INTERSPEECH 2009 BRIGHTON

Voice Conversion using K-Histograms and Frame Selection

Alejandro José Uriz[1], *Pablo Daniel Agüero*[1], *Antonio Bonafonte*[2], *Juan Carlos Tulli*[1]

[1]Facultad de Ingeniería,Universidad Nacional de Mar del Plata, Mar del Plata, Argentina
[2] Universitat Politècnica de Catalunya, Barcelona, Spain

ajuriz@fi.mdp.edu.ar, pdaguero@fi.mdp.edu.ar

Abstract

The goal of voice conversion systems is to modify the voice of a source speaker to be perceived as if it had been uttered by another specific speaker. Many approaches found in the literature work based on statistical models and introduce an oversmoothing in the target features. Our proposal is a new model that combines several techniques used in unit selection for text-to-speech and a non-gaussian transformation mathematical model. Subjective results support the proposed approach.

Index Terms: speech synthesis, voice conversion, frame selection, non-gaussian transformation

10.21437/Interspeech.2009-489

1. 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 1985. In this year Childers et al. [1] proposed to convert voices through mapping parameters from a speaker to other different speaker. 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 appearance of statistical methods based on gaussian mixture models (GMM) for spectral envelope transformation was an important breakthrough in voice conversion [2, 3]. The acoustic space of speakers was partitioned into overlapping classes and the weighted contribution of all classes was considered when transforming acoustic vectors. The spectral envelopes were successfully converted without discontinuities, but in exchange the quality of the converted speech was degraded by over-smoothing. This problem was faced in further works [4, 5], while the usage of GMM-based techniques became almost standard, up to the point that the research was focused on increasing the resolution of GMM-based systems through residual prediction [2, 6] in order to improve both the quality scores and the converted-to-target similarity.

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 is still a trade-off between the similarity of converted voices to target voices and the quality achieved by the different conversion methods.

Another interesting approach focused on improving target speaker identity is the frame selection proposal of Dutoit et al. [7]. In that paper the authors propose to find the optimal sequence of frame target features in training data reducing the distance between source converted features using GMM and target features by means of the Viterbi algorithm, which was also used in the work of Salor and Demirekler [8]. Sündermann [9] proposed a similar approach just using the source features without any conversion.

In this paper we propose two systems that work using a new approach based on a non-gaussian statistical transformation and frame selection. In order to compare the system's performances, we made experiments with other two state-of-the-art techniques: GMM [3](a gaussian statistical transformation) and Dutoit's method [7](a voice conversion algorithm based on frame selection).

This paper is organized as follows. In Section 2, the new voice conversion techniques are explained in detail. In Section 3 two voice conversion methods are proposed. In Section 4 the results of the objective and subjective tests are presented and discussed. Finally, the main conclusions are summarized in Section 5.

2. LSF conversion using k-histograms

In many voice conversion systems pairs of source-target LSF vectors are modelled using an approach of Gaussian Mixture Models (GMM). In some cases the initialization of the parameters of the model is done using the k-means clustering algorithm. In this paper we propose to cluster quantized LSF coefficients using k-histograms and transform source parameters into target parameters through a non-gaussian approach via the cumulative density function (CDF).

The k-means algorithm is one of the mostly 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 equations 1 and 2.

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

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

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

where W is an $n \times k$ partition matrix which assigns each vector X_i to one cluster, $Q = \{Q_1, Q_2, ..., Q_k\}$ is a set of

Copyright © 2009 ISCA

objects in the same object domain (usually known as centroids of the clusters), and $d(\cdot,\cdot)$ is the definition of distance between vectors.

2.1. Clustering using k-histograms

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

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

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 [10].

In this paper we propose to use k-histograms to partition the vectors of features (LSF parameters) used in voice conversion into sets. 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.

This approach intends to avoid the assumption made in GMM-based voice conversion system about the possibility to approximate the distribution of each LSF coefficient through a mixture of gaussians. In our proposal we do not include any assumption about a particular distribution of the parameters by estimating it using histograms.

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

$$\hat{y_i} = F_{y_j}^{-1}[F_{x_j}(x_i)] \qquad (4)$$

The LSF parameter x_i of 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 (Equation 5).

$$p_j = \sum_{i}^{K} log(f_{x_j}(x_i)) \qquad (5)$$

where 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 using Equation 4 are used to perform the synthesis of the target speech. In the next section two voice conversion methods will be explained based on the LSF transformation shown in this section.

