A flexible and modular crosslingual voice conversion system

Anderson Fraiha Machado
Universidade de São Paulo
dandy@ime.usp.br

Marcelo Queiroz
Universidade de São Paulo
mqz@ime.usp.br

ABSTRACT

A cross-lingual voice conversion system aims at modifying the timbral structure of recorded sentences from a source speaker, in order to obtain processed sentences which are perceived as the same sentences uttered by a target speaker. This work presents the cross-lingual voice conversion problem as a network of related sub-problems and discuss several techniques for solving each of these sub-problems, in the context of a modular implementation that facilitates comparisons between competing techniques. The implemented system aims at high-quality cross-lingual voice conversion in a text-independent setting, i.e. where the training sets of sentences recorded by source and target speakers are not the same. New strategies are introduced, such as artificial phonetic maps, N-likelihood clustering and normalized frequency warping, which are evaluated through numerical experiments.

1. INTRODUCTION

Voice conversion refers to the process of manipulation of acoustic parameters that transform sentences uttered by a *source speaker* into sentences that sound as if having been uttered by a *target speaker*, according to the original formulation of the problem by Childers et al. in 1985 [1]. The cross-lingual variant of the problem considers that source and target speakers are not required to speak the same language.

Among the applications of cross-lingual voice conversion a few are prominent, such as the personification of interactive systems with speech synthesis from text, so called *Text-To-Speech Systems (TTS)* [2] and personalized virtual interpreter systems [3, 4], which usually require subsystems for *speech recognition*, *text translation* and *speech synthesis*, possibly coupled with voice conversion.

A voice conversion system takes into account the timbre and prosody of the source and target speakers. These are easily recognizable perceptual aspects of speech which present difficulties in terms of definition and formalization in general terms. But in the specific context of voice conversion, timbre is usually considered in terms of the dynamic spectral envelope of the voice signal, whereas prosody is reduced to energy and melodic contours and rhythmic patterns of phonemes [5, 6, 7, 8].

Timbre and prosody transformations in a voice conversion system usually depend on a training stage, which may be text-dependent or text-independent. In the text-dependent setting source and target speakers are required to record the same sentences, which are time-aligned and corresponding segments are matched to create mappings from source to target acoustic parameter spaces. In the text-independent setting [9] the sentences uttered by source and target speakers are not necessarily the same; segments are mapped into an acoustical feature space and clustered according to *artificial phonetic classes*, which may or may not coincide with conventional phonetic classes of the corresponding language.

Artificial phonetic classes are one of the fundamental constructs that characterizes the representation of a speaker's identity within a voice conversion system, the so-called *corpus* of that speaker. Acoustic parameters of source sentences are then mapped to an artificial phonetic class before being transformed to a corresponding (artificial) phonetic class in the target acoustic feature space. The *corpora* of source and target speakers are then used to build a mapping between the acoustic feature spaces that allow the conversion of the meta-representations corresponding to artificial phonetic classes, which are ultimately used to render voice segments with the timbral and prosody qualities extracted from the corpus of the target speaker.

A distinctive aspect of voice conversion systems is related to the phonetic content of the languages used in the training and conversion stages. Cross-lingual voice conversion techniques usually presuppose that source and target speakers use different languages; although this is not a requirement, it guides the development of techniques that do not explore coincidences of the artificial phonetic classes obtained from source and target speakers. The central point of this work consists in performing a *genuine conversion* where each speaker uses her/his own language, in which case training is inevitably text-independent. It is important to emphasize that no kind of symbolic-textual processing (e.g. translation) is carried out in this type of conversion, which instead takes place entirely within the acoustic feature spaces, classes and clusters of speech segments that characterize both speakers.

Section 2 presents a short theoretical framework that underlies the main contributions of this work, which are in turn presented in Section 3, along with the complete cross-lingual voice conversion system. Experimental evaluation is reported in Section 4, and conclusions and further work

are delineated in Section 5.

2. THE FRAMEWORK OF THE SYSTEM

The success in the development of a cross-lingual voice conversion system depends intrinsically on the choice of reliable tools for the specific sub-tasks that comprise the conversion process.

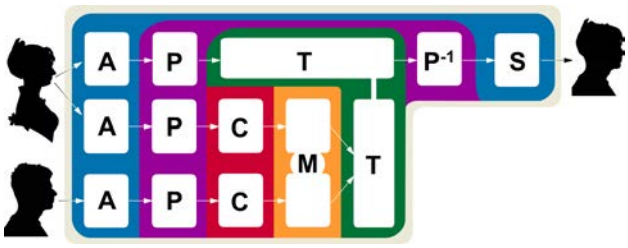

Figure 1. Functional modules comprising the voice conversion system.

