

VOICE CONVERSION USING K-HISTOGRAMS AND RESIDUAL AVERAGING

A.J. URIZ[†], P.D. AGÜERO[‡], J.C. TULLI[‡], J. CASTIÑEIRA MOREIRA[†],
E.L. GONZÁLEZ[‡] and A. BONAFONTE[§]

[†] CONICET -Facultad de Ingeniería, Universidad Nacional de Mar del Plata, 7600 Mar del Plata, Argentina.
ajuriz@conicet.gov.ar

[‡] Facultad de Ingeniería, Universidad Nacional de Mar del Plata, 7600 Mar del Plata, Argentina, pdaguero@fi.mdp.edu.ar

[§] Universitat Politècnica de Catalunya, Barcelona, Spain.

Abstract — The main goal of a voice conversion system is to modify the voice of a source speaker, in order to be perceived as if it had been uttered by another specific speaker. Many approaches found in the literature convert only the features related to the vocal tract of the speaker. Our proposal is to convert those characteristics, and to process the signal passing through the vocal chords. Thus, the goal of this work is to obtain better scores in the voice conversion results.

Keywords— Voice Conversion, K- Histograms, Residual Conversion, Voice Synthesis.

I. INTRODUCTION

The primary goal of voice conversion systems is to modify the voice of a source speaker, in order to be perceived as if it had been uttered by another specific speaker, the target speaker. For this purpose, relevant features of the source speaker are identified and replaced by the corresponding features of the target speaker.

Several voice conversion techniques have been proposed since the problem was first formulated in 1988. In this year, Abe *et al.* (1988) proposed to convert voices by mapping codebooks created from a parallel training corpus. Since then, many authors tried to avoid spectral discontinuities caused by the hard partition of the acoustic space, by means of fuzzy classification or frequency axis warping functions. The introduction of statistical methods based on gaussian mixture models (GMM) for spectral envelope transformation was an important breakthrough in voice conversion (Kain, 2001; Stylianou *et al.*, 1998). In these approaches, the acoustic space of speakers is partitioned into overlapping classes and the weighted contribution of all classes is considered when transforming acoustic vectors. As a result, the spectral envelopes are successfully converted without discontinuities, but as a downside, the quality of the converted speech was degraded by over-smoothing.

The most recent approaches (He *et al.*, 2002; 2005) use non-numerical clustering techniques to make the conversion. An example of these systems is Uriz *et al.* (2009), which uses a clustering algorithm based on k-histograms (KH): a non-numerical clustering algorithm presented by He *et al.* (2005). These systems have a better performance, with no conditions about a specific distribution for the data. Thus, the cluster is adjusted to the

data distribution.

Nevertheless, the problem of creating high-quality voice conversion systems that could be used in real-life applications has not been completely solved. At present, there still is a trade-off between the similarity of converted voices to target voices, and the quality achieved by different conversion methods. The best way to overcome this trade-off is to convert both the vocal tract features of the speaker, and the residual signal of the phonation.

This problem has been faced in other works (Chen *et al.*, 2003; Kain, 2001), where the research was focused on increasing the resolution of GMM-based systems through residual prediction (Duxans and Bonafonte, 2006; Hanzlicek, 2006; Sundermann *et al.*, 2005) in order to improve both the quality scores and the converted-to-target similarity.

This paper proposes a voice conversion (VC) system that combines clustering by means of K-Histograms to convert the vocal tract features with an averaging of residual signals obtained from a pre-recorded dataset for converting the excitation of the voice. This is made to improve both the quality scores and the converted-to-target similarity.

This paper is organized as follows. In Section II, the most important aspects of the voice conversion techniques used in the work are explained in detail. In Section III, a new voice conversion method is proposed. In Section IV, the results of the objective and subjective tests are presented and discussed. Finally, the main conclusions are summarized in Section V.

