type of training. For example, in text-dependent training, the mapping of classes is accomplished by a time alignment of the frames.

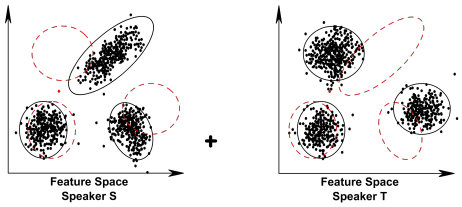

Figure 6. Example of paired acoustical spaces.

One of the major difficulties in CVC is the incompatibility of phonetic classes between the languages L_S and L_T of source and target speakers. In this case, it is extremely difficult to separate aspects of the formant structure of the spectrum of a speech signal that correspond to phonetic content from those that correspond to the voice timbre of an individual.

One possible solution is the use of a bilingual speaker which is fluent in both L_S and L_T. It is hoped that in this case there are more artificial phonetic classes in the feature space of the source speaker, simplifying the definition of the mapping.

C. Transformation of Acoustic Features

The transformation of acoustic parameters can be performed during runtime (dynamic transformation) or using an approximation function defined in the training phase (static transformation). For example, a dynamic real-time transformation can be realized as formant shifts, according to the phonetic class from the mapping of the previous phase. An example of static transformation is the association of each pair of phonetic classes with a transformation structure, such as a Gaussian Mixture Model defined during the mapping phase, to be used in the conversion phase.

The basic process of transforming a source speaker's voice signal to the voice of the target may be seen in Figure 7.

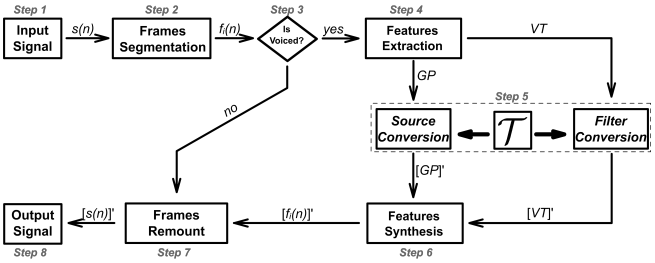

Figure 7. The process of converting a speech signal.

1) From Input Signal Vectors to Features: Analogously to the first four steps in Phase 1 of the construction of the corpus, the first steps here correspond to the basic preparation of the input, management and segmentation of frames, V/UV decision, and extraction of acoustic features for the vocal tract and glottal pulse, according to Section II-A.

Clearly, the method for choosing the frame size should be same as in the corpus acquisition phase, due to compatibility in the extraction of acoustic parameters. Unvoiced sounds are transmitted directly to the assembly module of the algorithm (Step 7).

2) Conversion of Acoustic Parameters: Given an acoustic parameter vector v_i extracted from i-th frame of a voiced speech signal belonging to a phonetic class S of the feature space of the source speaker (represented by the centroid $\mathcal{C}_S$), it is possible to find the centroid $\mathcal{T}(\mathcal{C}_S) = \mathcal{C}_T$ in the corpus of the target speaker, using the mapping table $\mathcal{T}$ (see Section II-B). The artificial vector $\mathcal{C}_T$ serves as a target in the approximation (or interpolation) of v_i, i.e. an approximation $\mathcal{A}$ applied to v_i is defined according to the pair of centroids associated with the phonetic classes S and T in such a way that

$$\mathcal{A}(\mathcal{C}_S) \approx \mathcal{C}_T.$$

This approximation is divided into two parts: the conversion $\mathcal{A}_{\text{source}}$ of the excitation source (glottal pulse) and the conversion $\mathcal{A}_{\text{filter}}$ of the resonance filter (vocal tract).

Conversion of the Source Component: The conversion of the source component follows two basic steps: harmonic adjustment and prosodic adjustment. The harmonic adjustment receives acoustic vectors GP which are normalized with respect to pitch and energy, and transforms them according to the transformation $\mathcal{A}_{\text{source}}$. It is expected that at the end of this transformation, the excitation source on the vector GP looks like a signal of the glottal pulse of the target speaker (without his/her prosodic aspects).

Prosodic adjustment in the glottal pulse transforms pitch and energy according to global parameters stored in the target corpus. The relationship between the prosodic parameters of the source and target speakers can be obtained from their respective corpus (see Figure 3), and thus one may obtain global control parameters for pitch and energy, such as the first two statistical moments describing these temporal profiles.