Figure 1 exhibits a general diagram, in which data flows for both speakers are identified. The upper lines start with the flow of sentences uttered by a source speaker (silhouette at the upper-left corner) and the lower line starts with target speaker (lower-left corner silhouette) sentences. Central and lower lines correspond to a training stage, fed by both speakers, whereas the upper line represents the conversion stage, where the input is formed only by sentences of the source speaker; the converted sentences (i.e. with message content of the source speaker and timbral identity of the target speaker) are produced at the upper-right end of the diagram. Sub-tasks are denoted by capital letters, according to the functional module, as follows:

[A] The **A**nalysis module takes a recorded sentence, divides it in short quasi-stationary segments and builds a representation of each segment according to a voice production model;

[P] The **P**arameterization module is responsible for modelling and quantization of the acoustic parameters;

[C] The **C**lustering module is responsible for organizing the *corpus* of each speaker into acoustic classes, i.e. modelling the acoustic feature space corresponding to each speaker;

[M] The **M**apping module associates acoustic classes between speakers, using optimized alignment techniques;

[T] The **T**ransformation module contains the acoustic parameter conversion functions, estimated from the alignment of acoustic classes;

[P^{-1}] The inverse **P**arameterization module readapts transformed (quantized) acoustic parameters to the analysis / synthesis model adopted; and

[S] The **S**ynthesis module assembles the output signal incorporating the transformed acoustic parameters.

These functional modules comprise the two-stage voice conversion system. The lower part of Figure 1 shows the sequences **A** $\rightarrow$**P** $\rightarrow$**C** $\rightarrow$**M** $\rightarrow$**T** for both speakers, corresponding to the training stage, whereas the upper part, whose sequence corresponds to **A** $\rightarrow$**P** $\rightarrow$**T** $\rightarrow$$\mathbf{P}^{-1}$ $\rightarrow$**S**, is the conversion stage. These modules are discussed in the sequel.

Conceptually, one may interpret the recorded voice signal as the result of an excitation source, the glottal pulse, which is modified by the vocal tract; this interpretation applies only to voiced sounds, e.g. vowels, whereas unvoiced sounds must be considered separately. Under a linearity hypothesis, the vocal tract is modelled as a filter whose transfer function modifies the spectral content of the glottal pulse. The harmonic part of this spectrum is modelled as a train of pulses with corresponding amplitudes; usually a stochastic component is added to the model in order to simulate the effect of the air column which freely traverses the phonatory system.

The vocal tract filter may be represented with few coefficients, which will describe the profile of the filter response or the *spectral envelope* of each quasi-stationary segment. Thus it is possible to characterize both harmonic and stochastic components through their respective spectral envelopes in acoustic feature vectors, for ulterior parametric modelling and transformation. Spectral envelopes may be obtained through LPC coefficients [10], Cepstrum coefficients [11] and interpolation methods [12]. The *STRAIGHT* [13] technique also uses a glottal pulse/vocal tract model, allowing manipulation of parameters such as the spectral envelope, pitch and other acoustic parameters.

Given a representation of voice segments through acoustic feature vectors immersed in an acoustic feature space, similar voice segments are clustered around a representative (usually the centroid of a set of similar vectors) to define an Artificial Phonetic Class (or APC for short). Two classes of clustering techniques are of widespread appearance in voice conversion literature: cognitive network methods and statistical methods, more prominently Gaussian Mixture Models or GMMs [14] which are considered state-of-the-art in voice conversion. A new clustering approach is introduced here, which normalizes classes in order to eliminate discrepancies between APCs of different languages.

A model of acoustic mapping between APCs is essential for the transformation module. The alignment of APCs of different speakers is based on the assumption that transposition of phonetic characteristics is performed locally, on the voice segment level, without taking the phonetic context of adjacent segments into account; naturally this assumption applies specifically to the APC alignment problem, not to the whole conversion process. A solution to APC alignment is proposed here which adapts a classical graph-theoretical algorithm for minimal cost graph node matching [15].

The alignment uses APC representatives from both speakers, from which transfer functions are derived (in the training stage). These transfer functions are used in the conversion stage to transform acoustic parameters of a segment belonging to an APC of the source speaker to acoustic parameters that correspond to a matched APC in the target speaker model. These parameters, now in the context of the target speaker acoustic feature space, are synthesized in the last module into a voice segment that will compose the converted sentence.

The technical aspects of these modules are detailed in the

following section.

3. NEW TOOLS FOR VOICE CONVERSION

According to the structure of the system, the Analysis and Synthesis modules correspond to the first stage, which should provide a sensible representation for voice segments (analysis) that will allow a truthful reconstruction of the converted voice signal (synthesis).

3.1 Module I: Harmonic-Stochastic Modelling (HSM)

The Harmonic+Noise model proposed by Stylianou [16] is a flexible model that offers a representation framework for high-fidelity voice reconstruction. It is efficient and transparent, representing harmonic and stochastic components, which simplifies manipulation and transformation of voice segments. The harmonic components correspond to a quasi-periodic portion of the signal, and are represented through individual frequencies, amplitudes and initial phases of sinusoidal oscillators. The stochastic component corresponds to the rest of the signal with essentially aperiodic behaviour, and is modelled as a source-filter system with white noise as source and LPC coefficients to define the spectral envelope of the filter.

