VOICE CONVERSION: A CRITICAL SURVEY

Anderson F. Machado*
Computer Science Dept. – University of São Paulo
dandy@ime.usp.br

Marcelo Queiroz*
Computer Science Dept. – University of São Paulo
mqz@ime.usp.br

ABSTRACT

Voice conversion is an emergent problem in voice and speech processing with increasing commercial interest, due to applications such as *Speech-to-Speech Translation (SST)* and personalized *Text-To-Speech (TTS)* systems. A *Voice Conversion* system should allow the mapping of acoustical features of sentences pronounced by a source speaker to values corresponding to the voice of a target speaker, in such a way that the processed output is perceived as a sentence uttered by the target speaker. In the last two decades the number of scientific contributions to the voice conversion problem has grown considerably, and a solid overview of the historical process as well as of the proposed techniques is indispensable for those willing to contribute to the field. The goal of this text is to provide a critical survey that combines historical presentation to technical discussion while pointing out advantages and drawbacks of each technique, and to bring a discussion of future directions, specially referring to the development of a perceptual benchmark process for voice conversion systems.

1. INTRODUCTION

Speech is an inherently human communication tool. The development of computational systems that process speech in various ways is a very interesting and important challenge. Systems that concentrate on the intelligible content of speech, such as speech recognition and text-to-speech systems, have received widespread attention due to important applications in providing accessibility for disabled individuals, as well as applications in human-computer interface design and in security systems. Some systems focus mainly on the timbral quality of speech, for instance speaker identification systems, whereas others are equally concerned about intelligible and timbral aspects, such as singing voice synthesis [1].

This paper concentrates on the *Voice Conversion (VC)* problem as introduced by Childers et al. [2], which is the task of converting a sentence uttered by a source speaker in such a way that the converted result appears to be the same sentence spoken with a different voice, i.e. that of a target speaker. It is important to make a few distinctions between (VC) and related problems clear.

Voice Transformation is a general problem that encompasses all tasks and methods that modify any features of a voice signal; for instance, pitch shifting or time-stretching a recorded sentence are examples of voice transformations. *Voice Morphing* is a term borrowed from image processing, and is the special case of voice transformation where two voices are blended to form a virtual third voice, where usually the two source voices speak or sing the same thing synchronously. Instances of these techniques have been made known to the general public through films such as *Farinelli* (where a soprano and a countertenor voices are blended to make up a pretended castratto voice) or *Alvin and the Chipmunks* (where chipmunks voices are pitch-shifted/formant-corrected actors voices).

A *Voice Conversion* system takes into account both the *timbre* and the *prosody* of the source and target speakers. While timbre and prosody are qualities that are easy to recognize and hard to define in general terms, in the specific context of voice conversion timbral features are usually associated with the dynamic spectral envelope of the voice signal, whereas prosody is related to pitch/energy contours and rhythmic distribution of phonemes.

In order to define the transformations related to timbre and prosody, VC systems usually depend on a training phase, which may be **text-dependent** or **text-independent**. In the first case, both source and target speakers have to record the same sentence; after that, both recordings are time-aligned (using for instance *Dynamic Time Warping* [3, 4, 5]), and acoustic features are mapped synchronously between recordings.

In the *text-independent* case [6], source and target speakers are not required to record the exact same sentences. Recordings are usually segmented into frames which are mapped into a feature space and clustered into groups of similar frames, defining *artificial phonetic categories*, which may or may not coincide with usual phonemes. Acoustical parameters of the source sentence are then mapped within each category, according to similarity of source and target frames.

Another distinctive aspect of VC systems is related to the *phonetic content* of the languages used in training and in actual conversion. In voice conversion within a single language, both text-dependent and text-independent trainings are feasible, and artificial phonetic classes are more likely to reflect actual phonetic classes, since the sets of phonemes present in source and target recordings are basically the same.

On the other hand, *Crosslingual Voice Conversion* [7] assumes that source and target subjects speak different languages (A and B, respectively), and a sentence (in lan-

**Acknowledgements: The authors would like to thank FAPESP and CAPES.*

guage A) from the source speaker should sound as if spoken (untranslated, of course) by the target speaker. This process involves a text-independent training strategy, and is predicated on the assumptions that similar phonemes exist within the languages A and B, and that substitution by similar phonemes would not instantly prohibit comprehension of the converted speech.

Some attempts at crosslingual conversion have been made using bilingual individuals [8], since they allow text-dependent training, by using sentences in language B spoken by the source speaker. This allows the specification of timbral transformations between similar phonemes (in language B) from source and target, which are later applied to phonemes in language A to obtain similar phonemes with the timbre of the target.

