

Voice Conversion of Non-aligned Data using Unit Selection

Helenca Duxans, Daniel Erro, Javier Pérez,
Ferran Diego, Antonio Bonafonte, Asunción Moreno

Signal Theory and Communication Dept., TALP Research Center
Universitat Politècnica de Catalunya (UPC), Barcelona, Spain

Abstract

Voice conversion (VC) technology allows to transform the voice of the source speaker so that it is perceived as the voice of a target speaker. One of the applications of VC is speech-to-speech translation where the voice has to inform, not only about what is said, but also about who is the speaker. This paper introduces the different methods submitted by UPC to the TC-STAR second evaluation campaign. One method is based on the LPC model and the other on the Harmonic+Noise Model (HNM). Unit selection techniques are employed so that the methods no longer require parallel sentences during the training phase. We have applied these methods both to *intra-lingual* and *cross-lingual* voice conversion. Results from the TC-STAR evaluation show that the speaker identity is successfully transformed with all the methods. Further work is required to increase the quality of the voice so that it achieve the quality of current TTS voices.

1 Introduction

Voice Conversion (VC) systems modify a speaker voice (source speaker) to be perceived as if another speaker (target speaker) had uttered it. Therefore, given two speakers, the goal of a VC system is to determine a transformation that makes the speech of the source speaker sounds as if it were uttered by the target speaker. Applications of VC systems can be found in several fields, such as TTS (text-to-speech systems) customization. Nowadays, high quality TTS are based on acoustic unit concatenation, i.e. to produce an utterance the most appropriated acoustic units are selected from speaker-dependent databases. In order to produce a high quality synthetic voice, a large amount of recorded and processed data is needed, making the development of a new speaker voice an expensive and time consuming task. VC techniques can be used as a fast and a cheap way of building new voices for TTS systems. It will make possible, for instance, to read e-mails or SMS with their sender's voice, to assign our and our friends voices to characters when playing on a computer game, or to apply different voices to different computer applications. VC can also be very useful in speech-to-speech translation, in applications that require that listeners identify the speaker. For example, when the speech to be translated has been generated by several speakers as in meetings, movies or debates. In such situations, it is important to be able to differentiate between speakers by their voices.

Many VC systems require that the source and target speakers utter the same sentences. Based on these aligned sentences, a transformation function is estimated. However, this is not possible in the speech-to-speech translation framework. First of all, the source speaker does not speak the target language, so it is not possible to have aligned sentences in the target language. Furthermore, the system has to be non-intrusive, i.e., it is not possible to get specific training sentences from the source speakers.

TC-STAR organizes periodic evaluations of the different technologies, open to external partners. Recently, the second evaluation campaign took place including the assessment of intra and cross-lingual voice conversion activities

in English, Mandarin and Spanish. This paper reports the three approaches followed by UPC. The methods have been applied both to intra and cross-lingual voice conversion and do not require aligned sentences (text-independent). Section 2 deals with the *text-independent* issue. Basically, the idea is to use the back-end of the TTS to generate sentences with similar prosody to the target. This approach can also be used as a voice conversion method (at least as a reference for voice conversion methods). The second method is presented in section 3: section 3.1 is devoted to vocal tract conversion and section 3.2 presents the residual transformation techniques. The third method is presented in section 4: section 4.1 gives an overview of the method, detailed in section 4.2. Section 5 presents and discusses the results which have been obtained in the TC-STAR evaluation. The final section summarizes the main conclusions of this paper.

2 Alignment using unit selection

Most of the works in voice conversion required aligned data, i.e., the transformation is estimated from pairs of sentences uttered by the source and target speaker. But this requirement can limit the use of voice conversion. Even in some cases this is not possible at all, as in the speech translation framework. Here we present our work in the unaligned training context. The approach we follow here is to synthesize source sentences which are parallel to the target sentences. Our goal is to transform the TTS voice so that it sounds as the target. In order to produce parallel data the front-end is based on the target samples and the back-end uses the unit selection module of the speech synthesizer. After this step, the different algorithms employed for the training process using parallel data can be applied, as will be explained in sections 3 and 4.