There are several models for speech synthesis which completely discard all phase information and try to rebuild phases a posteriori from the magnitude spectrum alone, using spectral modelling under the hypothesis of minimum phase [17]. Instead of discarding phase information, an alternative representation for the phases of the harmonic components is here introduced, in order to provide for a more robust transformation which is less prone to phase distortion. Consider the harmonic part s_h obtained from the HSM decomposition with $L+1$ sinusoidal components of an arbitrary voice segment, indexed as k, of length $N+1$ (N even) and with samples n indexed from $-\frac{N}{2}$ to $+\frac{N}{2}$ and sampling frequency R:

$$s_h^{(k)}[n] = \sum_{m=0}^{L} A_m^k \cos\left(\frac{2\pi m f_0^k n}{R} + \varphi_m^k\right); \qquad (1)$$

it is considered that the central samples of each segment are $\frac{N}{2}$ samples apart from each other. The overlap rate is $\frac{\frac{N}{2}}{N+1}$, slightly less than 50%.

The phase φ_m^k of the m-th harmonic component in Equation (1) refers to the sample $n = 0$ at the center of the segment, and can be seen as a sum of two components, the first that corresponds to the predicted phase propagated from segment $k-1$, and a second component that accounts for the difference between this prediction and the actual phase obtained in the analysis of segment k:

$$\varphi_m^k = \left(\varphi_m^{k-1} + \frac{2\pi m f_0^{k-1}\frac{N}{2}}{R}\right) + \eta_m^k. \qquad (2)$$

This η_m^k is actually *defined* by the above equation, and will be used to redefine the frequencies of the model in a manner similar to the well-known phase vocoder method.

Rewriting $f_m^{k-1} = m f_0^{k-1}$ and refactoring,

$$\varphi_m^k = \varphi_m^{k-1} + \frac{2\pi\left(f_m^{k-1} + \frac{R\eta_m^k}{\pi N}\right)\frac{N}{2}}{R} \qquad (3)$$

which corresponds to

$$\varphi_m^k = \varphi_m^{k-1} + \frac{2\pi \hat{f}_m^{k-1}\frac{N}{2}}{R}, \qquad (4)$$

where

$$\hat{f}_m^{k-1} = f_m^{k-1} + \frac{R\eta_m^k}{\pi N}. \qquad (5)$$

This process uses information obtained from the HSM model of segment k to redefine the frequencies of segment $k-1$, and as a consequence the phases for segment k become unnecessary, since they can be recovered from the phases of segment $k-1$ by Equation (4). By iterating this mechanism, only the initial phases are required to reconstruct the whole sequence of segments corresponding to a single voiced emission. The reconstruction based on the above recodification is thus

$$\hat{s}_h^k[n] = \sum_{m=0}^{L} A_m^k \cos\left(\frac{2\pi \hat{f}_m^k n}{R} + \varphi_m^k\right), \qquad (6)$$

which is evidently different from the original segment, but nevertheless of enormous utility in this context, because it has no phase discontinuities between overlapped segments. This representation might be used to reconstruct a perceptually similar signal, but most importantly it allows several transformations of the signal, such as pitch shift and time stretch, to be performed easily.

The recurrence relation involving phases and re-estimated frequencies poses an initial value problem: in principle, the first segment $k = 0$ could be represented explicitly through the phases φ_m^0 of the HSM analysis; but since phases are going to be propagated according to the sequence of re-estimated frequencies, it is reasonable to drop the initial phases for the first segment, and use $\varphi_m^0 = 0$ in the reconstruction.

It should be noted that this recodification takes this model away from the original HSM premises, because the re-estimated frequencies $\{\hat{f}_m^k,\ m = 1, \ldots, L\}$ are no longer guaranteed to be in harmonic relation. Instead, the model becomes a mixture of the additive synthesis and stochastic models. It should also be noted that in pitch-shifted segments phase coherency between inharmonic components is lost, but the resulting artifacts can be masked in the overlap-add reconstruction.

3.2 Module II: Parametric Decomposition

The Parameterization module normalizes the acoustic parameters using a fixed number of coefficients in order to characterize segments, creating a correspondence of values that allows the direct comparison of segments within a normalized feature space.

The central idea of this module is to provide two log-magnitude spectral envelopes, the harmonic and stochastic

spectral envelopes, and decompose them into sums of radial basis functions (bandpass filters) with flexible shapes and bandwidths. Such spectral envelopes can be easily estimated using classical methods such as LPC and Cepstrum, among others [10], and the radial basis functions decomposition follows the strategy presented in [12], by constructing sub-bands E_k and initializing basis parameters conveniently.

The sum of L general parametric functions is denoted by

$$\hat{E}(f; a, \mu, \sigma) = \sum_{m=1}^{L} \psi(f; a_m, \mu_m, \sigma_m) \qquad (7)$$

where each component $\psi(f; a_m, \mu_m, \sigma_m) : \Re \longrightarrow \Re$ corresponds to a continuous basis function with amplitude a_m, central frequency μ_m and bandwidth σ_m, evaluated in the frequency range $f = [0, R/2]$.

The problem of radial basis function modelling is to find an approximation $\hat{E}(f; a, \mu, \sigma)$ of a discrete function $E(f)$ such that the error of estimation is minimal. The fitting of the given function $E(f)$ to the parametric model $\hat{E}(f; a, \mu, \sigma) \approx E(f)$ could in principle be obtained by any iterative optimization method (e.g. gradient descent or Newton) to obtain the least-squares solution that minimizes the error of the model. Other methods for parametric decomposition can also be found in the literature [18, 12].