1.1 Applications

There has been an increasing interest in voice conversion systems, specially in telecommunication companies such as *CENET* [9]. Some of the applications of voice conversion with a commercial interest are:

- Customization of Text-To-Speech (TTS) interactive systems [6, 10].
- Personalized virtual interpreters: this is a combination of speech recognition followed by automatic translation and finally TTS in the destination language using the voice of the original speaker. Some examples are the **Verbmobil** project (German/English and German/Japanese real-time voice translation), and **TC-STAR** [11, 12, 6, 13, 14, 15, 16, 17] which aims at providing *Speech-to-Speech Translation (SST)* in several languages.
- Biometric voice authentication systems: the development of voice conversion techniques has a natural interplay with the development of voice authentication systems, which are subject to attacks (in this case using voice disguise) as any other security system.
- Voice restoration systems: these are aimed at people who suffered some voice-impairing pathology.

1.2 A Typical Voice Conversion System

Figure 1 presents a sketch of a typical voice conversion system. The system receives sentences from the source speaker (S_0) and from the target speaker (T_0), which are used in a training phase to define a transformation (T) from source speaker features (which may be local or global) to target speaker features. Afterwards, the system receives a new sentence S_f from the source speaker and synthesizes a sentence T_f, which should carry the same message as S_f but with the vocal qualities of the target speaker.

The training phase is generally the first necessary step for voice conversion. This is the stage where data from both speakers is collected and processed, in order to obtain a reasonable characterization of the acoustic features of each speaker, thus allowing the definition of the transformation to be used in the subsequent stage. According to figure 1, in this phase the VCS:

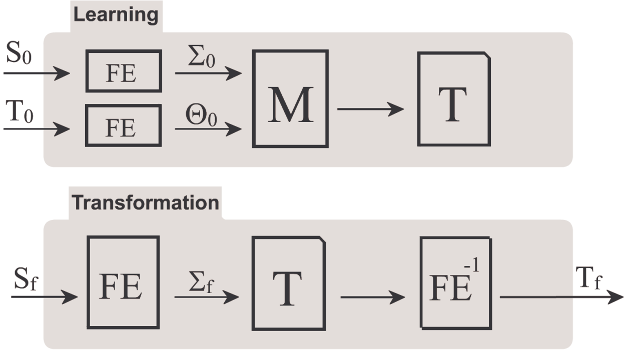

Figure 1. Typical Voice Conversion System.

1. Receives input sentences S_0 and T_0.
2. Extracts relevant acoustic features for each speaker, creating alternative representations Σ_0 and Θ_0 for source and target speakers, respectively.
3. Processes the acoustic features Σ_0 and Θ_0 in order to obtain a database of local (frame-based) and global (sentence-based) descriptors for both speakers.
4. Defines a transformation from local and global source descriptors to local and global target descriptors.

The transformation phase has a similar structure: input sentences S_f are represented as Σ_f in an acoustical feature space, which is then converted by the transformation defined in the training phase into a representation Θ_f, which is finally inverse transformed into a sentence T_f.

The alternative representation of the sentences in a space of acoustical features (Σ and Θ in the diagram) is supposed to preserve enough information so as to allow not only plain resynthesis but also manipulation of timbral and prosodic aspects of the signal. Since many of these are time-varying attributes of the signal, the extraction of acoustic features is usually done on a frame-by-frame basis. Acoustic descriptors are said to be local if they describe a feature of a single frame, and global if they correspond to a whole sentence or to a model of the speaker. Examples of local descriptors are instantaneous pitch, energy, and spectral envelope, or the artificial phonetic category to which a particular frame belongs. Examples of global descriptors are means and standard deviations of pitch or energy measurements, or estimates of the glottal pulse and vocal tract for each speaker.

There is no general consensus with respect to frame size. In theory, frames smaller than $10\ ms$ may be considered stationary due to the inertia of larynx and vocal tract within such timespans [18]. In practice, frames of $15\ ms$ or $25\ ms$ are frequently used [19, 20], and are considered stationary in a broader sense. Some authors [7, 13] prefer to use variable-sized frames defined by an integral number of periods of the quasi-stationary voice signal, which are called *pitch-synchronous frames*.

The choice of frame size is also related to the choice of sample rate, since both combined determine spectral accuracy. In theory, sample rates should never be smaller than 6 kHz, since the human voice has important formant frequencies below 3 kHz, but in practice the sample rates of 8 kHz and 16 kHz are frequently used, and larger sample rates are advised for high-quality voice conversion.

The reconstruction of a voice signal from processed frames must be carefully planned, since changes in spectral content may introduce audible artifacts due to phase differences between adjacent frames, and may be perceived as high-frequency noise, clicks or ringing frequencies that did not exist in the input signal.

The following section presents a historical overview of articles dealing with voice conversion, and also the main techniques introduced. Section 3 presents a comparative overview of these contributions based on perceptual tests. Section 4 discusses some of the frequent difficulties faced when developing VC systems, and some of the proposed solutions. Finally, section 5 summarizes possible future directions in VC development, as presented in the recent literature, and discusses comparison standards for future VC systems.