The method we propose performs a resynthesis of the input utterance, corresponding to the source speaker. The prosody (fundamental frequency contour, duration and energy) is copied from the source utterance, and the selection module is forced to use a database corresponding to the target speaker. In the evaluation task the source voice was not the TTS voice, but a speaker with limited data. Therefore,

we build a TTS based on this data. Some of the constraints of the unit selection algorithm needed to be relaxed, since by default it works with either diphones or triphones, and in our case the reduced size of the database implied that some units were missing. We also forced the selected speech segments to belong to a different utterance than that of the input. This is necessary since during the training stage all the database was available, and when analyzing one file, the unit selection module would find that the best candidate units were those belonging to this file. The training data available in this campaign consist on parallel sentences but we wanted to test a text-independent method.

At the output of this module we have the selected units of the target speaker, and the automatic phonetic segmentation of the source utterance. Hence, we have obtained the alignment of the source and target phonetic units and are able to use the same voice conversion algorithm as in the aligned case. Next sections present the particularities of the different VC algorithms using this alignment information.

3 VC using LPC and phonetic information

All the methods that deal with vocal tract conversion are based on the idea that each speaker has his/her own way of uttering a specific phone. Therefore, the spectral mapping function has to take into account some phonetic/acoustic information in order to choose the most appropriate relationship for converting the vocal tract (LSF, Line Spectral Frequencies, coefficients) of each speech frame. To complete the conversion from the source speaker to the target speaker, a target LPC residual signal prediction from the converted LSF envelopes is carried out. This strategy assumes that the residual is not completely uncorrelated with the spectral envelope, making the prediction possible (Kain, 2001).

Next section deals with the vocal tract conversion, and section 3.2 gives the details of the residual transformation.

3.1 Decision tree based vocal tract conversion

Generally, a vocal tract conversion system may be divided in three components: a model of the acoustic space with a structure by classes, an acoustic classification machine and a mapping function (see figure 1).

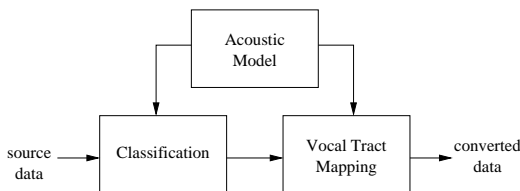


Figure 1: Vocal tract conversion block diagram.

Previous Gaussian Mixture Models (GMMs) based vocal tract conversion systems (Stylianou et al., 1998; Kain, 2001) use only spectral features to estimate acoustic models by maximum likelihood. CART (classification and regression tree) allow working with numerical data (such as spectral features) as well as categorical data (such as phonetic features) when building an acoustic model. Phonetic data is available for TTS voices and may be very useful in

the classification task, because the acoustics are somehow related to the phonetics.

The procedure to grow a CART for vocal tract conversion is as follows. First, the available training data is divided into two sets: the training set and the validation set. A joint GMM based conversion system (Kain, 2001) is estimated from the training set for the parent node t (the root node in the first iteration), and an error index $E(t)$ for all the elements of the training set belonging to that node is calculated. The error index used is the mean of the Inverse Harmonic Mean Distance between target and converted frames, calculated as:

$$E(t) = \frac{1}{|t|} \sum_{n=0}^{|t|-1} D(\tilde{\mathbf{y}}_n, \mathbf{y}_n), \quad (1)$$

where $|t|$ is the number of frames in the node t , \mathbf{y} is a target frame and $\tilde{\mathbf{y}}$ its corresponding converted frame. The distance $D(\tilde{\mathbf{y}}, \mathbf{y})$:

$$D(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{p=1}^P c(p)(x(p) - y(p))^2} \quad (2)$$

$$c(p) = \frac{1}{w(p) - w(p-1)} + \frac{1}{w(p+1) - w(p)} \quad (3)$$

with $w(0) = 0$, $w(P+1) = \pi$ and $w(p) = x(p)$ or $w(p) = y(p)$ so that $c(p)$ is maximized (p is the vector dimension), weights more the mismatch in spectral peaks than the mismatch in spectral valleys when working with LSF vectors.