3.3 Module III: Data Clusterization

The Clusterization module groups segments according to similarity in the acoustic feature space, which will later define the artificial acoustic classes in the phonetic map of the speaker corpus. First, acoustic vectors with harmonic component amplitudes are quantized and then grouped, based on a likelihood criterion, for further analysis.

3.3.1 N-likelihood clustering

The N-likelihood clusterization method introduced here groups vectors in an acoustic feature space in clusters with N elements (typically between 5 and 10, depending on window length). Each of these clusters is obtained by minimizing the dispersion of a vector set, measured as the norm of the diagonal of the covariance matrix of the vector set.

For each segment k with an harmonic amplitude vector A^k, a temporally-smoothed version v^k is defined as

$$v^k = \left[\frac{1}{4}A^{k-1} + \frac{1}{2}A^k + \frac{1}{4}A^{k+1}\right]. \qquad (8)$$

Then for each pair (v^m, v^n) a similarity matrix E is built using the euclidean distance $E(m, n) = \|v^m - v^n\|$, and for each line m of this matrix the indices of the N smaller values are selected in a sub-matrix J in such a way that $J(m, 1 \dots N)$ are the N smallest indices in $\{E(m, n),\ \forall n\}$. Then a selection structure $K(m)$ is defined as

$$K(m) = \|\text{diag}\,(\text{Cov}\{v_k \mid k \in J(m, 1 \dots N)\})\|. \qquad (9)$$

In its final stage the clusterization algorithm loops through the selection structure K searching for the index m^* with the smallest value, in order to define an N-likelihood cluster as the set $\upsilon = \{v^i \mid i \in I\}$ where $I = \{J(m^*, 1 \dots N)\}$. Then all indices $i \in I$ are removed from the selection structure K (e.g. set $K(i) = +\infty,\ \forall i \in I$). This process is repeated until the selection structure is emptied (e.g $K(i) = +\infty,\ \forall i$). After that, an optional filtering is performed to eliminate clusters with a relatively high dispersion measure.

3.3.2 Artificial Phonetic Map

In order to establish a starting point for the clusterization of voiced segments in artificial phonetic classes, the use of a discrete polygon which simulates a phonetic map is here proposed. This polygon is a trapeze where each point corresponds to a pair of formantic centers (f_1, f_2), obtained from a (mel-frequency) spectral envelope c_k corresponding to a centroid of an N-likelihood cluster υ_k.

The process of estimation of each point, or formantic center, (f_1, f_2) in the artificial phonetic map from a centroid c_k considers several possible candidates $(\hat{f}_1, \hat{f}_2)$ with $\hat{f}_1 \leq \hat{f}_2$, $\hat{f}_1 \in F_1 = \{\hat{f}_1^1, \hat{f}_1^2, \dots, \hat{f}_1^r\}$ and $\hat{f}_2 \in F_2 = \{\hat{f}_2^1, \hat{f}_2^2, \dots, \hat{f}_2^r\}$. For any candidate $\hat{f}$ a triangular bandpass filter $\xi_{\hat{f}}$ is built, with center $\hat{f}$ and a bandwidth ranging from 300 mels to 1800 mels, adjusted so as to not overlap neighbouring formantic centers in the spectrum. After filtering (as $\tilde{c}_k^{\hat{f}} = c_k^t * \xi_{\hat{f}}$) the centroid c_k through all candidate filters $\xi_{\hat{f}}$ with $\hat{f} \in F_1 \cup F_2$, the two spectra $\tilde{c}_k^{f_1^*}$ and $\tilde{c}_k^{f_2^*}$ with maximum total cumulative energy are selected.

The pair of formants (f_1^*, f_2^*) thus built is the phonetic map entry that corresponds to the N-likelihood cluster υ_k. The phonetic map can be seen as a kind of hash table indexed by (f_1^*, f_2^*), giving access to all the acoustic vectors $v \in \upsilon_k$ belonging to the same N-likelihood cluster. This grouping of acoustic vectors $v \in \upsilon_k$ indexed by a pair (f_1^*, f_2^*) corresponds to an artificial phonetic class (APC), which will be denoted by $\mathbb{C}(f_1^*, f_2^*)$. For each APC, the means and covariance matrices of the elements that compose each set are obtained, which will be used in the transformation module.

3.4 Module IV: Alignment of Non-Parallel Corpora

The goal of this module is to create a canonical and unified representation for both source and target corpora, in a way that an alignment between them is established even in the case of distinct languages. A normalization of the phonetic maps is realized in order to diminish the linguistic differences of the speakers, by warping the acoustic map around the origin and with a normalized dispersion in all directions. To weigh the contributions of each class $\mathbb{C}(f_1, f_2)$ the cardinality $|\mathbb{C}(f_1, f_2)|$ is considered; classes $\mathbb{C}(f_1, f_2)$ with few elements are excluded from the phonetic map.