3. Voice conversion systems

In this paper we show two different implementations of voice conversion using k-histograms. In the first one we perform speech synthesis after parameter conversion. On the other hand, the second proposed method also includes a frame selection process using dynamic programming to search the optimal sequence of target feature vectors, avoiding the smoothing introduced by statistical mapping via k-histograms.

3.1. Voice conversion using k-histograms

The voice conversion algorithm using k-histograms (**KH**) has four steps in our experiments: windowing and parameterization, inverse filtering, parameter transformation and resynthesis.

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. The source excitation (the residual of LPC estimation) is calculated via inverse filtering with the LPC parameters obtained in each frame.

During the training process source and target LSF parameter vectors are aligned to obtain the mapping function using k-histograms. The alignment information is extracted from phone boundaries provided by a speech recognizer. Inside the boundaries of a frame, the alignment is linear.

The LSF parameters are transformed using the CDF estimated for the set with the highest probability calculated as shown in Equation 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 the counts of the histograms.

The transformed LSF parameters are converted 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 PSOLA [11].

Figure 1 shows the scheme of our proposal. In this case we preferred to use the target excitation to study the accuracy of LSF parameter conversion without the influence of an inaccurate excitation estimation.

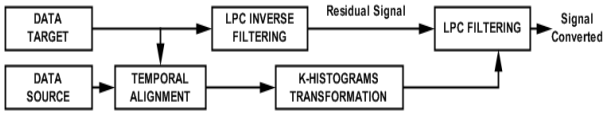

Figure 1: First Method proposed

Although the proposal is an aproximation that uses statistical tools likewise the GMM model [3], we expect to obtain a better conversion with this non-gaussian approach, without introducing assumptions about the distribution of the LSF coefficients. The main drawback of our proposal is the discretization of LSF parameters that introduces noise in the estimation. We studied in the experiments the influence of such quantization.

3.2. Voice conversion using k-histograms and frame selection

As stated in the introduction many systems transform the LSF vectors to find the transformed envelope. However, a novel approach used the transformed LSF to select real frames from the training data of the target speaker. As the motivation of using k-histograms instead of k-means does not depend on the final use of the transformed vectors, in this section we apply the k-histograms based transformation to the method proposed by Dutoit [7], and it is named **FSKH**.

In this case the transformation is divided in two stages (as shown in Figure 2:

- The first stage makes a transformation using the k-histograms method as explained in Section 3.1. The LSF parameters of source speaker x are transformed into $\hat{y}$.
- Then, the converted LSF parameters $\hat{y}$ are converted using a second stage based on frame selection to obtain a new set of transformed parameters $\hat{y'}$.

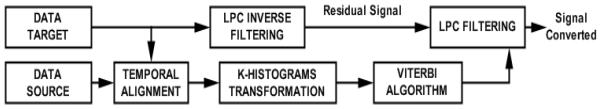

Figure 2: Second Method proposed

3.2.1. Frame selection stage

Given a sequence of converted feature vectors (via k-histograms) of source speaker ($\hat{y}$) we may find an optimal sequence of feature vectors of target speaker in training data ($\hat{y'}$). The optimal sequence is obtained using the formulation of Equation 6. This optimization problem is solved using the viterbi algorithm.

$$min_{\hat{y'}} \left[\sum_i \alpha d_t(\hat{y}_i, y_j^{train}) + (1-\alpha) d_c(\hat{y'}_{i-1}, \hat{y'}_i) \right] \quad (6)$$

In this expression, $d(\hat{y}_i, y_j^{train})$ represents the target cost which measures the distance between the converted source parameters of i^{th} frame and the target parameters of j^{th} frame in the training set. In this way we find appropriate converted target parameters according to converted source parameters. The acoustic parameters included in the target cost are LSF, energy, fundamental frequency and phone identity. Each phone is divided in three zones: start, medium and end. The phone identity is concatenated with the zone code to preserve the dynamics of phone evolution both for source and target frames.

The concatenation cost $d(\hat{y'}_{i-1}, \hat{y'}_i)$ minimizes the discontinuities between adjacent frames, and also favours the selection of consecutive frames. The parameters listed above are weighted to normalize their effects, and the weights are calculated using an automatic adjustment: MultiLinear Regression (MLR) [12].