II. VOICE CONVERSION

The goal of voice conversion systems is to convert the voice of a source speaker, so it is perceived as being pronounced by another specific speaker, who is called the target speaker. The next subsections describe the most important aspects of voice conversion systems.

A. Source-Filter Model

The **source-filter model** (Huang *et al.*, 2001) is a representation of the phonatory system as a filter being excited by a source signal. The filter that represents the vocal tract of the speaker is modeled by using a series of coaxial tubes. This is made by using an n degree polynomial, where n is the number of tubes used in the model. The coefficients of this polynomial are called Linear Predictive Coding (LPC). The source of the system is

the air flow that comes from the lungs and passes through the vocal cords. On the other hand, the inverse model of this filter is a widely used tool to analyze speech, because it decomposes the voice into the excitation and the vocal tract model using LPC coefficients.

The next subsections describe the techniques used in this work to convert each component of the vocal tract model.

B. Vocal Tract Conversion

Line Spectral Frequencies (LSF) transformation

In the most popular voice conversion systems pairs of source-target Line Spectral Frequencies (LSF) parameters (Huang *et al.*, 2001) are modeled using an approach of Gaussian Mixture Models (GMM) (Kain, 2001; Stylianou *et al.*, 1998). In some cases, the initialization of the parameters of the model is done applying the k-means clustering algorithm. In this work, quantized LSF coefficients are clustered using k-histograms and source parameters are transformed into target parameters through a non-gaussian approach via the cumulative density function (CDF).

The k-means algorithm is one of the most widely used clustering algorithms. Given a set of numeric objects $X_i \in D$ and an integer number k , the k-means algorithm searches for a partition of D into k clusters that minimizes the within groups sum of squared errors (WGSS). This process can be formulated as the minimization of the function $P(W, Q)$ with respect to W and Q , as shown in Eqs. 1 and 2.

$$\text{Minimize } P(W, Q) = \sum_{i=1}^k \sum_{l=1}^n w_{i,l} d(X_i, Q_l), \quad (1)$$

$$\text{Subject to } \sum_{l=1}^k w_{i,l} = 1, 1 \leq i \leq n, \quad (2)$$

$$w_{i,l} \in \{0,1\}, 1 \leq i \leq n, 1 \leq l \leq k,$$

In Eqs. 1 and 2, W is an $n \times k$ partition matrix which assigns each vector X_i to one cluster. $Q = \{Q_1, Q_2, \dots, Q_k\}$ is a set of objects in the same object domain (usually known as centroids of the clusters), and $d(.,.)$ is the definition of distance between vectors.

Clustering using k-histograms

K-histograms are an interesting approach to cluster categorical data. Each cluster is represented by the histograms of the elements of that cluster. Assuming that each element X_i is a vector of m categorical values $x_{i,1} \dots x_{i,m}$, Eq. 1 can be adapted to categorical data defining a distance based on the histograms of the cluster, as shown in Eq. 3.

$$\text{Minimize } P(W, H) = \sum_{i=1}^k \sum_{l=1}^n w_{i,l} d(X_i, H_l), \quad (3)$$

where $w_{i,l}$ is the partition matrix. The distance d compares the histograms of the cluster of each element. The clustering algorithm is explained in detail by He *et al.* (2005).

In this paper, k-histograms are used to partition into sets the vectors of features (LSF parameters) utilized in voice conversion. The LSF parameters are discretized to estimate the counts in the histograms of each set. The

source and target LSF vectors are aligned in the training set, and they are jointly partitioned using k-histograms. Then, when estimating by using histograms we make no assumptions about a particular distribution of the parameters.

The conversion between source and target parameters using histograms is performed applying a non-gaussian to non-gaussian mapping via the cumulative distribution function (CDF) coefficient by coefficient, as shown in Eq. 4.

$$\hat{y}_i = F_{y_j}^{-1} [F_{x_j}(x_i)], \quad (4)$$

The LSF parameter x_i of the source speaker is mapped into the target LSF parameter \hat{y}_i using the CDF of source and target i^{th} LSF parameter and j^{th} set (F_{x_j} and F_{y_j} respectively). The different available sets are obtained using the partition of the LSF parameter space via the k-histograms clustering technique.