In general, the transposition of average pitch and average energy is made using a linear function, whose scale is defined by these external variables. There are specific techniques for implementing pitch shifts such as the methods TD-PSOLA or FD-PSOLA [26], or by changing the fundamental frequency in a resynthesis using the Harmonic plus Noise Model [10]. Among the alternatives for converting the glottal pulse, an experimental evaluation is required to get the best representation and transformation of the excitation source.

Conversion of the Filter Component: Similarly, the acoustic feature vectors VT are transformed using an approximation function. Given a vector v, an approximation technique $\mathcal{A}_{\text{filter}}$ basically interpolates a vector u, so that $\mathcal{A}_{\text{filter}}(u) \approx v$.

The Gaussian Mixture Model [14] (GMM) is the most commonly used technique in voice conversion due to its properties of smooth interpolation of spectral envelopes of consecutive frames in time, keeping the criterion of minimizing mean square error. One potential problem of using GMM is excessive smoothing [17], which may occur due to a poor choice of the number of Gaussian components of the model.

Desai [7] uses neural networks as universal models of function approximation, by training a neural network $\mathcal{R}_{(S,T)}(s)$ for each pair of phonetic classes $\mathcal{C}_S$ and $\mathcal{C}_T$ of source and target, respectively, that maps each vector $s \in \mathcal{C}_S$ to a vector $t \in \mathcal{C}_T$. Thus, such networks $\mathcal{R}_{(S,T)}(s)$ are trained and subsequently selected in the processing stage, according to the input vector s.

Basically, the Frequency Warping technique [25] is a mapping function that modifies the scale on the axis of frequencies, mapping frequencies ω of a frequency spectrum to $\tilde{\omega}$ in the corresponding distorted spectrum. Daniel Erro [10] recently proposed a Weighted Frequency Warping method that combines frequency warping and GMM.

3) Voice Signal Resynthesis from the Acoustic Features Processed: The resynthesis of the output speech signal is carried out by transforming the modified acoustic vectors $[VT]'$ and $[GP]'$ of the vocal tract and glottal pulse, respectively, to a voice signal in time domain. This voice signal is an audio frame $[f_i(n)]'$ with the same length as the input frame $f_i(n)$ (see Figure 7).

4) Assembling the Voice Signal from Converted Frames: It often happens that the voice conversion is successful on the frame level, but the improper assembly of resynthesized speech frames affect the audio quality leading to degradation of voice signal. The concatenation of frames of converted voice without any care results in many defects, such as clicks arising from the discontinuity of periodic components in consecutive frames, phase distortion or even the creation of nonexisting frequencies because of interference from the windowing function that is also transformed during the conversion.

Among the most common solutions to solve these problems are the technique known as Phase Bashing [20] and the use of pitch-synchronous frames [26], along with Overlap-Add methods [6]. It is also possible to use modeling techniques based on sinusoidal models, in order to use vocoders to re-synthesize the output voice signal [11].

5) Finishing Process for the Output Voice Signal: This output voice signal obtained by the concatenation of individual frames is further refined in order to increase its audio quality, eliminating unpleasant effects such as noise, missing frequencies and any clicks present in the signal. In some cases, you can choose to perform an overall adjustment (EQ) and an enhancement of unvoiced regions in order to increase the intelligibility of sentences.

III. Conclusion

This paper presented some of the most important techniques used in crosslingual voice conversion. The steps in a typical crosslingual voice conversion system have been detailed in order to discuss important aspects of system design, such as the construction of a corpus for each speaker, the representation models for vocal tract and glottal pulse and the transformation techniques that allow the imprint of a different vocal timbre to a speech signal.

One of the open issues in crosslingual voice conversion that should be addressed in a future work is the specification of a benchmark for subjective evaluation of different voice conversion systems that allows the comparison of different representation and tranformation strategies in a unified setting [16].

References

[1] M. Abe, K. S., and H. K. Cross-language voice conversion. In *ICASSP*, 1990.

[2] S. Ahmadi, A. S. Spanias, N. M. P. Inc, and C. A. San Diego. Cepstrum-based pitch detection using a new statistical V/UV classification algorithm. *IEEE TSAP*, 7(3):333–338, 1999.

