PhD Dissertation: Intra-lingual and Cross-lingual Voice Conversion using Harmonic plus Stochastic Models

Daniel Erro[1,2]*, supervised by Asunción Moreno*[1]

[1] TALP Research Center, Technical University of Catalonia (UPC), Barcelona
[2] AHOLAB Signal Processing Laboratory, University of the Basque Country (UPV/EHU), Bilbao
derro@aholab.ehu.es, asuncion.moreno@upc.edu

Abstract

This PhD dissertation, written by Daniel Erro and supervised by Asunción Moreno, was defended on June 16th 2008 at UPC. The committee members were Antonio Bonafonte (UPC), Helenca Duxans (TID), Inma Hernáez (EHU), Eduardo Rodríguez-Banga (UVIGO), and Xavier Serra (UPF). Qualification: "sobresaliente cum laude".
Index Terms: voice transformation and conversion, speech synthesis, harmonic plus stochastic model

1. The Voice Conversion Problem

Voice conversion (VC) is the technology used to transform the voice of one speaker (the source speaker) for it to be perceived by listeners as if it had been uttered by a different specific speaker (the target speaker). Among all the speaker-dependent voice characteristics, voice conversion focuses mainly on the acoustic ones: the spectral characteristics and the fundamental frequency. During training, given a certain amount of training data recorded from specific source and target speakers, the system determines the optimal transformation for converting one voice into the other one. Then, the system can apply this transformation to convert new input utterances of the source speaker.

VC has a wide variety of applications, including the design of multi-speaker speech synthesis systems without strong memory requirements, the customization of speaking devices, the design of speaking aids for people with speech impairments, film dubbing using the original actors' voices, the creation of virtual clones of famous people for videogames, masking identities in chat rooms, etc.

2. State-of-the-Art before this Thesis

At the time of beginning this thesis, state-of-the-art VC systems were mainly based on the statistical framework proposed by Stylianou [1] and modified by other authors [2][3][4]. Such systems produced satisfactory results in terms of similarity between converted and target voices, but degraded noticeably the quality of the speech signals. In fact, considering different VC systems and methods, a trade-off could be observed between these two performance dimensions (i.e. converted-target similarity and quality). Therefore, the design of new VC methods characterized by a good similarity-quality balance was one of the main challenges.

The versatility of state-of-the-art VC systems was often limited by their requirements for estimating adequate transformation functions from the training data. A vast majority of them could operate only when a parallel training corpus was available (in other words, when the training sentences uttered by the source and target speaker were exactly the same and therefore showed a clear phonetic correspondence). Although some techniques compatible with non-parallel corpora (different source and target utterances) had been already proposed, being some of them valid for cross-lingual VC, they required some extra information for a correct performance (other pre-trained VC functions [5], phonetic labels [4], etc.) or either the performance scores of the VC system decayed under some conditions [6]. In order to build more versatile VC systems capable of learning transformation functions flexibly under different training conditions, a new source-target alignment procedure was desirable.

Other unsolved problems related to VC were prosody conversion and robustness against data reduction.

3. Objectives and Methodology

The general objective of this thesis was to research into VC systems and methods in order to improve their quality and versatility. The specific objectives were the following:

- Design of new VC methods that succeeded at converting the source voices into the target voices without degrading significantly the quality of the signals.
- Design of training methods that made the VC system capable of operating in all possible training scenarios: intra-lingual scenario with parallel corpus, intra-lingual scenario without parallel corpus, and cross-lingual scenario.
- Integration of the resulting VC system into a text-to-speech (TTS) synthesis system, so that it could operate not only as a conversion device, but also as a stand-alone TTS that generated different voices from a single synthesis database.

The fulfillment of the described objectives was to be verified by means of perceptual tests: both the similarity between converted and target voices and the quality of the converted speech had to be rated by human listeners. The Mean Opinion Scores (MOS) were chosen as figure of merit.