The decision about the set j used in the transformation of a given source feature vector x is performed calculating the joint probability of each component of the vector (of dimension K) for each possible set (Eq. 5).

$$p_j = \sum_i^K \log(f_{x_j}(x_i)), \quad (5)$$

In Eq. 5, f_{x_j} is the probability that the coefficient x_i belongs to set j . The vector belongs to the set j with the highest probability p_j .

The parameters estimated by means of Eq. 4 are used to perform the synthesis of the target speech. The next subsection explains the proposed conversion method based on the LSF transformation shown in this section.

Voice conversion using k-histograms

The voice conversion algorithm using k-histograms can be described in four steps: windowing and parameterization, inverse filtering, parameter transformation and resynthesis. In the first step, each utterance is divided into overlapping pitch synchronous frames with a width of two periods. An asymmetrical Hanning window is used to minimize boundary effects. The parameterization consists of a 20^{th} order LSF vector. Then, the source excitation (the residual of LPC estimation) is calculated via inverse filtering with the LPC parameters obtained in each frame.

In the third step, the LSF parameters are transformed using the CDF estimated for the set with the highest probability calculated as shown in Eq. 5. The transformation includes a discretization of the LSF parameters that span from 0 to π . The degree of discretization is an adjustable parameter and it is directly related to the amount of available data to estimate histogram counts. The estimated CDF is obtained by means of the training process, where source and target LSF parameter vectors are aligned to obtain the mapping function using k-histograms. The alignment information is extracted from phoneme boundaries provided by a speech recognizer. Inside the boundaries of a frame, the alignment is proportional.

Finally, the transformed LSF parameters are transformed into LPC coefficients, and they are used to obtain the target converted voice by filtering the source excitation. The fundamental frequency is transformed using a mean and standard deviation normalization and the signal is resynthesized using TD-PSOLA (Moulines and Chanpentier, 1990).

In Figure 1 it is shown a scheme of a simple system that makes voice conversion only using K-Histograms, without any transformation of the source residual signal (e_s).

Although K-Histograms is an approximation that uses statistical tools likewise the GMM model (Stylianou *et al.*, 1998), Uriz *et al.* (2009) obtained a better conversion with this non-gaussian approach, without introducing assumptions about the distribution of the LSF coefficients. The main drawback of this proposal is the discretization of LSF parameters that introduces noise in the estimation. This is reduced by using a high quantity of levels to discretize the cumulative distribution function (3140 bins), and consequently, this error is negligible.

C. Residual Conversion

Although methods based on transformation of vocal tract obtain good performances, some works (Duxans and Bonafonte, 2006; Sundermann *et al.*, 2005) establish that a certain percentage of the identity of the speaker is contained into the residual signal resulting from decomposing the voice using the source-filter model. Consequently, it is necessary to find a system capable of converting voices that not only uses information of the vocal tract to map the acoustic space between speakers, but also it is necessary to take into account the residual signal for the conversion. There are several research lines, and the trivial solution is to resynthesize the voice by copying the source residual signal. This solution is used in some popular research works, and it corresponds to systems similar to that shown in Fig. 1. However, the residual signal has information about the identity of the speaker. Then, it is possible to establish a temporal correlation between the features of the vocal tract and the residual signal.

Consequently, it is important to resynthesize the voice using converted residual signals. In the literature, several works (Duxans and Bonafonte, 2006; Hanzlicek, 2006; Sundermann *et al.*, 2005) utilize parameters related to the vocal tract to obtain a residual signal from a prerecorded database in a training stage. Then, the baseline residual selection method is based on a prerecorded database, where there are stored pairs of target residual signals and LSF vectors of the source speaker tempo-

Figure 1: System based on K-Histograms, without modification of the residual signal e_s .