All the possible questions of the set Q are evaluated at node t and two child nodes t_L and t_R are populated for each question q . The left descendant node is formed by all the frames which fulfill the question and the right node by the rest. The set Q is formed by binary questions of the form $is \{\mathbf{x} \in A\}$, where A represents a phonetic characteristic of the frame \mathbf{x} , in particular: the vowel/glide/consonant category, the point and manner of articulation for consonants, the height and the backness for vowels and glides, and the voicing.

For each child node, a joint GMM conversion system is estimated, and the error figures $E(t_L, q)$ and $E(t_R, q)$ for the training vectors corresponding to the child nodes t_L and t_R obtained from the question q are calculated. The increment of the accuracy for the question q at the node t can be calculated as:

$$\Delta(t, q) = E(t) - \frac{(E(t_L, q)|t_L|) + (E(t_R, q)|t_R|)}{(|t_L| + |t_R|)}. \quad (4)$$

To decide if a node will be split or not, the increment of accuracy for the training set is evaluated for each question and the question q^* corresponding to the maximum increment is selected. Then, the increment of accuracy for the validation set for the question q^* is calculated, and only if it is greater than zero the node will be split. The tree is grown until there is no node candidate to be split. The decision tree constructed by this procedure can be used to divide the acoustic space into overlapping classes determined by

phonetic properties. Each leaf represents a hidden acoustical class and has defined a conversion function. To estimate a conversion function for each leaf, all the available data (training set plus validation set) is classified by the tree. Then, the data of each class is used to estimate a joint GMM with one component and the transformation function related is derived as:

$$\hat{y}_i = \mu_i^y + \sum_i^{y^x} \sum_i^{x^x} \Sigma_i^{x^x} (x - \mu_i^x). \quad (5)$$

It must be remarked that, although the transformation function of each leaf is estimated with data of a single phonetic class, the transformation is continuous and defined in all the acoustic space. Both properties are a requirement to assure a high quality of the converted speech. To transform new source vectors, they are classified into leafs according to their phonetic features by the decision tree. Then, each vector is converted according to the GMM based system belonging to its leaf.

3.2 Residual Selection and Fixed Smoothing

In the residual selection technique, which has been proved to be a better approach than the residual conversion technique (Duxans and Bonafonte, 2006), residuals are selected from a database extracted from the target training data. In the current work, each entry of the database is formed by a target LSF vector \mathbf{y} and its corresponding residual signal \mathbf{r} . Only voiced residuals with a length l in the interval $\mu_T - 1.5 * \sigma_T \leq l \leq \mu_T + 1.5 * \sigma_T$, where T is the pitch period length, have been used to build the database.

To produce the converted speech, once the vocal tract has been transformed, a residual signal is selected from the database. The criteria used to select the residual r_k for the converted envelope \tilde{y} is to choose that residual which associated LSF vector \mathbf{y}_k minimizes the Inverse Harmonic Mean Distance between \tilde{y} and \mathbf{y}_k . For unvoiced frames, white noise samples are used as residuals.

The output signal is generated by filtering the selected residual signals with the inverse LSF filter. The prosody is manipulated using TD-PSOLA. Since no similarity criteria over neighbor residual signals is imposed, concatenation problems appear. Thus we need to smooth the voiced residual signals once they are selected from the database. The smoothing applied in this work is a weighted average over neighbor frames, with weights equal to a normal distribution centered in the current frame. Unlike previous works (Sündermann et al., 2005), the average is applied only to voiced residuals, and the normal weighting window has a fixed duration.

4 VC using a harmonic/stochastic method

In voice conversion systems not only the spectral characteristics of the voice are considered, but also some prosodic aspects, so it is important to use a synthesis system capable of modifying all these features in a flexible way. Furthermore, the output signal of the TTS block may have some acoustic discontinuities or artifacts caused by the concatenation of units containing slight spectral differences. A good synthesis system should minimize this kind of effects before passing the signals to the voice conversion module.