2. STATE-OF-ART TECHNIQUES IN VOICE CONVERSION SYSTEMS

This section brings a historical overview of techniques used in Voice Conversion, as well as a classification of the techniques according to their interrelations.

2.1 Historical Overview of VC techniques

Many voice conversion (VC) techniques have been proposed since the original formulation of the voice conversion problem by Childers et al. [2] in 1985. Childers proposed solution involved a mapping of acoustical features from a source speaker to a target speaker. A year later, Shikano [21] proposes to use *vector quantization (VQ)* techniques and *codebook* sentences.

A few years later, in 1990, Abe [19] introduces the idea of crosslingual voice conversion systems (CVCS) using bilingual subjects, and in 1991, Valbret [22] rekindles the discussion by proposing personalized *Text-to-Speech* systems using the idea of *Dynamic Frequency Warping (DFW)*.

In 1995, Childers [20] introduces the idea of VC based on the physiological model of glottal pulse and vocal tract, and Narendranath [23] adds *Artificial Neural Networks (ANN)* to the list of VC techniques.

By the end of the 90's, Arslan [24] proposes a model using *Line Spectral Frequencies* for spectral envelope representation, which results in the *STASC (Speaker Transformation Algorithm using Segmental Codebooks)* algorithm, and Stylianou [9] uses *Gaussian Mixture Models (GMMs)* combined with *Mel-Frequency Cepstral Coefficients (MFCCs)* as an alternative to spectral envelope representation.

In 2001, Toda [25] proposes a combined spectral representation and voice conversion technique named *STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum)*, which allows the manipulation of spectral, acoustical and rhythmic parameters. A year later Türk [26] proposes a variation of Arslan's *STASC* algorithm using the *Discrete Wavelet Transform (DWT)*.

Sündermann [7] made a series of contributions since 2003. He has established the concept of text-independent voice conversion and has been the first to propose a text-independent crosslingual voice conversion system that did not require bilingual subjects for training the system. He also brought up to the field of voice conversion a technique known as *Vocal Tract Length Normalization (VTLN)*, which had been originally proposed in 1995 by Kamm et al. [27] in the context of speech recognition.

More recent contributions by Rentzos [28], Ye [29] [29], Rao [30] and Zhang [31] have focused in probabilistic techniques, such as *GMMs*, *codebook sentences*, and techniques such as *ANN* and *DFW*, among others.

In the next section, a more detailed view of voice signal representation and transformation techniques frequently used in voice conversion systems is given.

2.2 Classification of VC techniques

Voice conversion techniques may be classified according to the acoustical features used in the alternative representation of the signals, as well as according to the transformation techniques employed in conversion.

2.2.1 Representation Models

There are a few parameters that are usually computed for each frame, such as pitch (F0), energy (rms), and some representation of frequency content, which is fundamental both for classification and transformation of voice quality. Besides the Fourier spectrum and its envelope, voice conversion systems use many other representation models for a voice signal, such as:

Voice-based models: Vocal Tract Length Normalization (VTLN), Formant Frequencies, and Glottal Flow models.

Mixed Voice/Signal Models: Linear Prediction Coding (LPC), Line Spectral Frequencies (LSF), Cepstral Coefficients, and Speech Transformation and Representation using Adaptative Interpolation of weiGHTed spectrum (STRAIGHT).

Signal-based models: Improved Power Spectrum Envelope (IPSE), Discrete Wavelet Transform (DWT), Harmonic plus Noise Model (HNM).

Voice-based models are based on representations of human voice-producing mechanisms, using concepts such as glottal pulse, which is the raw signal produced by vocal folds, and vocal tract, which comprises the oral and nasal cavities, palate, tongue, jaw and lips, and is responsible for many timbral voice qualities.

Mixed voice/signal models are actually signal models that provide compact representations for the signals. Since they are largely used by the speech recognition community, they acquired many helpful voice-related interpretations. For instance, parts of the cepstrum are often related to formant regions (and thus to vocal tract contribution) or to the fundamental frequency, and LPC coefficients and LPC residuals can also sometimes be associated to vocal tract and glottal pulse (viewed in a subtractive synthesis context).

Purely signal-based models are based on general time-domain and frequency-domain signal representations, and are usually devoid of specific voice-related or phonetic-related semantics. The harmonic-plus-noise model is more specific than the others, but is specially useful in tracking voiced portions (e.g., vowels) of the signal.

Besides the usual linear frequency spacing that is common to Fourier-based methods, many of these techniques also allow the use of alternative frequency scales such as *BARK* and *MEL* [32].

2.2.2 *Transformation Techniques*

The transformation phase in voice conversion systems is concerned with every acoustic feature used in the representation of the voice signal. This includes pitch shifting and energy compensation, but also the transformation of frequency content in such a way that both timbral aspects and intelligibility are taken into account. Transformation techniques are intrinsically tied to representation models. Some of the common techniques are:

Statistical Techniques: Gaussian Mixture Models (GMM), Hidden Markov Models (HMM), Multi-Space Probability Distributions, Maximum Likelihood Estimators (MLE), Principal Component Analysis (PCA), Unit Selection (US), Frame Selection (FS), K-means, K-histograms.