Given two sets of artificial phonetic classes (or phonetic maps) from corpora $\mathcal{C}^X$ and $\mathcal{C}^Y$ represented by the respective sets of normalized switches $\bar{\mathbb{C}}^X$ and $\bar{\mathbb{C}}^Y$, it would be convenient to try to define a bijection mapping $M : \bar{\mathbb{C}}^X \rightarrow$

$\bar{\mathbb{C}}^Y$ as

$$M(x_i) = \arg \min_{y_i \in \bar{\mathbb{C}}^Y} \left\{ \sum_{x_i \in \bar{\mathbb{C}}^X} \mathbf{d}(x_i, y_i) \right\}, \quad (10)$$

where $\mathbf{d}$ corresponds to the euclidean distance. But an important issue in the alignment of corpora is the fact that the cardinalities of the sets $\bar{\mathbb{C}}^X$ and $\bar{\mathbb{C}}^Y$ are typically different, making it impossible to define such a bijection. To resolve this issue, the largest of the two sets will be shrunk in order to force a correspondence between the remaining classes of the two corpora. This reduction is performed at the end of the pairing algorithm, when all unpaired classes are excluded from the database.

The *Optimal Acoustic Mapping Problem* between phonetic classes $\bar{\mathbb{C}}^X$ and $\bar{\mathbb{C}}^Y$ is then reduced to the *Minimum Cost Perfect Matching Problem* in bipartite graphs. Two vertices x_i and y_j are paired in the acoustic mapping if, and only if, the edge $a_{i,j}$ belongs to the perfect matching. This problem can be efficiently solved by the classical algorithm of *Hopcroft and Karp* [15] in time $\mathcal{O}(|E|\sqrt{|V|})$ where V and E are the sets of vertices and edges, respectively.

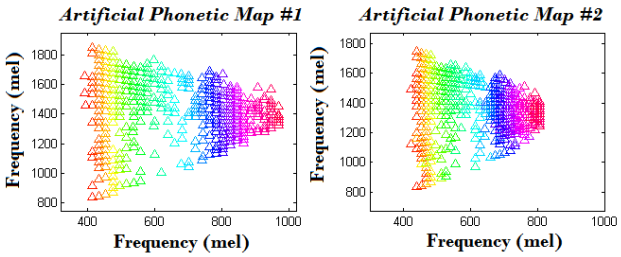

Figure 2. The perfect matching method applied to the alignment of parallel sentences.

Figure 2 illustrates the result of matching between two different corpora using a color code for aligned data. Given a sentence uttered by speaker #1, the correspondence between the classes of corpus #1 and corpus #2 (with the same color) allow the definition of a (piecewise) linear mapping function that warps the spectral contents of the sentence to adjust it to the map #2. Using the above alignment, a new technique of transformation of acoustic parameters is presented in the sequel.

3.5 Module V: Transformation of the Acoustic Parameters

Two types of transformations are defined at this stage: (1) a global transformation which corresponds to a linear transform based on means and variances of global prosodic parameters; and (2) a local transformation for the spectral conversion of each input segment using a mapping between the artificial phonetic classes developed in the prior module. Prosody conversion transposes pitch (F_0) and energy (E_0) contours of the harmonic components, and pure energy contours of the stochastic components, into corresponding values obtained from the pitch and energy models of the target corpus. Local spectral conversion applies only to harmonic components, and reshape spectral content of input segments, conforming them to the target's phonetic map.

Each input segment is dynamically associated to a set of probable APCs $\mathbb{C}_j^X$ in the source corpus $\mathcal{C}^X$ through a pertinence criterion applied to the harmonic components in this segment's representation. The composition of weighted local transformation is commonly used in GMM-based linear transformations, since it avoids discontinuity when a sequence of segments transits between APCs [19]. Let $\mathbb{C}_i^Y = M(\mathbb{C}_i^X)$ be an APC in the target corpus $\mathcal{C}^Y$ corresponding to the APC $\mathbb{C}_i^X$ in the source corpus $\mathcal{C}^X$. The local (segment-wise) transformation uses the means and covariance matrices of the corresponding APCs in order to build a segment in the target acoustic vector space.

The pertinence criterion adopted is based on the mel-cepstral distortion [20]; i.e. the pertinence of vector v_k with respect to centroid c of class $\mathbb{C}^X$ is defined as

$$\mathbf{d}(v, c) = 1 - \frac{\|\text{dct}[v] - \text{dct}[c]\|}{\sum_{\forall c_k \in \bar{\mathbb{C}}^X} \|\text{dct}[v] - \text{dct}[c_k]\|}, \quad (11)$$

where dct is the well known Discrete Cosine Transform. Then a set of m probable APCs $c_i,\ i \in [1, m]$ with highest pertinence values $\mathbf{d}(v, c_i)$ is taken. The weighted transformation is defined as

$$\hat{v} = \frac{\sum_{i=1}^{m} \mathbf{d}(v, c_i) \mathcal{T}_{loc}(v, c_i, M(c_i))}{\sum_{i=1}^{m} \mathbf{d}(v, c_i)}, \quad (12)$$

where $\mathcal{T}_{loc}$ is a chosen segment conversion function and M is the mapping between phonetic maps obtained in module IV. One example of a segment conversion function is the linear transform using full covariance matrices of each APC involved, defined as

$$\mathcal{T}_{loc}^{LT}(v, c, M(c)) = \mu_{M(c)} + \Sigma_{M(c)}^{0.5} (\Sigma_c^{-0.5})(v - \mu_c), \quad (13)$$

where μ_c and Σ_c are the mean and covariance matrix of the vectors in class c.