With regard to the prosody, in most of the voice conversion systems found in the literature, a pitch-synchronous synthesis system is used to generate the converted waveform and modify the prosodic parameters (Stylianou et al., 1998; Kain, 2001; Duxans et al., 2004). The main advantage of this kind of systems is that the frames correspond to the signal periods, so the prosodic modifications can be performed by means of any PSOLA technique, and each frame can be processed individually without losing the phase coherence in the regenerated signal. The main disadvantage is the need of a robust method for the accurate separation of all the signal periods. The use of constant-length frames can induce significant artifacts if the phase envelopes are altered in any way. However, if the problem of the phase manipulation is solved successfully, several advantages can be obtained:

- The errors coming from the separation of periods are avoided. In addition, it is not necessary to use pseudo-periods in the unvoiced regions. The pitch and the voiced/unvoiced decision are not necessary a priori.
- The use of constant length frames is desirable for the analysis of signals in real-time applications. It is easier and more reliable to measure the pitch than to locate the exact position of the pitch marks.
- The analysis rate can be controlled manually, so more parameters can be extracted from the same amount of audio data.

With regard to the flexibility and capability of spectral manipulation, methods like TD-PSOLA, frequently found in the TTS systems, may be not appropriated for the voice conversion task, because they assume no model for the speech signal. In addition, if the unit database is small, the noise caused by the spectral discontinuities at the concatenation points can seriously affect the quality of the synthetic signal. The quality provided by other systems based on LPC or residual-excited LPC is not as high as desirable, but in exchange the LPC parameters are easy to convert. The different variants of sinusoidal or harmonic models provide good knowledge of the signal from the perceptual point of view, and allow manipulating many characteristics of the signal by changing its parameters in a very flexible way. Furthermore, they minimize the concatenation artifacts, and can operate in a pitch-asynchronous way. For all these reasons, the synthesis system presented in (Erro and Moreno, 2005), based on the decomposition of a speech signal into a harmonic and a stochastic component, has been applied to develop a new voice conversion system.

4.1 Synthesis system overview

The deterministic plus stochastic model assumes that the speech signal can be represented as a sum of a number of sinusoids and a noise-like component (Erro and Moreno, 2005). In the analysis step the signal parameters are measured at the so called analysis points, located in samples $n=kN$, $k=1, 2, 3, \dots$. N is a constant number of samples corresponding to a time interval of 8 or 10ms. At each analysis point, the following parameters are extracted:

- Fundamental frequency. If the analysis point is inside an unvoiced region, the fundamental frequency is considered to be zero.
- Amplitudes and phases of all the harmonics below 5KHz, only in voiced regions. Note that the voicing decision employed is binary.
- The LPC coefficients that characterize the power spectral density of the stochastic component.

In order to resynthesize the signal from its measured parameters, both the deterministic and the stochastic components are rebuilt using the overlap-add technique. A frame of $2N$ samples centered at each analysis point k is built by summing together all the detected sinusoids with constant amplitudes, frequencies and phases. For the generation of the stochastic component, $2N$ -length frames of white Gaussian noise are shaped in frequency by the previously calculated LPC filters. A triangular $2N$ -length window is then used to overlap and add the frames in order to obtain the time-varying synthetic signal.

The duration modification of the signal can be carried out by increasing or decreasing the distance N between the different analysis points, so that the amplitude and fundamental frequency variations get adapted to the new time scale. The change in N needs to be compensated with a phase manipulation in a way that the waveform and pitch of the duration-modified signal are similar to the original.

When the pitch of the signal is modified, the amplitudes of the new harmonics are obtained by a simple linear interpolation between the measured amplitudes in dB. The new phases can be obtained by means of a linear interpolation of the real and imaginary parts of the measured complex amplitudes, but the interpolation has to be done in the same conditions for all the analysis points, in order to guarantee the coherence. Finally, a new phase term has to be added to compensate the modification of the periodicity, because the relative position of the analysis point within the pitch period has changed.

Different analyzed units can be concatenated together in order to synthesize new utterances. The deterministic and stochastic coefficients inside each unit are transformed to match the energy, duration and pitch specifications. A phase shift is added to the harmonics of the second unit to make the waveforms match properly. Another adjustment is carried out in the amplitudes of the sinusoids near the borders between units, to smooth the spectrum in the transition.

4.2 The voice conversion method

The speaker modification is performed in several steps:

- Prosodic scaling, in which only the F_0 and the frequencies are changed according to a simple transformation.
- Vocal tract conversion, which is linked to the amplitudes of the sinusoids.
- Phase calculation, because the phase variations are tied to the amplitude variations, and if this equilibrium is broken, a significant loss of quality is induced.