Cognitive Techniques: Artificial Neural Networks (ANN), Radial Basis Function Neural Networks (RBFNN), Classification And Regression Trees (CART), Topological Feature Mapping, and Generative Topographic Mapping.

Linear Algebra Techniques: Bilinear Models, Singular Value Decomposition (SVD), Weighted Linear Interpolations (WLI) and Perceptually Weighted Linear Transformations, and Linear Regression (LRE, LMR, MLLR).

Signal Processing Techniques: Vector Quantization (VQ) and Codebook Sentences, Speaker Transformation Algorithm using Segmental Codebooks (STASC), and Frequency Warping (FW, DFW, WFW).

These techniques basically differ with respect to the way they look at data. For instance, statistical techniques usually assume that data such as feature vectors or vocal parameters have a random component and may be reasonably described by means and standard deviations (Gaussian model), or that they evolve over time according to simple rules based on the recent past (Markov models).

Cognitive techniques are based on learning processes using abstract neuronal structures, and usually depend on a training phase (where both inputs and outputs are available). Frequently they are used for decision problems (where only 2 possible output values are available), for instance in speech recognition, where a separate network is trained for every specific phoneme or word or sentence that is going to be recognized.

Linear algebra techniques are based on geometrical interpretations of data, for instance in finding simplified models by orthogonal projection (linear regression), in obtaining convex combinations of input data (weighted interpolations), or in decomposing transformations into orthogonal components (SVD).

Signal processing techniques define transformations based on time-domain or frequency-domain representations of the signal. These may try to encode a signal using a library of frequently found signal segments or codewords, or to convert timbre-related voice qualities by modifying frequency scale representations (warping).

The above categories also frequently overlap in the case of voice-conversion techniques. For instance, LRE and SVD are frequently used as statistical tools, and both PCA and vector quantization are built using linear algebra framework.

In the next section some of the usual methods for evaluating voice conversion systems are presented and discussed.

3. EVALUATIONS

The two fundamental questions related to the evaluation of voice conversion addresses the ideas of *similarity* of the timbre of the converted voice with respect to the target voice, and of the *quality* of the result with respect to sound artifacts or intelligibility. Several attempts have been made in trying to answer those questions both objectively and subjectively.

3.1 Objective Evaluation

In this setting it is required that the target sentence be also recorded by the target speaker, thus providing a golden standard to which the converted sentence is compared. Both target and converted sentences are time-aligned and then a global distance between the time-aligned sentences is computed. This global distance can be computed by accumulating differences between time-aligned frames, or using other acoustical measures, such as cepstral distortion (CD) [25], among others.

It has been observed that such objective measures are not necessarily correlated to human perception or to human preferences [33]. In fact, some works report large objective distances and good subjective evaluations [26].

3.2 Subjective Evaluation

Due to the difficulty in defining good objective distance measures which are perceptually meaningful, and also due to the difficulty in comparing objective values using different metrical distances, some authors have preferred to evaluate the performance of their systems by standard subjective tests such as *MOS* and *ABX*.

The *MOS* or *Mean Opinion Score* test is basically an evaluation process using 5 values for grading the output, in this case the quality of the converted voice and its similarity to the target voice. The values are standardized as **5**=excellent, **4**=good, **3**=fair, **2**=poor, **1**=bad. The project TC-STAR [11, 12] proposes a standard perceptual test using *MOS* as a measure of both quality and similarity.

Tables 1 and 2 bring a collection of MOS results for quality and similarity, respectively, of several voice conversion systems, as presented by their authors. In experimental voice conversion tests, a distinction is usually made between *intra-gender* conversion (indicated by $M \rightarrow M$ and $F \rightarrow F$ in the tables) and *inter-gender* conversion ($M \rightarrow F$ and $F \rightarrow M$). The method of Rao [30] has received the highest grading for quality of conversion, whereas Rentzos [28] methods have been graded higher for similarity of timbres. Some authors [16, 3] reported that variants of their methods were able to significantly improve quality gradings at the expense of lower similarity gradings, probably due to excessive smoothing issues (see section 4).