Another segment conversion function introduced here is the Normalized Frequency Warping (NFW). The first step in this method is obtaining *Cumulative Distribution* (CD) functions of the (parametric) spectral envelope $E(f) = \sum_{\forall k} \psi(f, a_k, \mu_k, \sigma_k),\ \forall f \in [f_{\min}, f_{\max}]$ (of module II), defined as

$$\text{CD}_E(f) = \frac{\sum_{k=f_{\min}}^{f} \{E(k)\} - E(f_{\min})}{\sum_{k=f_{\min}}^{f_{\max}} \{E(k)\} - E(f_{\min})}, \quad (14)$$

where $f_{\min}$ and $f_{\max}$ are the minimum and maximum frequencies considered in the parametric decomposition of the spectral envelope. Then a Normalized Frequency Distribution (NFD) is obtained, which associates to each normalized amplitude value $e \in [0, 1]$ a frequency f in such a way that

$$\text{DFN}(e) = \frac{\sum_{k=e_0}^{e} \text{CD}(k)}{\sum_{k=e_0}^{e_1} \text{CD}(k)} (f_{\max} - f_{\min}) + f_{\min}. \quad (15)$$

Values between e_0 and e_1 are linearly-spaced amplitudes between 0 and 1 according to the discretized version of the

spectral envelope E. The Normalized Frequency Warping (NFW) then transforms a vector v into $\hat{v}$ based on the alignment of DFN values for source and target spectra representing the phonetic classes c_i and $M(c_i)$.

The NFW method is described through the following algorithm:

1. The centroids c_i and $M(c_i)$ of the APCs $\mathbb{C}_i^X$ and $\mathbb{C}_i^Y = M(\mathbb{C}_i^X)$ are used as input for the definition of a local tranformation;

2. The DFNs $\text{DFN}_{c_i}(f)$ and $\text{DFN}_{M(c_i)}(f)$ of c_i and $M(c_i)$ are obtained;

3. A warping pattern $NFW(v, c_i, M(c_i))$ is defined as the cubic splines interpolation of the pairs $(\text{DFN}_{c_i}(f), \text{DFN}_{M(c_i)}(f))$, for all (discretized) frequencies in the input vector v;

Finally, the above warping is applied to the difference between the input vector v and the centroid c_i, as

$$\mathcal{T}_{loc}^{\text{NFW}}(v, c_i, M(c_i)) = M(c_i) + NFW\,(v - c_i, c_i, M(c_i))\,. \quad (16)$$

The voice conversion system containing the above strategies has been subjected to objective and subjective experimental evaluations, which are reported in the next section.

4. EXPERIMENTAL RESULTS

Two fundamental questions are related to voice conversion evaluation: (1) timbre *similarity* between the target speaker voice and the converted sentence; and (2) *quality* of the transformed signal, comprising sound quality aspects such as intelligibility and naturalness. This section includes an objective and a subjective evaluation of these questions in the proposed system.

This research has had the support of the Universitat Politècnica de Catalunia, which provided a database used in the TC-STAR project [4]. Sentences in this database had been recorded by several speakers using time cues to facilitate alignment, and with a rigorous prosody control achieved by imitation, i.e. each speaker had an example sentence whose prosody should be reproduced as closely as possible. Furthermore, all recruited speakers were bilingual (English and Spanish), providing for examples of each timbral identity in both phonetic spaces.

Eight speaker corpora form the TC-STAR database, categorized by language and gender, for a total of 10 hours of audio recorded sentences, sampled at 96 kHz and quantized at 24 bits/sample, obtained in a nearly noise-free environment. Signals have been resampled at 16 kHz for computational efficiency reasons. In order to put the proposed system to test in a broader scenario with much fewer training data available, a random selection of 50 sentences, each approximately 5 seconds long, has been taken to compose the experimental database in the following evaluation.

4.1 Objective Evaluation

In order to assess objectively the quality of the sentence obtained from the system, a pair of input sentences is required as in text-dependent training. The converted sentence in the real voice of the target speaker works as a ground-truth sentence. This allows a direct comparison between the sentence obtained from the system through voice conversion and the expected result (ground-truth).

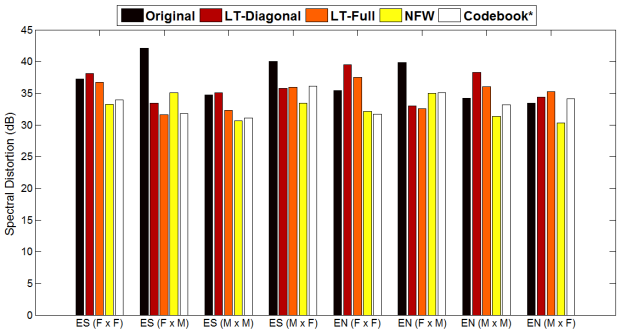

Figure 3. Average Spectral Distortion Rates: each joint set of bars represent a language (EN or ES) and a gender conversion scenario (F=Female, M=Male), and colors represent transformation methods.