- Stochastic component prediction.

4.2.1 Fundamental frequency scaling

The fundamental frequency is characterized by a log-normal distribution. During the training phase, an estimate of the average value μ and variance σ of $\log F_0$ is calculated for each speaker. The only prosodic modification consists of replacing the source speaker's μ and σ by the values of the target speaker. The frequencies of the sinusoids are then scaled according to the new pitch values.

4.2.2 Transformation of the amplitudes

Three different types of parameters were considered to model the vocal tract: line spectral frequencies (LSF), discrete cepstral coefficients, and some points of the amplitude envelope, obtained from the amplitudes of the sinusoids measured in dB by linear interpolation. The LSF coefficients were considered the most suitable, for several reasons:

- They are a good representation for the formants structure, and have been shown to possess very good interpolation characteristics. Furthermore, a bad estimation of one of the coefficients affects only one small portion of the spectrum.
- If the amplitudes of the sinusoids are substituted by the sampled amplitude response of the all-pole filter associated with the LSF coefficients, keeping the phases and the stochastic parameters without variation, there is not a perceptually important quality loss. This fact means that the codification is not an important source of errors.
- The all-pole filter associated with the LSF coefficients provides not only a magnitude envelope but also a phase envelope, whose information can be used as an estimate of the phase envelope of the speaker. The other types of parametrization are magnitude-only.
- As the stochastic component is parametrized by means of LPC, the same type of codification can be easily used for both components of the speech.

For each vector of amplitudes, the optimal all-pole filter is obtained by the Discrete All-Pole Modeling technique, in which the Itakura-Saito distortion measure between the measured amplitudes and the envelope of the filter is minimized (El-Jaroudi and Makhoul, 1991). The resolution given by a 14th order filter is accurate enough for a sampling frequency of 16 KHz. The aligned LSF vectors of both the source and the target speaker are used to estimate the parameters of an 8th order GMM, and the converted amplitudes are obtained by sampling the envelope of the all-pole filter associated with the converted LSF vector. An attempt was made to convert also the residual amplitudes of the codification, but no significant improvements were reached, and in some cases the quality of the converted speech got worse.

4.2.3 Phase envelope adjustment

It must be taken into account that in order to avoid unpleasant artifacts, the variations in the magnitude envelope must

entail appropriate variations in the phase envelope, but at the same time the phase coherence must be maintained during the consecutive frames. To satisfy these two objectives, the phase of the sinusoids is calculated in two steps. In the first step, the phase of the j th sinusoid at frame k is obtained from the value at frame $k-1$:

$$\varphi_j^{(k)} = \varphi_j^{(k-1)} + j\pi NT_s \left(f_0^{(k-1)} + f_0^{(k)} \right). \quad (6)$$

This equation assumes that the frequency of the j th harmonic varies linearly from $k-1$ to k . At the beginning of the voiced regions, all the phases are initialized to zero. There are no phase discontinuities at the end of the first step, but also no variations in the phase envelope from one frame to the next, so an annoying metallic noise appears in the resynthesized signal. During the second step, the phase envelope of the converted filter H is added as a new contribution to the final phases.

$$\varphi_j^{(k)} = \varphi_j^{(k)} + \arg \left\{ H \left(f_j^{(k)} \right) \right\}. \quad (7)$$

The phase of H does not represent the real phase envelope of the converted speech, but it provides small phase variations from one frame to the next, tied to the amplitude variations, and the metallic noise disappears.

4.2.4 Stochastic component prediction

It can be proved that the conversion of the stochastic component is not as important as the previous one. When the signals are analyzed using sinusoids and noise, it is very difficult to completely extract the non-sinusoidal component from the voiced regions of the original sound. In fact, the sinusoids beyond the voicing frequency treated as part of the stochastic component, do not form a part of it. Therefore, there is a strong dependence between some portions of the stochastic spectrum and the vocal tract. Other problems can be caused by inaccurate pitch detection, imprecise measurement and interpolation of the amplitude and instantaneous phase of the detected sinusoids between two consecutive frames, etc. In this paper, we have worked under the assumption that the stochastic component obtained at the voiced regions is in general highly correlated with the vocal tract. Then, a new GMM can be estimated from the LSFs corresponding to the amplitude envelopes and to the stochastic component. This modeling has a smoothing effect over the different stochastic instances that are measured for each phoneme at the analysis step, so the breathing noise and other irregularities are minimized. For the unvoiced regions, no transformation is performed.