Year	**Author**	**Quality *MOS***	**Technique**
1997	Kim [34]	3.42	VQ
1998	Kain [10]	4.20 ($M \rightarrow M$) 2.70 ($M \rightarrow F$)	GMM
1998	Stylianou [9]	≈ 2.70	GMM
2001	Toda [25]	≈ 4.20 ($M \rightarrow M$) ≈ 2.70 ($F \rightarrow F$)	GMM DFW
2004	Pfitzinger [5]	≈ 1.50	WLI
2005	Toda [35]	≈ 3.10 ($F \rightarrow M$) ≈ 3.30 ($M \rightarrow F$)	MLE
2006	Nurminen [14]	2.09	GMM
2006	Duxans [6]	2.37	GMM CART
2006	Sündermann [17]	2.7 (Txt-Dependent) 2.6 (Txt-Independent)	US
2006	Rao [30]	4.56 ($M \rightarrow F$) 4.71 ($F \rightarrow M$)	WLI
2006	Shuang [15]	4.09 (UK English) 3.68 (CN Mandarin)	FW
2007	Dutoit [3]	2.56	FS
2007	Erro [13]	3.27 ($M \rightarrow M$) 3.00 ($M \rightarrow F$) 3.60 ($F \rightarrow M$) 4.20 ($F \rightarrow F$)	WFW
2007	Fujii [4]	3.03 ($F \rightarrow F$) 2.75 ($M \rightarrow F$)	US
2008	Shuang [16]	3.48	FW
2008	Zhang [36]	3.00 ($M \rightarrow M$) 2.70 ($M \rightarrow F$) 3.10 ($F \rightarrow M$) 2.80 ($F \rightarrow F$)	VQ
2008	Desai [37]	≈ 2.70	ANN
2009	Zhang [31]	2.70 ($F \rightarrow M$) 2.50 ($F \rightarrow F$)	VQ

Table 1. Experimental Results for Quality *MOS* in Voice Conversion systems.

These tests have all been made using source and target speakers of the same language. Some MOS results for crosslingual conversion between English and Spanish have been reported by Duxans [6] in 2006. In his study, *MOS* gradings for quality were 2.37 for Spanish-to-Spanish conversion and 2.33 for Spanish-to-English conversion, and similarity MOS gradings were 3.18 (Spanish-to-Spanish) and 2.79 (Spanish-to-English).

Year	**Author**	**Similarity *MOS***	**Technique**
2003	Rentzos [28]	3.65	HMM
2006	Nurminen [14]	3.10 ($F \rightarrow F$) 3.05 ($F \rightarrow M$) 2.20 ($M \rightarrow F$) 1.77 ($M \rightarrow M$)	GMM
2006	Duxans [6]	3.18	GMM CART
2006	Rao [30]	2.92 ($M \rightarrow F$) 3.23 ($F \rightarrow M$)	WLI
2006	Shuang [15]	1.87 (UK English) 2.77 (CN Mandarin)	FW
2007	Dutoit [3]	2.77	FS
2007	Erro [13]	2.93 ($M \rightarrow M$) 3.27 ($M \rightarrow F$) 2.53 ($F \rightarrow M$) 3.00 ($F \rightarrow F$)	WFW
2008	Shuang [16]	2.20	FW
2008	Zhang [36]	2.20 ($M \rightarrow M$) 2.30 ($M \rightarrow F$) 2.50 ($F \rightarrow M$) 2.10 ($F \rightarrow F$)	VQ

Table 2. Experimental Results for Similarity *MOS* in Voice Conversion systems.

The *ABX* test is a two-alternative test that is often used in comparing similarity between converted and target sentences. In this test, experimental subjects have to decide whether a given sentence X is closer in vocal quality to one of a pair of sentences A and B, where one of them is the source and the other is the target, not necessarily in that order. Success is measured by the percentage of answers of the type $X \approx T$ where $T \in \{A, B\}$ is the target.

Table 3 shows a set of *ABX* results of several voice conversion systems, as reported by their authors. Among these, a method by Fujii [4] stands out with extremely high scores. The main problem with interpreting *ABX* scores is the fact that subjects are not allowed to answer that the sentence X is not similar to neither A nor B, if that is the case. It can be inferred that a method that is successful according to an *ABX* test might in fact have a very low similarity-*MOS* value, as long as the similarity of the X sentences to their respective source speakers were even lower.

Another use of the *ABX* test is the comparison of two different techniques applied to the same problem. In this setting X is the target sentence, and A and B are the converted sentences using both techniques. Subjects are required to answer which of the sentences is closer to the target, but here the subject is allowed to answer “neither”. Success rates are computed for each technique as the percentage of sentences for which that technique has been chosen as closer to the target.