The database used in the experiments, which had been specifically recorded for research on voice conversion in the framework of the TC-STAR project, contained around 200 sentences in Spanish and 170 in English, uttered by 4 different professional bilingual speakers (2 male + 2 female speakers). The average duration of the sentences was 4 seconds (10-15 minutes of audio per speaker and language).

The research work carried out during the thesis can be summarized in the following steps:

- Design and implementation of a speech model suitable for speech analysis-reconstruction, pitch and duration modification, synthesis, and spectral manipulation.
- Implementation and subjective evaluation of a baseline VC system using the mentioned speech model, state-of-the-art VC techniques, and parallel training corpora.
- Research on strategies for improving the quality scores of the baseline system without worsening the conversion scores. Subjective evaluation of the resulting method.
- Research on strategies to allow the system to train VC functions from non-parallel corpora while maintaining its subjective performance scores. Evaluation under both, intra-lingual and cross-lingual conditions.
- Implementation and evaluation of a multi-speaker TTS system and optimization of the interaction between the synthesis process and the VC process.

The next section presents an overview of the main contributions of this thesis.

4. Main Contributions of the Thesis

4.1. Flexible Harmonic/Stochastic Model

A new speech model based on a harmonic plus stochastic decomposition was developed during the thesis. This model allowed the manipulation of all kind of signal features with a high degree of flexibility, which was desirable for implementing VC systems. The novelty of the model lay in the algorithms for time-scale manipulations, pitch-scale manipulations, and concatenation of units, which were compatible with a non-pitch-synchronous analysis scheme. The reasons for preferring a constant analysis frame rate rather than a pitch-synchronous rate were augmenting the flexibility and simplifying the analysis (because the accurate separation of the signal periods was not a previous requirement). In exchange, in order to make artifact-free speech modification possible, the problem of estimating and manipulating the linear-in-frequency phase term of the speech frames without producing artifacts was faced. In contrast to previous non-pitch-synchronous models based on sinusoidal or hybrid decompositions, it was not necessary to use onset times or pitch-synchronous epochs as a reference. The use of computationally expensive techniques such as inverse filtering was also avoided. Instead, amplitude and phase envelopes were used as estimators of the vocal tract, assuming a simplified speech production model. A new method for removing the linear phase term from a set of measured harmonics was also proposed.

In order to validate the suitability of the new model for high-quality speech transformations, it was integrated into the waveform generation module of a concatenative TTS system which was then compared to an equivalent TD-PSOLA-based system. The results showed that the listeners had a clear preference for the new system when the synthesis required applying high modification factors. It was concluded that the speech model was valid for high-quality speech transformation and concatenation and provided a very good framework for research on voice conversion.

4.2. Weighted Frequency Warping

As mentioned before, state-of-the-art VC systems were based on statistical methods. Most of them followed Kain's approach [2], in which a set of paired vectors was extracted from parallel utterances of the source and target speakers and a joint Gaussian mixture model (GMM) was fitted to them. The mean vectors and covariance matrices provided by the GMM were used to define the conversion function, a weighted combination of linear transforms:

$$F(\mathbf{x}) = \sum_{i=1}^{m} p_i(\mathbf{x})\left(\boldsymbol{\mu}_i^y + \boldsymbol{\Sigma}_i^{yx}\boldsymbol{\Sigma}_i^{xx\,-1}\left(\mathbf{x} - \boldsymbol{\mu}_i^x\right)\right) \quad (1)$$

where $\mathbf{x}$ is the source acoustic vector to be converted; $p_i(\mathbf{x})$ is the probability that $\mathbf{x}$ belongs to the i-th mixture of the GMM, given by the mean $\boldsymbol{\mu}_i^x$ and the covariance matrix $\boldsymbol{\Sigma}_i^{xx}$; $\boldsymbol{\mu}_i^y$ is the target mean vector and $\boldsymbol{\Sigma}_i^{yx}$ is the cross-covariance matrix. This method was known to yield a successful conversion of voices but significant speech quality degradation due to statistical over-smoothing.