5 TC-STAR Evaluation

TC-STAR organizes periodical evaluations in all the speech-to-speech translation technologies, including speech synthesis and voice conversion. In the second campaign (March 2006), voice conversion has been evaluated in English, Mandarin and Spanish. For Spanish-English, one specific track was cross-lingual voice conversion.

5.1 Language resources

UPC produced the language resources for supporting the evaluation of English/Spanish. Basically, 4 bilingual speakers English/Spanish recorded around 200 sentences in each

language. To ease the alignment (for those methods that require it), a mimic style was used, as proposed by (Kain, 2001). Ten sentences were reserved for testing and the others for training. The sentences were designed to be phonetically rich. The recordings are of high quality (96kHz, 24 bits, three channels, including laryngograph). Details about the LR can be found in Bonafonte et al. (2006).

5.2 Evaluation metric

The evaluation was based on subjective rating by human judges. 20 judges were recruited to complete the evaluations. The judges were between 18 and 40 years old native speakers with no known hearing problem. They were not experts in speech synthesis; they were paid for the task. Perceptual tests were carried out via the web. Judges were required to have access to highspeed/ ADSL Internet connection and good listening material.

Two metrics were used in the evaluations: one for rating the success of the transformation in achieving the desired speaker identification, and one for rating the quality. This is needed since strong changes usually achieve the desired identity at the penalty of degrading the quality of the signal. To evaluate the performance of the identity change, the human judges were presented with examples from the transformed speech and the target one. They have to decide using a 5-points scale if the voices comes from different speakers (1) or from the same speaker (5). Some natural source-target examples were also presented as a reference. The judges rate the transformed voice quality using a 5-points MOS scale, from bad (1) to excellent (5).

5.3 Evaluation results

We have submitted three systems to the TC-STAR evaluation. The first method (UPC1) consists on a TTS-back-end that uses the phonetic and prosodic representation of the source speech. The synthetic speech is produced using a concatenative synthesizer built using the training data (approx. 200 sentences). The second method (UPC2) is the method explained in section 3. The third method (UPC3) uses the approach explained in section 4, using the UPC1 method in the training phase to avoid the use of parallel sentences. UPC1 was presented only to the intra-lingual evaluation, while UPC2 and UPC3 were submitted to both intra and cross-lingual evaluation. To apply the methods to the cross-lingual condition we rely on bilingual speakers: the transformation was learnt in one language and applied to the other language. This requires that the source speaker (the TTS in the final application) is bilingual. Both UPC1 and UPC3 were trained using non-parallel data, but we were not in time to present to the TC-STAR evaluation the UPC2 method trained using non-parallel data.

Table 1 shows the results for the Spanish and English evaluations, both for intra and cross-lingual voice conversion. From the last line, we can see that the original source and target speaker voices are judged to be different (rated 1.96 and 1.52 for Spanish and English respectively), and to have a good quality (> 4.5 in both cases).

The results for Spanish show how the three methods perform similarly when changing the voice identity in intra-lingual conversion, with UPC2 slightly outperforming the

	SPANISH				ENGLISH			
	Intralingual		Crosslingual		Intralingual		Crosslingual	
	Identity	Quality	Identity	Quality	Identity	Quality	Identity	Quality
UPC1	3.35	3.2	X	X	3.83	1.61	X	X
UPC2	3.47	2.25	3.2	1.63	3.47	1.78	2.21	1.58
UPC3	3.18	2.37	2.79	2.33	2.92	2.23	2.59	2.13
SRC-TGT	1.96	4.8/4.6			1.52	4.8		

Table 1: Evaluation results in intra and cross-lingual voice conversion, for Spanish and English.

other two. In terms of quality, UPC1 clearly outperforms the other two methods, that are rated very similarly. For the Spanish cross-lingual evaluation, the performance in identity degrades for both UPC2 and UPC3. As in the intralingual case, UPC2 performs better than UPC3 in the identity evaluation. However, the quality of UPC2 decreases down to a non-acceptable degree.