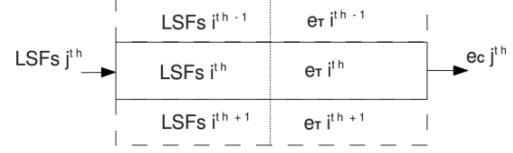


Figure 2: Structure of the database.

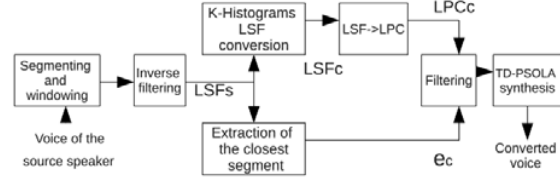


Figure 3: Baseline System. e_c is obtained from the closest element into the database.

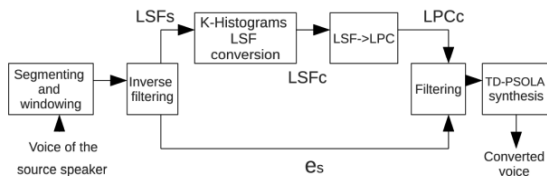
rally aligned. Then, each LSF source vector is used to select the closest LSF source vector from the database by computing the Euclidean distance. Thus, the correspondent residual signal is extracted of the database, and the voice signal is resynthesized using it.

In Fig. 2 the scheme of the database is shown. It is composed of pairs of LSF source vectors and target residual signals aligned during a training stage. Then, when the j^{th} segment is processed, the corresponding LSF source vector is compared with all the LSF vectors contained into the database applying the euclidean distance. Once the LSF vector that minimizes that distance is selected, the target residual signal associated with it is selected to resynthesize the j^{th} segment of speech. A scheme of the described system is shown in Fig. 3.

Figure 3 shows the proposed baseline system. This is divided into two blocks, the first block obtains the converted vocal tract parameters (LSF_c), and the second block, the converted residual signal (e_c). Once obtained LSF_c and e_c , the signal is filtered and resynthesized by means of the Time Domain Pitch Synchronous Overlapping and Add technique (TDPSOLA). Although the proposed method has a better identity than the method that uses the source residual signal, discontinuities appear due to the procedure of frame selection. It has the disadvantage of choosing consecutive frames from different places of the data base, which were pronounced in different phonetic contexts. Thus, voice signals resynthesized using these methods have a lower quality due to differences between consecutive frames. In the next Section a smoothed version of the residual selection is presented, which uses residual averaging to reduce the discontinuities between consecutive frames.

III. PROPOSED SYSTEM

As was mentioned in the previous Section, the voice signal resynthesized using residual conversion by means of codebooks has artifacts and discontinuities due to the procedure of frame selection. This process selects residual frames from the data base by computing an euclidean distance between the LSFS and each LSF source vec-



tor of the database. Once the vector that minimizes that distance is selected, the target residual signal associated with them is used to resynthesize the speech. As a result the procedure is based on an Euclidean distance that only selects frames that minimize the local distance, and the concatenation cost is not taken into account. Then, it is possible to take consecutive frames from distant places of the database, which may be pronounced in different phonetic contexts. This produces problems in the synthesized audio. In order to reduce this problem, the literature (Abe *et al.*, 1988; Dutoit *et al.*, 2007) takes into account a concatenation cost jointly with the local cost to choose the residual frame from the database. Although this method reduces the artifacts, the quality of the synthetic voice is low, and the computational cost associated with this kind of system is higher than others. Our proposal is based on averaging techniques, which are widely used in image processing to reduce the noise in pictures. The goal of these techniques is to reduce the noise of an image by averaging several copies of it. These methods are based on the supposition that the image is contaminated by Gaussian noise with a mean equal to zero. Consequently, the averaging of the image allows to obtain the signal plus the mean value of the noise (zero).