Among authors using this type of test are Pozo [42] and Desai [37]. Pozo compared his *Joint Estimation Analysis-Synthesis (JEAS)* method to the *Pitch-Synchronous Harmonic Model (PSHM)*, obtaining the following success rates: 41% ($M \rightarrow M$), 37% ($M \rightarrow F$), 33% ($F \rightarrow M$) and 36.5% ($F \rightarrow F$). Desai [37] compared his method using ANN to traditional GMM methods, obtaining an ABX

Year	Author	ABX Index	Technique
1998	Kain [10]	52.5% ($M \rightarrow M$) 97.5% ($M \rightarrow F$)	GMM
1998	Stylianou [9]	97%	GMM
1999	Arslan [24]	78% ($M \rightarrow M$) 100% ($M \rightarrow F$)	STASC
2001	Toda [25]	$\approx$ 77% ($M \rightarrow M$) $\approx$ 83% ($F \rightarrow F$)	GMM DFW
2004	Orphanidou [38]	79.5% ($M \rightarrow M$) 86.3% ($M \rightarrow F$) 88.6% ($F \rightarrow M$) 77.3% ($F \rightarrow F$)	RBFNN
2005	Toda [35]	$\approx$ 84% ($M \leftrightarrow F$)	MLE
2005	Zhang [39]	87.5%	GMM
2006	Ye e Young [29]	91.8%	GMM
2007	Fujii [4]	100% ($M \rightarrow M$) 100% ($M \rightarrow F$) 100% ($F \rightarrow M$) 98.0% ($F \rightarrow F$)	US
2007	Hanzlicek [40]	87.2% ($F \rightarrow M$) 70.8% ($F \rightarrow F$)	GMM
2008	Yue [41]	92.0%	GMM HMM
2008	Zhang [36]	$\approx$ 62% ($M \rightarrow M$) $\approx$ 80.5% ($M \rightarrow F$) $\approx$ 78.5% ($F \rightarrow M$) $\approx$ 55% ($F \rightarrow F$)	VQ
2009	Zhang [31]	68% ($M \rightarrow F$) 84% ($F \rightarrow F$)	VQ

Table 3. Experimental Results for *ABX* indices in Voice Conversion systems.

success rate for similarity of 65.0%. Likewise, Türk [26] compared his *Subband* conversion to *Full-band* methods, reporting an *ABX* index of 92.9% for similarity.

Comparing empirical results such as those reported in this section without considering the details of the experimental settings makes little (if any) sense. There are many concurrent factors that can significantly influence the outcome of an experiment, such as the number of sentences, number of subjects, subject listening sensitivity, original audio quality, unambiguity of the questions, among many others.

For instance, experimental settings must be carefully defined and questions must be carefully explained in order to obtain consistent experimental data. The description of the experiment should be detailed enough so as to enable independent replication of experimental results. Also, statistical analysis should be taken seriously, including hypothesis testing and measured significance levels, in order to derive statistically significant conclusions.

Another related difficulty is the lack of a standard database for voice conversion. For instance, Zhang [39] uses the *MSRA Mandarin Database*, Toda [35] uses the *MOCHA Database*, Ye e Young [29] uses the *VOICES Database*, whereas Türk [26] and Guido [43] use the *TIMIT Database*, which is probably the most popular choice.

These observations reinforce the need for some kind of *Benchmark* for subjective experimental evaluations of voice conversion systems, enabling researchers to set up similar experiments and to obtain comparable experimental data. This would benefit the community by helping to correctly identify advantages and limitations of each technique, which are the subject of the next section.

4. LIMITATIONS AND CHALLENGES

There are many open problems in voice conversion, which have been identified in several previous articles:

Phonetic Issues: in crosslingual voice conversion it is well-known that many phonemes in the source language may not exist in the target language. Bilingual subjects have been used [8] in order to derive phonetic transformations that allow similar but not identical phonemes to be converted between languages. Whether such transformations might be successfully used in other (not bilingual) subjects is an open question.

Prosody Issues: there are many global acoustic aspects that are decisive in order to obtain good conversions, such as average pitch and standard pitch deviation, average and standard deviation of energy, statistics related to the rhythmic flow of speech, and so on. Some (but not all) of these issues are discussed in Helander [44] and Hanzlicek [45].

Quality Issues: some of the problems that are perceived as a lack of quality are hissing noise, ringing tones, clicks and also timbral aspects that may be described as a synthetic or unnatural voice. For instance, large pitch shifts without formant correction may degrade quality (and even intelligibility) of the converted voice. These issues have been reported many times [9, 10, 4] and are easily detected in subjective tests.

Similarity Issues: these are related to the timbral quality and vocal identity, mainly correlated to phonetic aspects of speech, although they may easily be confused with quality and prosody issues in experimental tests. In theory, a purely synthetic voice might be perceived as unnatural but similar in timbre to a target voice. Intergender voice conversion is particularly susceptible to this type of problem [10, 23]

Evaluation Issues: this has been discussed in the previous section. Objective measures such as spectral distance or cepstral distortion may be uncorrelated to human perceptual measures [33, 26], whereas subjective measures such as *MOS* or *ABX* may be useful if some sort of experimental benchmark is agreed upon.

Excessive Smoothing Issues: this is a technical issue caused by interpolation methods in the transformation phase, which degrade the spectrum by eliminating details and reducing the similarity of target and converted voices [35, 46, 47].

Overfitting Issues: this is a counterpart of the previous issue, and is caused by using excessive data in training and obtaining an excessively fine-grained transformation which might produce discontinuities between adjacent frames [47].