When the above mentioned method was implemented as baseline system, a high correlation was observed between the spectral envelopes given by the source and target mean vectors, $\boldsymbol{\mu}_i^x$ and $\boldsymbol{\mu}_i^y$, for all the mixtures of the GMM. A simple frequency-warping transformation (a mapping between source and target frequency axes) of the source mean envelopes appeared to yield good estimates of the target mean envelopes. This was an interesting finding, since frequency warping of spectra was known to introduce very little quality degradation. Since vectors inside each Gaussian component of the GMM contained parametric representations of phonemes with similar formant structures, it was assumed that a single frequency warping function could be valid for all the spectra belonging to that class.

The resulting VC method, which was called Weighted Frequency Warping (WFW), can be described as follows. During training, once a joint GMM has been estimated from a set of paired vectors, the spectra given by $\boldsymbol{\mu}_i^x$ and $\boldsymbol{\mu}_i^y$ are used to estimate a warping function $W_i(f)$ for each mixture of the GMM (see Figure 1). During conversion, given the spectrum of the k-th frame, $S^{(k)}(f)$, and its parametric representation $\mathbf{x}^{(k)}$, the following transformation is applied to obtain the converted spectrum $S'^{(k)}(f)$:

$$S'^{(k)}(f) = G^{(k)}(f) \cdot S_{fw}^{(k)}(f) \quad (2)$$

where $S_{fw}^{(k)}(f)$ is a frequency-warped version of the original spectrum, calculated by applying a weighted combination of the trained warping trajectories $\{W_i(f)\}$,

$$S_{fw}^{(k)}(f) = S^{(k)}\left(W^{(k)\,-1}(f)\right) \quad (3)$$

$$W^{(k)}(f) = \sum_{i=1}^{m} p_i\left(\mathbf{x}^{(k)}\right) \cdot W_i(f) \quad (4)$$

and $G^{(k)}(f)$ is a smoothed correction filter that compensates the differences in amplitude between the warped spectrum $S_{fw}^{(k)}(f)$ and the real target spectrum, estimated from vector $F(\mathbf{x}^{(k)})$ given by expression (1).

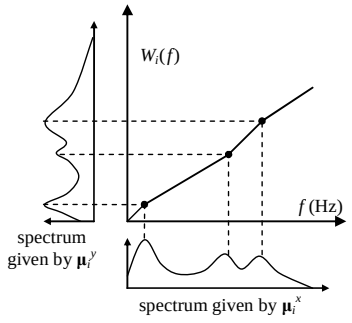

Figure 1: *Estimating a frequency warping function from the mean spectra of the i-th mixture of a GMM.*

The subjective evaluation of a WFW-based VC system confirmed that the method provided much better quality than standard GMM-based systems. The average improvement was around 0.7 points in a 1-to-5 MOS scale, whereas the similarity scores remained almost unaltered. The average quality level achieved by WFW in these experiments was close to 3.5 in a 1-to-5 MOS scale, which could be considered acceptable for real voice conversion applications. The system achieved also excellent results in the 3rd international TC-STAR evaluation campaign.

4.3. Iterative Alignment Method

A new solution was proposed to allow the system to train VC functions when no parallel corpus was available. In other words, it made possible the correct alignment between source and target acoustic vectors extracted from text- and language-independent utterances. The idea came from the observation that when a simple nearest neighbour (NN) criterion was used to pair the source and target vectors during training, an intermediate converted voice was obtained (closer to the target, but not very different from the source). If a NN alignment was carried out again on the intermediate converted vectors and the target vectors, the resulting voice got slightly closer to the target voice. Therefore, the proposed method consisted of iteratively repeating these two steps: 1) NN alignment; 2) training of a VC function, and conversion of the source vectors. This idea is described graphically in Figure 2.