The quality of both intra-lingual UPC1 and UPC2 methods applied to the English database severely degrades with respect to the Spanish case. On the contrary, UPC3 only shows a minor degradation. The identification capabilities of the UPC2 and UPC3 methods do not significantly change, and UPC1 gets better results according to the MOS results. In cross-lingual conversion, UPC3 suffers a small degradation of the identity capabilities, while maintaining the same degree of quality. UPC2, on the other hand, suffers a severe degradation of the identity, and a lighter decrease in quality.

6 Conclusions

This paper reports the different methods for voice conversion presented to the second evaluation campaign of the TC-STAR project. The main goal was to make the systems text-independent, so that they did not require aligned sentences. Our first method, the back-end of our TTS system, is based on a *small* amount of source training data. It is also used to create sentences aligned to the training target data to be used by the other two methods. In our second method, CART are used to split the acoustic space based on phonetic features. For each class, a linear regression is applied to transform the LSF coefficients. Then, the appropriated residual is selected from the residuals found in the training data based on the similarity of the associated LSF and the transformed LSF. The third method is based on the deterministic plus stochastic speech model, where the speech signal can be represented as a sum of a number of sinusoids and a noise-like component. Vocal tract conversion is linked to the amplitudes of the sinusoids, and special care is taken to avoid phase variations. The last step involves the prediction of the stochastic component. In all cases, a prosody scaling is performed to adequately change the F_0 . It is somewhat unexpected that the TTS back-end (UPC1) is not rated highest in terms of speaker identity, since the speech waveforms are derived directly from the target training data. This could be explained considering the artifacts introduced during the concatenation process, due to the reduced size of the database. The degradation of the UPC1 and UPC2 methods for English compared to the Spanish evaluation, could be due to the automatic seg-

mentation of the databases. Both methods use phonetic information to perform the conversion, and are then highly dependent on the segmentation quality. UPC3, on the other hand, has achieved a good balance between speech quality and speaker identity transformation for both intra and cross-lingual voice conversion, using non-aligned data. In future works, it is expected that a deeper study of the stochastic component will lead to important improvements. Although UPC2 was presented to the TC-STAR evaluation using parallel data, informal results show that the use of non-aligned sentences does not degrade the performance neither on speaker identity nor in speech quality significantly.

7 Acknowledgements

This work has been funded by the European Union under the integrated project TC-STAR - Technology and Corpora for Speech to Speech Translation (IST-2002-FP6-506738, <http://www.tc-star.org>).

8 References

- A. Bonafonte, H. Höge, I. Kiss, A. Moreno, U. Ziegenhain, H. van den Heuvel, H.U. Hain, X. S. Wang, and M. N. Garcia. 2006. TC-STAR: Specifications of language resources and evaluation for speech synthesis. In *LREC*, Genoa, Italy.
- H. Duxans and A. Bonafonte. 2006. Residual Conversion versus Prediction on Voice Morphing Systems. In *International Conference on Acoustics, Speech, and Signal Processing*.
- H. Duxans, A. Bonafonte, A. Kain, and J. van Santen. 2004. Including dynamic and phonetic information in voice conversion systems. In *International Conference on Spoken Language Processing*.
- Amro El-Jaroudi and John Makhoul. 1991. Discrete all-pole modeling. *IEEE Transactions on Signal Processing*, 39(2):411–423, February.
- D. Erro and A. Moreno. 2005. A pitch-asynchronous simple method for speech synthesis by diphone concatenation using the deterministic plus stochastic model. In *Proc. SPECOM*.
- A. Kain. 2001. *High resolution voice transformation*. Ph.D. thesis, OGI school of science and engineering.
- D. Sündermann, A. Bonafonte, H. Ney, and H. Höge. 2005. A Study on Residual Prediction Techniques for Voice Conversion. In *International Conference on Acoustics, Speech, and Signal Processing*.
- Yannis Stylianou, Olivier Cappé, and Eric Moulines. 1998. Continuous Probabilistic Transform for Voice Conversion. *IEEE Transactions on Speech and Audio Processing*.