The TC-STAR database with its parallel set of sentences in both languages is perfect for this type of evaluation. Both converted and ground-truth sentences are time-aligned and compared through a spectral distance measure applied segment-wise. Alignment is obtained with the classical Dynamic Time Warping (DTW) algorithm [10], adapted to the HSM model.

The local (segment-wise) transformation function is considered to use one of the following methods: (1) Linear transformation using the full covariance matrix (LT-Full); (2) Linear transformation using only the diagonal of the covariance matrix (LT-Diag); (3) Normalized Frequency Warping (NFW) proposed in this work; or (4) Resynthesis of the target signal through weighted reassembly of centroids for the selected target corpus classes (Codebook, described in reference [21]).

An experiment has been conducted using all the corpora for English-speaking and Spanish-speaking speakers in TC-STAR, which have been arranged to cover for all gender combinations in intra-lingual voice conversion, namely $M \times M$, $M \times F$, $F \times M$ and $F \times F$. It should be noted that, although intra-lingual voice conversion is not the main focus of this work, it provides a controlled setting for assessing voice conversion quality across genders, which is one of the well-known issues in general voice conversion.

A set of histograms is shown in Figure 3, which presents spectral distortion values between the converted signal and the ground-truth signal, averaged over segments. The black bars labelled "Original" refer to the spectral distortion between the aligned original source and ground-truth target voice signals. It can be seen from this objective comparison that there is no method that wins on every situation, although either NFW or Codebook have achieved lowest distortions in 75% of the scenarios.

Three of these methods have been chosen for the subjective evaluation to be presented in the sequel: NFW, LT-Diag and Codebook; LT-Full has been left out because of evident noisy artifacts produced, probably due to the re-

Paired Sentences	*MOS* – Naturalness			*MOS* – Similarity		
	LT-DIAG	NFW	COD	LT-DIAG	NFW	COD
F → F	**3.8059**	3.7411	1.8974	2.9645	**3.0821**	3.0411
F → M	**2.3838**	2.2332	1.4696	3.5733	**3.6716**	3.5932
M → M	3.3801	**3.5123**	1.9197	3.3441	**3.6158**	3.4370
M → F	2.9620	**3.0545**	1.6743	3.6674	**3.7323**	3.6029

Table 1. Results for the perceptual evaluation (*MOS*).

duced dataset in the training stage.

4.2 Subjective Evaluation

An online interactive system has been developed in order to apply the required perceptual tests for the subjective evaluation of the cross-lingual voice conversion system proposed. The Mean-Opinion Score (MOS) [3, 4] has been adopted as a standard for measuring naturalness and similarity of the converted voice signals. This experiment had a participation of 50 volunteers.

An interview with each participant was composed by a series of 16 rounds of questions each referring to particular set of paired sentences; presentation order has been randomized. In each round the user was presented with three versions of the converted sentence (through the NFW, LT-Diag and Codebook techniques, in randomized order) and the target sentence, rating converted sentences in a 5-value scale from 5=excellent to 1=bad. All scenarios of cross-lingual and cross-gender voice conversion were covered in this experiment.

In order to gain a general view of the transformation methods, average MOS scores for naturalness and similarity were taken, according to Table 1, for each gender-grouped pair of corpora: $F \rightarrow F$, $F \rightarrow M$, $M \rightarrow M$ and $M \rightarrow F$.

4.3 Discussion

The results corresponding to the objective and subjective evaluations can be seen to agree in the comparison between the NFW and LT-Diag methods, in the sense that NFW has obtained better objective and subjective scores in the cases with a male source speaker (and any target); in the cases with a female source speaker, the cross-gender case $F \rightarrow M$ favoured LT-Diag in the objective evaluation, whereas in the subjective evaluation MOS-Naturalness values favoured LT-Diag but MOS-Similarity values favoured NFW for female source speakers. The Codebook technique appeared to be a serious competitor in the objective evaluation, but scored low in the subjective evaluation with respect to MOS-Naturalness values.

The difficulty of the particular case of female to male voice conversion has been observed in many of the methods here considered, and also in other studies. Zorilă and co-authors [22] believe that this phenomenon is related to the difficulty of obtaining spectral envelopes for female voices, due to the wider spacing between spectral harmonic components. Although several alternative experiments have been conducted in order to try to better understand this phenomenon and tailor methods to this particular situation, no alternatives have been found that consistently achieved higher scores in this voice conversion scenario.

The observed interactions between naturalness and similarity in voice conversion are believed to correspond to a competitive relationship, according to previous studies by several authors [23, 24, 14]. The success in one metric is supposed to be inversely tied to the success in the other, making it unlikely for a method to be highly scored in both metrics at the same time. This conflict is partially explained by the duality between over-fitting in the analysis of sentence segments and over-smoothing of the transformation due to the contribution of many artificial phonetic classes.

5. CONCLUSIONS

This paper presented a complete voice conversion system [1] with an orientation towards cross-lingual voice conversion. The method presupposes a training stage which does not depend on parallel-produced corpora, bilingual speakers or labelled recordings. The system includes a set of tools for analysis, spectral manipulation, clustering and classification of voice segments in terms of artificial phonetic classes that could be also used independently for tasks other than voice conversion.