Our proposal is to use signal averaging to reduce the noise in the residual signal. The literature (Huang *et al.*, 2001) establishes that the residual signal can be modeled depending on the phonation characteristics. Then, for voice sounds, the residual signal can be represented as the sum of a train of impulses and white noise, while an unvoiced sound is represented by means of white noise. Then, if a series of similar residual signals are extracted from the database, and they are adjusted in length and averaged, a smoothed residual signal is obtained. If this procedure is repeated for all resynthesized segments, the resulting smoothing will reduce the differences between consecutive segments. This will be reflected in a better quality in the converted audio.

In this paper we propose to average target residual signals to obtain a vector, which will be used to resynthesize the voice signal. The averaging is made using the m closest target residual signals, extracted by the database through the LSF source vectors. The number m of averaged frames determines the degree of smoothing of the obtained vector. The main blocks of the system are presented in Fig. 4.

In Fig. 4 it is possible to see that the LSF source vectors (LSF_s) are used to obtain an approximation of the LSF target vector (LSF_c), which is made applying the K-Histograms algorithm. Also the LSF_s are used to select the m closest LSF source vector (LSF_s) from the database. Thus, the correspondent m target residual signals are adjusted in length so that they can be added and averaged. Then, the resulting residual signal is filtered using the LPC parameters obtained from LSF_c , and through TD-PSOLA, the signal is resynthesized. An important factor is the number m of averaged residual signals. A low value of m will not contribute to a good

quality because the smoothing is poor. On the other hand, a high value of m generates an over smoothed signal,

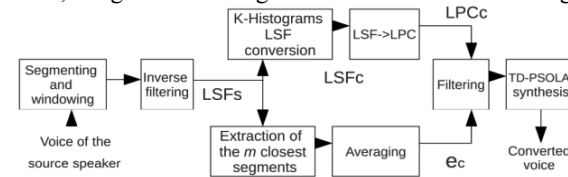


Figure 4: Proposed System. e_c is obtained from the database by averaging the k closest elements.

nal, which can affect the identity of the obtained speaker and the content of the message. For example, excessive filtering would eliminate sounds associated with fricatives.

In order to test the performance of the system, three implementations are presented. The first generates the frame to resynthesize the voice by averaging two segments of the data base. The second one uses five frames to obtain the averaged segment. Finally, an implementation that uses twenty segments from the database to obtain, by averaging, a residual converted signal. In the next Section, objective and subjective experiments analyze the performance of each system.

IV. EXPERIMENTS

The audio database used for the experiments contained 205 sentences in Spanish uttered by two male and two female speakers. The sampling frequency was 16 KHz and the average duration of the sentences was 4 seconds. 55% of the sentences were used to train the conversion functions, while 30% were kept as development set (to tune model parameters) and 15% were used to perform the objective and subjective tests.

One male and one female speaker were chosen as source, and the other two speakers were used as targets. Four different conversion directions were considered: male to male, female to female, male to female and female to male. 31 unseen sentences during training were converted and resynthesized for all methods. The results will be shown by merging all speakers.

For the proposed method, we will consider a quantization resolution of 3140 bins for the histograms, since this ensures a low level of noise quantization.

Four averaging levels are used to test the system: 1 vector (without averaging) ($m=1$), 2 vectors ($m=2$), 5 vectors ($m=5$) and 20 vectors ($m=20$). A fifth system that employs privileged information was developed. This system, noted in this paper as REFERENCE, which was presented in Uriz *et al.* (2009), resynthesizes the signal using the converted LSF parameters, and the residual signal corresponding to the real target speaker. It is a measure of the highest achievable quality and identity by the proposed method. The main aim of the averaging is to reduce the inter-frame discontinuities, which may produce an improvement in the quality of the synthetic voice. Figure 5 shows the proposed reference system.