5. CONCLUSION

This text has presented the voice conversion problem and discussed some of its application contexts, such as TTS customization [10, 6] and virtual interpreters [11, 12]. Major contributions in the recent literature, as well as comparative results, have also been presented and discussed.

High-level representation models for voice signals are a critical aspect of any voice conversion system, since they define and also constrain the available transformation techniques. Aspects such as timbre, prosody and intelligibility should all be taken into account for better results in terms of naturalness and fluency of virtual interpreters and of customized TTS systems [10, 6]. The challenge of crosslingual voice conversion has brought interest to studies of phonetic similarities and differences across languages and automatic phonetic transformation, specially resorting to bilingual individuals [19, 8].

The choice of training model for a voice conversion system usually depend on specific requirements (for instance, attempts at breaking a voice security system may require text-dependent training in order to minimize conversion errors) or on availability of data (for instance, if crosslingual conversion between non-bilingual subjects is desired, then text-independent training is the only available option).

Transformation techniques should be considered not only in relation to compatible representation models, but also with respect to the prosodic and timbral aspects that will be converted, since they define how a voice conversion system views and manipulates such high-level representation data.

Among the open research problems, the definition of a benchmark for the subjective comparison of voice conversion systems for quality and similarity assessment seems to be one the most urgent issues. Some progress has already been made in this respect through the TC-STAR project [11, 12], but a more thorough specification of reproducible experiments is desirable.

6. REFERENCES

[1] P. R. Cook, "Singing voice synthesis: History, current work, and future directions," *Computer Music Journal*, vol. 20, pp. 38–46, 1996.

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

[3] T. Dutoit, A. Holzapfel, M. Jottrand, A. Moinet, J. Perez, and Y. Stylianou, "Towards a voice conversion system based on frame selection," in *ICASSP*, pp. 513–516, 2007.

[4] K. Fujii, J. Okawa, and K. Suigetsu, "High-individuality voice conversion based on concatenative speech synthesis," *IJCSE*, vol. 2, p. 1, 2007.

[5] H. Pfitzinger, "Unsupervised speech morphing between utterances of any speakers," in *SST 2004*, pp. 545–550, 2004.

[6] 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.

[7] D. Sündermann, H. Ney, and H. Hoge, "VTLN-based cross-language voice conversion," in *ASRU*, 2003.

[8] M. Mashimo, T. Toda, H. Kawanami, H. Kashioka, K. Shikano, and N. Campbell, "Evaluation of cross-language voice conversion using bilingual e non-bilingual databases," *Interspeech*, pp. 293–296, 2002.

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

[10] A. Kain and M. Macon, "Text-to-speech voice adaptation from sparse training data," in *5th ICSLP*, 1998.

[11] D. Sündermann, H. Höge, A. Bonafonte, H. Ney, and J. Hirschberg, "TC-STAR: Cross-language voice conversion revisited," *TC-STAR WSST*, pp. 231–236, 2006.

[12] A. Bonafonte, H. Höge, I. Kiss, A. Moreno, U. Ziegenhain, H. van den Heuvel, H. Hain, X. Wang, and M. Garcia, "TC-STAR: Specifications of language resources and evaluation for speech synthesis," in *LREC*, 2006.

[13] D. Erro and A. Moreno, "Weighted frequency warping for voice conversion," in *Interspeech*, 2007.

[14] J. Nurminen, V. Popa, J. Tian, Y. Tang, and I. Kiss, "A parametric approach for voice conversion," in *TC-STAR WSST*, pp. 225–229, 2006.

[15] Z. Shuang, R. Bakis, and Y. Qin, "Voice conversion based on mapping formants," in *TC-STAR WSST*, pp. 219–223, 2006.

[16] Z. Shuang, R. Bakis, and Y. Qin, "IBM voice conversion systems for 2007 TC-STAR evaluation," *Tsinghua Science & Technology*, vol. 13, no. 4, pp. 510–514, 2008.

[17] D. Sündermann, H. Hoge, A. Bonafonte, H. Ney, A. Black, and S. Narayanan, "Text-independent voice conversion based on unit selection," in *ICASSP*, 2006.

[18] H. Kawahara and T. Irino, "Exploring temporal feature representations of speech using neural networks," Technical Report SP88-31 (in Japanese), IEICE, Tokyo, 1988.

[19] M. Abe, K. Shikano, and H. Kuwabara, "Cross-language voice conversion," in *ICASSP*, pp. 345–348, 1990.

[20] D. Childers, "Glottal source modeling for voice conversion," *Speech Communication*, vol. 16, pp. 127–138, 1995.

[21] K. Shikano, K. Lee, and R. Reddy, "Speaker adaptation through vector quantization," in *ICASSP*, vol. 11, 1986.

[22] H. Valbret, E. Moulines, and J. Tubach, “Voice tranformation using PSOLA technique,” in *2nd ECSCT*, 1991.