Some experiments using parallel test corpora and several objective measures were conducted in order to check that the alignment got more accurate (compared to the parallel corpus case) as the number of iterations increased. At the same time, several aspects of the method were studied using the same objective measures: configuration of the NN search, convergence, stopping criterion, initialization, etc. The objective performance of the alignment system resulted to be close to that of parallel alignment in most cases. Although the speech frames were being aligned using only acoustic information extracted directly from the signal, the experiments showed that the alignment was acceptable also from a phonetic point of view.

In order to evaluate the new alignment method through perceptual tests, it was integrated into the training module of a VC system. In an intra-lingual context, the results were found to be highly satisfactory, similar to those obtained by an equivalent voice conversion system using parallel training corpora. The performance under cross-lingual conditions was slightly worse, due to the phonetic differences between languages. A cross-lingual VC system that resulted from the combination of the described alignment method and WFW (see section 4.2) participated in a public evaluation campaign organized under the authority of the European TC-STAR project. It achieved the best results considering both the similarity of voices and the quality of the converted speech.

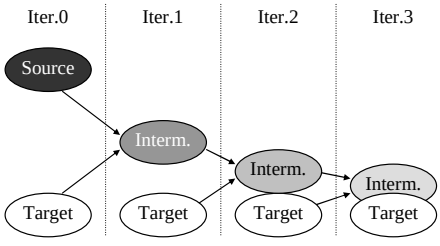

Figure 2: *Underlying idea of the iterative alignment method.*

4.4. Multi-speaker TTS System

A multi-speaker synthesis system was built by combining the UPC TTS system with a VC system based on the methods and algorithms presented throughout this thesis. The full system was evaluated by means of perceptual tests. As expected, the scores indicating the similarity between converted and target voices were very close to those obtained when converting natural speech utterances (approximately 3.1 in a 1-to-5 MOS scale). The quality of the converted synthetic signals was affected by both synthesis and conversion, so the resulting average score was 2.5 (approximately 1 point below the one obtained for converted natural speech and 0.5 points below the one obtained for non-converted synthetic speech). The results revealed also that the choice of the source voice had a direct influence on the performance of the system.

5. General Conclusions

As a result of this thesis, contributions were made in many parts of the voice conversion process. The analysis, manipulation and reconstruction of speech signals were improved through the design of a new speech model based on a harmonic/stochastic decomposition. A new spectral conversion method, WFW, was proposed to increase the quality of the converted speech with respect to state-of-the-art methods and alleviate the over-smoothing effect. An iterative alignment method was also proposed to overcome the lack of versatility of VC systems when no parallel training corpus was available. All these ideas were put into practice in a multi-speaker TTS system. All these methods and systems yielded highly satisfactory performance scores as shown by different subjective listening tests.

6. Publications and Merits

The main contributions of the thesis [7], summarized in section 4, were published in two journal papers in IEEE Trans. Audio, Speech, and Lang. Proc. [8][9], and also in several papers in the most important conferences related to speech technologies, being the most relevant ones the following: [10][11][12][13]. In addition, the author was invited by the Music Technology Group of the Pompeu Fabra University to give a talk on the results of this thesis [14]. Some indirect results of the thesis contributed to the publication of two more journal papers: in [15] the methods described in section 4.1 and 4.2 were used in an emotion conversion application; in [16] the VC system was used to evaluate the robustness of a speaker recognizer. In one of the two papers awarded during last JTH [17] (best paper award), the parameter extraction module of the VC system was applied to build an HMM-based synthesizer, outperforming other traditional methods. Moreover, the VC system developed during this thesis took part in different public evaluation campaigns. It is worth mentioning the diploma obtained in Albayzin 2006 evaluation (which consisted in cheating a biometric system based on speaker recognition), and the excellent results achieved in the 3rd international evaluation campaign of the EU funded project TC-STAR [18] (best results in cross-lingual voice conversion categories). Finally, the findings in this thesis contributed to make progress in the framework of three different funded projects: TC-STAR (FP6-506738, European Comission), AVIVAVOZ (TEC2006-13694-C03, Spanish Ministry of Science and Education) and Tecnoparla (Catalan Government).