Some results will be shown using boxplots (Tukey, 1970). This representation is an useful statistical tool to compare several statistical distributions. In our case we will use it to compare the distribution of the scores of

the different systems to study the significance of the differences.

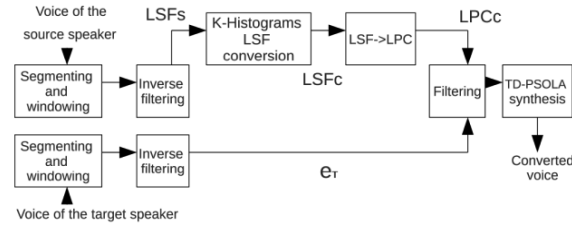


Figure 5: Architecture of the reference system. The excitation used in the resynthesis (e_t) is obtained from the target speaker.

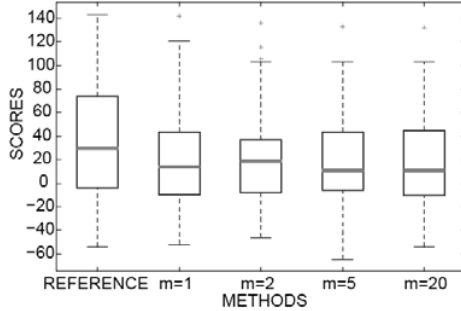


Figure 6: Objective test.

A. Objective Results Using a Small Speaker Verification System

In this work, we evaluate the proposed methods using a small speaker verification system (SVS) based on a GMM model (Reynolds, 1995). MFCC coefficients are used to code the voice signal using frames of 20ms. Two GMM models were trained using evaluation data to build source and target models. Given an utterance of a converted voice, these models may be utilized to establish the closeness to source and target. The subtraction of the log-likelihood of source and target models is an indicator (score) of the performance of the conversion. A positive score indicates a good conversion, while a negative score is an indicator of closeness to source voice model.

Figure 6 shows the results of the different algorithms. Methods are ordered according to the median of the scores. The results show that the method with $m=2$ has the higher performance without using privileged information. These are within expectations. Although the filtering is performed to reduce noise, the over smooth generated by filtering in the cases where $m=5$ and $m=20$ causes a decreasing in the resynthesized speaker identity.

The most important result obtained in this experiment is the fact that the identity of the resynthesized voice is not degraded by using averaged segments.

B. Subjective Experiments

The subjective tests were conducted with 31 sentences unseen during training. 15 volunteers were asked to listen to the target converted sentences in random order. Listeners were asked to judge the similarity of the voices to the target using a 5-point scale, from 1 (totally identical to source) to 5 (totally identical to target). On the other hand, the listeners were also asked to rate the

quality of the converted sentences from 1 point (bad) to 5 points (excellent).

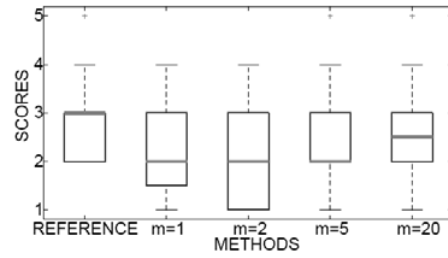


Figure 7: MOS test of Quality.

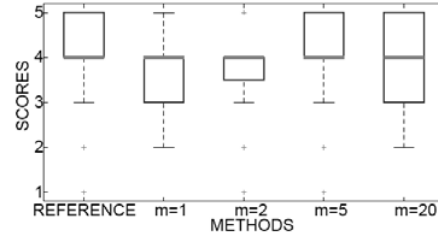


Figure 8: MOS test of Identity.

Figure 7 shows the scores for quality MOS test using boxplots. In this figure it can be seen that except for the reference system, the proposed system has increasing scores according to the degree of averaging. Also, the distance between the first and the third quartiles is reduced when the averaging ins increased, thus the residual averaging increases the consistence of the quality of the systems.