[3] L. M. Arslan. Speaker transformation algorithm using segmental codebooks (STASC). *Speech Communication*, 28, 1999.

[4] B. Atal and L. Rabiner. A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition. *IEEE TASSP*, 24(3):201–212, 1976.

[5] D. G. Childers, B. Y., and Ke W. Voice conversion: Factors responsible for quality. *ICASSP*, 1985.

[6] R. Crochiere. A weighted overlap-add method of short-time Fourier analysis/synthesis. *IEEE TASSP*, 28(1):99–102, 1980.

[7] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. Black, and K. Prahallad. Voice conversion using artificial neural networks. In *IEEE SLT*, 2008.

[8] S. Desai, B. Yegnanarayana, and K. Prahallad. A Framework for Cross-Lingual Voice Conversion using Artificial Neural Networks. In *7th ICON*, 2009.

[9] H. Duxans, D. Erro, J. Pérez, F. Diego, A. Bonafonte, and A. Moreno. Voice conversion of non-aligned data using unit selection. In *TC-STAR WSST*, 2006.

[10] D. Erro and A. Moreno. Weighted frequency warping for voice conversion. In *Interspeech*, 2007.

[11] J. L. Flanagan, D. I. S. Meinhart, R. M. Golden, and M. M. Sondhi. Phase vocoder. *J. ASA*, 38:939, 1965.

[12] T. Sone T. Nimura H. Matsumoto, S. Hiki. Multidimensional representation of personal quality of vowels and its acoustical correlates. *IEEE TAE*, 21(5):428–436, 1973.

[13] J. C. Hardwick and J. S. Lim. Voiced/unvoiced estimation of an acoustic signal, June 1 1993. US Patent 5,216,747.

[14] A. Kain and M. W. Macon. Spectral voice conversion for text-to-speech synthesis. In *IEEE ICASSP*, volume 1, 1998.

[15] A. Kumar and A. Verma. Using phone and diphone based acoustic models for voice conversion: a step towards creating voice fonts. In *ICASSP*, pages 720–723, 2003.

[16] A. F. Machado and M. Queiroz. Voice conversion: A critical survey. In *SMC*, 2010.

[17] L. Mesbahi, V. Barreaud, and O. Boeffard. Comparing GMM-based speech transformation systems. In *Interspeech*, pages 1989–1992, 2007.

[18] M. Narendranath, H. A. Murthy, S. Rajendran, and B. Yegnanarayana. Transformation of formants for voice conversion using artificial neural networks. *Speech Communication*, 16, 1995.

[19] M. Nishiguchi, J. Matsumoto, and S. Ono. Voiced/unvoiced decision based on frequency band ratio, September 28 1999. US Patent 5,960,388.

[20] M. S. Puckette. Phase bashing for sample-based formant synthesis. In *ICMC*, pages 733–736, 2005.

[21] A. Rinscheid. Voice conversion based on topological feature maps and time-variant filtering. In *4th ICSLP*, 1996.

[22] D. Sundermann, H. Hoge, A. Bonafonte, H. Ney, A. Black, and S. Narayanan. Text-independent voice conversion based on unit selection. In *ICASSP*, 2006.

[23] D. Sundermann, H. Ney, and H. Hoge. VTLN-based cross-language voice conversion. In *ASRU*, 2003.

[24] K. Tanaka and M. Abe. A new fundamental frequency modification algorithm with transformation of spectrum envelope according to F0. In *IEEE ICASSP*, volume 2, 1997.

[25] A. J. Uriz, P. D. Aguero, A. Bonafonte, and J. C. Tulli. Voice Conversion using K-Histograms and Frame Selection. *Interpeech*, 2009.

[26] H. Valbret, E. M., and J.P. T. Voice Tranformation Using PSOLA Technique. In *2nd ECSCT*, 1991.

[27] M. Benbouchta W. Yang and R. Yantorno. Performance of the modified bark spectral distortion as an objective speech quality measure. In *ICASSP*, 1998.

[28] J. Wouters and M. W. Macon. A perceptual evaluation of distance measures for concatenative speech synthesis. In *ICSLP*, 1998.

[29] M. Zhang, J. Tao, J. Nurminen, J. Tian, and X. Wang. Phoneme cluster based state mapping for text-independent voice conversion. In *ICASSP*, pages 4281–4284, 2009.