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

[24] L. Arslan, “Speaker transformation algorithm using segmental codebooks (STASC),” *Speech Communication*, vol. 28, no. 3, pp. 211–226, 1999.

[25] T. Toda, H. Saruwatari, and K. Shikano, “Voice conversion algorithm based on gaussian mixture model with dynamic frequency warping of straight spectrum,” *Power [dB]*, vol. 30, p. 40, 2001.

[26] O. Türk and L. Arslan, “Subband based voice conversion,” in *7th ICSLP*, 2002.

[27] T. Kamm, G. Andreou, and J. Cohen, “Vocal tract normalization in speech recognition: Compensating for systematic speaker variability,” in *15th ASRS*, 1995.

[28] D. Rentzos, S. Vaseghi, E. Turajlic, Q. Yan, and C. Ho, “Transformation of speaker characteristics for voice conversion,” in *IEEE WASRU*, pp. 706–711, 2003.

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

[30] K. Rao and B. Yegnanarayana, “Voice conversion by prosody and vocal tract modification,” in *9th ICIT*, pp. 111–116, 2006.

[31] M. Zhang, J. Tao, J. Nurminen, J. Tian, and X. Wang, “Phoneme cluster based state mapping for text-independent voice conversion,” in *ICASSP*, pp. 4281–4284, 2009.

[32] F. Nolan, “Intonational equivalence: an experimental evaluation of pitch scales,” in *ICPhS*, pp. 771–774, 2003.

[33] D. Sündermann, “Voice conversion: State-of-the-art and future work,” *FORTSCHRITTE DER AKUSTIK*, vol. 31, p. 735, 2005.

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

[35] T. Toda, A. Black, and K. Tokuda, “Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter,” in *ICASSP*, pp. 9–12, 2005.

[36] M. Zhang, J. Tao, J. Tian, and X. Wang, “Text-independent voice conversion based on state mapped codebook,” in *ICASSP*, 2008.

[37] S. Desai, E. Raghavendra, B. Yegnanarayana, A. Black, and K. Prahallad, “Voice conversion using artificial neural networks,” in *IEEE WSLT*, 2008.

[38] C. Orphanidou, I. Moroz, and S. Roberts, “Wavelet-based voice morphing,” *WSEAS Transactions on Systems*, vol. 3, no. 10, pp. 3297–3302, 2004.

[39] J. Zhang, J. Sun, and B. Dai, “Voice conversion based on weighted least squares estimation criterion and residual prediction from pitch contour,” *Lecture notes in computer science*, vol. 3784, p. 326, 2005.

[40] Z. Hanzlíček and J. Matoušek, “Voice conversion based on probabilistic parameter transformation and extended inter-speaker residual prediction,” *Lecture Notes in Artificial Intelligence*, vol. 4629, pp. 480–487, 2007.

[41] Z. Yue, X. Zou, Y. Jia, and H. Wang, “Voice conversion using HMM combined with GMM,” in *CISP’08*, vol. 5, 2008.

[42] A. del Pozo and S. Young, “The linear transformation of LF glottal waveforms for voice conversion,” in *Interspeech*, pp. 1457–1460, 2008.

[43] R. Guido, L. Sasso Vieira, S. Barbon Június, F. Sanchez, C. Dias Maciel, E. Silva Fonseca, and J. Carlos Pereira, “A neural-wavelet architecture for voice conversion,” *Neurocomputing*, vol. 71, no. 1-3, pp. 174–180, 2007.

[44] E. Helander and J. Nurminen, “A novel method for prosody prediction in voice conversion,” in *ICASSP*, vol. 4, pp. 509–512, 2007.

[45] Z. Hanzlicek and J. Matousek, “F0 transformation within the voice conversion framework,” in *ICSLP*, pp. 1961–1964, 2007.

[46] T. Toda, A. Black, and K. Tokuda, “Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory,” *IEEE TASLP*, vol. 15, no. 8, p. 2222, 2007.

[47] L. Mesbahi, V. Barreaud, and O. Boeffard, “Comparing GMM-based speech transformation systems,” in *Interspeech*, pp. 1989–1992, 2007.

[48] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, “Voice conversion through vector quantization,” in *ICASSP*, pp. 655–658, 1988.

[49] O. Türk and M. Schröder, “A comparison of voice conversion methods for transforming voice quality in emotional speech synthesis,” *Interspeech*, 2008.

[50] H. de Paula, H. Yehia, D. Shiller, G. Jozan, K. Munhall, and E. Vatikiotis-Bateson, “Linking production and perception through spatial and temporal filtering of visible speech information,” in *6th ISSP*, pp. 37–42, 2003.

[51] L. Arslan and D. Talkin, “Voice conversion by codebook mapping of line spectral frequencies and excitation spectrum,” in *5th ECSCT*, 1997.

[52] Y. Stylianou, “Voice transformation: A survey,” *ICASSP*, pp. 3585–3588, 2009.