Figure 8 shows the MOS test of identity using boxplots, which presents similar results of the objective test of identity. In the figure it can be seen that the most consistent method is $m=2$, while the method with $m=20$, is the less consistent system because the distance between the first and the third quartiles is the highest.

V. CONCLUSIONS AND FUTURE WORK

In this paper, a voice conversion algorithm based on K-Histograms and averaging of residual signals stored in a database was presented.

Objective and subjective experiments show that the proposed method has a higher performance in the converted voices. Since the level of similarity of the resulting audios remains the same, a better trade-off of similarity and quality than the system that resynthesizes the voice using only one residual segment is obtained.

This paper shows that the results of the system using residual averaging are slightly lower than the reference system, which uses privileged information for the excitation of the converted voice. This is important because the main limitation of the performance of this system is the size of the pre-recorded database. Thus, when increasing the size of the database, the performance of the system would increase to become closer to the reference value.

In the future work we will include another state-of-the-art methods to enhance the processing of the excitation. Consequently, it is expected that the quality of the complete voice conversion will be improved.

REFERENCES

- Abe, M., S. Nakamura, K. Shikano and H. Kuwabara, "Voice conversion through vector quantization," *Proc. of ICASSP*, New York, USA, 655-658 (1988).
- Chen, Y., M. Chu, E. Chang, J. Liu and R. Liu, "Voice conversion with smoothed GMM and MAP adaptation," *Proc. of the European Conf. on Speech Communications and Technology*, Geneva, Switzerland, 2413-2416 (2003).
- Dutoit, T., A. Holzapfel, M. Jottrand, A. Moinet, J. Perez and Y. Stylianou, "Towards a voice conversion system based on frame selection," *Proc. of ICASSP*, Honolulu, USA, 513-516 (2007).
- Duxans, H. and A. Bonafonte, "Residual conversion versus prediction on voice morphing systems," *Proc. of the IEEE ICASSP*, 85-88 (2006).
- Hanzlicek, Z., "On residual prediction in voice conversion tasks," *Speech Processing, Institute of Radio Engineering and Electronics*, Prague, 90-97 (2006).
- He, Z., Xu, X. and Deng, S., "Squeezer: An Efficient Algorithm for Clustering Categorical Data," *J. Comput. Sci. Technol.*, **17**, 611-624 (2002).
- He, Z., Xu, X. and Deng, S., "K-Histograms: An Efficient Clustering Algorithm for Categorical Dataset," *Proceedings of CoRR* (2005).
- Huang, X., A. Acero and H.W. Hon, *Spoken language processing. A guide of theory, algorithm, and system development*, Prentice Hall, New York (2001).
- Kain, A., *High resolution voice transformation*, PhD Thesis OGI-OHSU, Oregon, USA (2001).
- Moulines, E. and F. Chanpentier, "Pitch synchronous waveform processing techniques for Text-to-Speech synthesis using diphones," *J. Speech Communication*, **9**, 453-467 (1990).
- Reynolds, D.A., "Speaker identification and verification using gaussian mixture speaker models," *J. Speech Communication*, **17**, 91-108 (1995).
- Stylianou, Y., O. Cappe and E. Moulines, "Continuous probabilistic transform for voice conversion," *Proc. of ICASSP*, Seattle, USA, 131-142 (1998).
- Sundermann, D., H. Hoge, A. Bonafonte and H. Duxans, "Residual prediction," *Proc. of the IEEE IS SPIT*, Athens, Greece, 512-516 (2005).
- Tukey, J.W., *Exploratory Data Analysis*, Addison Wesley, Boston, (1970).
- Uriz, A.J., P.D. Agüero, B. Bonafonte and J.C. Tulli, "Voice conversion using K-Histograms and frame selection," *Proc. of Interspeech*, Brighton, U.K., 1639-1642 (2009).

Received: May 5, 2012

Accepted: October 12, 2012

Recommended by Subject Editor: Gastón Schlotthauer, María Eugenia Torres and José Luis Figueroa