7. References

[1] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion", IEEE Trans. Speech and Audio Proc., vol. 6, pp. 131–142, 1998.

[2] A. Kain, "High resolution voice transformation," Ph.D. dissertation, OGI School of Sci. and Eng., Beaverton, OR, 2001.

[3] T. Toda, A.W. Black, K. Tokuda, "Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter", Proc. IEEE ICASSP, vol.1 pp.9-12, 2005.

[4] H. Ye and S. Young, "Quality-enhanced voice morphing using maximum likelihood transformations," IEEE Trans. Audio, Speech and Lang. Proc., vol. 14, no. 4, pp. 1301–1312, 2006.

[5] A. Mouchtaris, J. Van der Spiegel, P. Mueller, "Nonparallel training for voice conversion based on a parameter adaptation approach", IEEE Trans. Audio, Speech and Lang. Proc., vol. 14, no. 3, pp. 952–963, 2006.

[6] D. Sündermann, H. Höge, A. Bonafonte, H. Ney, A. W. Black, S. Narayanan, "Text-independent voice conversion based on unit selection", Proc. IEEE ICASSP, vol. 1, pp. 81–84, 2006.

[7] D. Erro, "Intra-lingual and Cross-lingual Voice Conversion using Harmonic plus Stochastic Models", PhD thesis, Technical University of Catalonia (UPC), 2008.

[8] D. Erro, A. Moreno, A. Bonafonte, "Voice Conversion based on Weighted Frequency Warping", IEEE Trans. Audio, Speech, and Lang. Proc., vol. 18(5), pp. 922-931, 2010.

[9] D. Erro, A. Moreno, A. Bonafonte, "INCA Algorithm for Training Voice Conversion Systems from Nonparallel Corpora", IEEE Trans. Audio, Speech, and Lang. Proc., vol. 18(5), pp. 944-953, 2010.

[10] D. Erro, T. Polyakova, A. Moreno, "On combining statistical methods and frequency warping for high-quality voice conversion", Proc. IEEE ICASSP, pp. 4665-4668, 2008.

[11] D. Erro, A. Moreno, "Weighted Frequency Warping for Voice Conversion", Proc. Interspeech, pp. 1965-1968, 2007.

[12] D. Erro, A. Moreno, "Frame Alignment Method for Cross-Lingual Voice Conversion", Proc. Interspeech, pp. 1969-1972, 2007.

[13] D. Erro, A. Moreno, A. Bonafonte, "Flexible Harmonic/ Stochastic Speech Synthesis", Proc. 6th ISCA Speech Synthesis Workshop, 2007.

[14] D. Erro, "Voice conversion: state of the art, recent improvements and remaining challenges", seminar at Pompeu Fabra University, Barcelona, 2009.

[15] D. Erro, E. Navas, I. Hernáez, I. Saratxaga, "Emotion Conversion based on Prosodic Unit Selection", IEEE Trans. Audio, Speech, and Lang. Proc., vol. 18(5), pp. 974-983, 2010.

[16] M. Farrús, M. Wagner, D. Erro, J. Hernando, "Automatic Speaker Recognition as a Measurement of Voice Imitation and Conversion", International Journal of Speech, Language and the Law, vol. 17(1), pp. 119-142, 2010.

[17] E. Banos, D. Erro, A. Bonafonte, A. Moreno, "Flexible Harmonic/Stochastic Modelling for HMM-based Speech Synthesis", Proc. V Jornadas en Tecnologías del Habla, pp. 145-148, 2008.

[18] D. Mostefa, O. Hamon, N. Moreau, K. Choukri, Evaluation Report, Deliverable D30 of the EU funded project TC-STAR, http://www.tc-star.org, 2007.