Experimental results based on the TC-STAR dataset suggest that these methods offer good alternatives for voice conversion across different languages, even though the problem of achieving perfect voice conversion in perceptual terms is still far from being completely understood, let alone solved. Some of the difficulties in objective and subjective evaluations have been presented, which hopefully will shed some light on paths for extension and improvement of the proposed methods in future work.

Among the improvements that are going to be explored, a higher resolution in the quantization of acoustic vectors and artificial phonetic maps should be considered in an attempt to increase the quality of the conversion, as should some more compact alternative to the phase configuration in the representation of the initial phases of harmonic components. Another important challenge is performing conversion in real-time; many conversion modules are parallelizable, such as the parameterization module which can treat different frequency bands simultaneously. This suggests the implementation of the voice conversion system on parallel machines, such as GPGPU [25].

6. REFERENCES

[1] D. Childers, B. Yegnanarayana, and K. Wu, "Voice conversion: Factors responsible for quality," *Interna-*

[1] The code for this system can be obtained at `http://www.ime.usp.br/~dandy/VoiceConversionSystem.zip`.

tional Conference on Acoustics, Speech and Signal Processing, ICASSP-85, pp. 748–751, 1985.

[2] H. Duxans, "Voice conversion applied to text-to-speech systems," *Ph. D. thesis, Universitàt Polytecnica de Catalunia, Barcelona, Spain*, 2006.

[3] D. Sündermann, H. Höge, A. Bonafonte, H. Ney, and J. Hirschberg, "Tc-star: Cross-language voice conversion revisited," *Proceedings of the TC-STAR Workshop on Speech-to-Speech Translation*, pp. 231–236, 2006.

[4] 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.

[5] E. Helander and J. Nurminen, "A novel method for prosody prediction in voice conversion," in *International Conference on Acoustics, Speech and Signal Processing, ICASSP-07*, vol. 4, 2007, pp. 509–512.

[6] A. Kain and M. Macon, "Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction," in *International Conference on Acoustics, Speech and Signal Processing, ICASSP-01*, vol. 2. IEEE, 2001.

[7] Y. Stylianou, "Voice transformation: A survey," *International Conference on Acoustics, Speech and Signal Processing, ICASSP-09*, pp. 3585–3588, 2009.

[8] A. Machado and M. Queiroz, "Voice conversion: A critical survey," *Proceedings of the 7th Sound and Music Computing Conference - SMC*, pp. 291–298, 2010.

[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 Workshop on Speech to Speech Translation*. Citeseer, 2006.

[10] X. Huang, A. Acero, and H. H., *Spoken Language Processing: A guide to Theory, Algorithms, and System Development*. Prentice Hall, 2001.

[11] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj, "Voice conversion using partial least squares regression," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 912–921, 2010.

[12] A. Machado, A. Bonafonte, and M. Queiroz, "Parametric decomposition of the spectral envelope," in *International Conference on Acoustics, Speech and Signal Processing, ICASSP-13*, 2013.

[13] 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.

[14] Y. Stylianou, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, no. 6, pp. 131–142, 1998.

[15] J. Hopcroft and R. Karp, "An $n^{5/2}$ algorithm for maximum matchings in bipartite graphs," *SIAM Journal on ComSputing*, vol. 2, no. 4, pp. 225–231, 1973.

[16] Y. Stylianou, "Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification," *PhD Thesis, Ecole Nationale Supérieure des Télécommunications*, 1996.

[17] R. Balan, P. Casazza, and D. Edidin, "On signal reconstruction without phase," *Applied and Computational Harmonic Analysis*, vol. 20, no. 3, pp. 345–356, 2006.

[18] A. Machado, A. Bonafonte, and M. Queiroz, "Spectral envelope representation using sums of gaussians," in *IberSPEECH 2012 - VII Jornadas en Tecnología del Habla and III Iberian SLTech Workshop*, 2012.

[19] H. Ye and S. Young, "Perceptually weighted linear transformations for voice conversion," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[20] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Communications, Computers and Signal Processing, 1993., IEEE Pacific Rim Conference on*, vol. 1. IEEE, 1993, pp. 125–128.

[21] M. Zhang, J. Tao, J. Tian, and X. Wang, "Text-independent voice conversion based on state mapped codebook," in *International Conference on Acoustics, Speech and Signal Processing, ICASSP-08*, 2008.

[22] T. Zorilă, D. Erro, and I. Hernaez, "Improving the quality of standard gmm-based voice conversion systems by considering physically motivated linear transformations," *Advances in Speech and Language Technologies for Iberian Languages*, pp. 30–39, 2012.

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

[24] D. Sündermann, H. Hoge, A. Bonafonte, H. Ney, A. Black, and S. Narayanan, "Text-independent voice conversion based on unit selection," in *International Conference on Acoustics, Speech and Signal Processing, ICASSP-06*. Citeseer, 2006.

[25] D. Luebke, M. Harris, N. Govindaraju, A. Lefohn, M. Houston, J. Owens, M. Segal, M. Papakipos, and I. Buck, "Gpgpu: general-purpose computation on graphics hardware," in *IEEE Conference on Supercomputing*. ACM, 2006, p. 208.