A problem of computational load arises with the proposed conversion method: the size of the search space. The amount of frames in the lookup table is around $60,000$ for a 15 minutes database using pitch synchronous analysis. As a consequence, we decided to use the clustering provided by k-histograms to reduce the search space. Only the target frames in the training data that belong to the cluster assigned to the converted source frame are considered.

The fundamental frequency contour of the target speaker is obtained in the same way than the method proposed in Section 3.1.

4. Experiments

The audio database used for our experiments contained 200 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. 50% of the sentences were used to train the conversion functions, while 30% were kept as development set (to tune model parameters) and 20% were used to perform the objective test.

One male and one female speaker were chosen as source, and the other two speakers were used as target, so four different conversion directions were considered: male to male (m2m), female to female (f2f), male to female (m2f) and female to male (f2m). 38 sentences unseen during training were converted and resynthesized for all methods. The results will be shown by merging all speakers, because separate results show a correspondence with the global results.

For each of the two proposed methods we will consider two quantization resolutions: 314 and $3,140$ bins for the histograms. The original methods GMM and Dutoit's proposal are included as a reference.

A seventh voice conversion method was included in the experiments. It consists of finding the closest feature vector of target speaker in training data to the real feature vector of target speaker. This voice conversion method based on frame selection that uses privileged information is named **FSOPT**. It is a measure of the highest achievable quality and identity by the frame selection method.

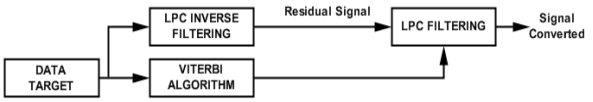

Figure 3: Architecture of FSOPT

Some results will be shown using box-plots [13]. 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.

4.1. Experimental results

In this work we evaluate the proposed methods using the P distance (see Equation 7). It was used to measure the closeness of the converted voice to the target voice using the seven voice conversion methods included in the experiments. The P distance was already used in several works about voice conversion [3].

$$P = 1 - \frac{d(y, \hat{y})}{d(x, y)} \quad (7)$$

The closer the converted parameters ($\hat{y}$) to the parameters of the target speaker (y) produces that P approaches to one. The distance between source parameters (x) and target parameters (y) allows to scale the P distance in the virtual path that goes from source to target parameters.

The differences in P score shown in Table 1 of the methods under study were not statistically relevant. For that reason we decided to focus in the subjective results.

The subjective test was conducted with 35 sentences unseen during training. 15 volunteers were asked to listen to the converted-target sentence in random order. Listeners were asked to judge the similarity of the voices to the target using a 5-point scale, from 1 (totally different to target) 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

	P	MOS-S	MOS-Q
FSOPT	0.41	3.6	2.8
KH3140	0.18	3.6	3.0
KH314	0.17	3.6	2.8
FSKH3140	0.17	3.1	2.3
FSKH314	0.16	3.4	2.3
DUTOIT	0.03	3.4	2.4
GMM	0.28	2.7	2.1

Table 1: P, MOS-S and MOS-Q scores for all systems under evaluation, target and source voices.

(bad) to 5 points (excellent). The resulting scores for similarity are shown in the box-plot of Figure 4.

The MOS of similarity (MOS-S) shows that the methods based on k-histograms have a better similarity to target voice than GMM and DUTOIT methods. The MOS of similarity and quality (MOS-Q) is identical to FSOPT. It is an important result taking into account the privileged information used by FSOPT.

The use of frame selection tends to degrade the similarity and quality of k-histograms methods, as shown columns MOS-S and MOS-Q of Table 1. However, there is not an important degradation depending on the different resolution of our experiments (314 and $3,140$ bins).

In the case of GMM transformation the use of frame selection improves its performance, as shown by Dutoit's proposal in our experiments. The similarity improves in 0.7 points and quality in 0.3 points.

The Wilcoxon test shows only statistical relevant differences ($p < 0.01$) in similarity scores of FSOPT, KH314, KH3140, FSKH314 and DUTOIT with respect to the other methods. The quality scores of all methods show statistical relevant differences ($p < 0.01$) in the Wilcoxon test, except between FSOPT, KH314 and KH3140.

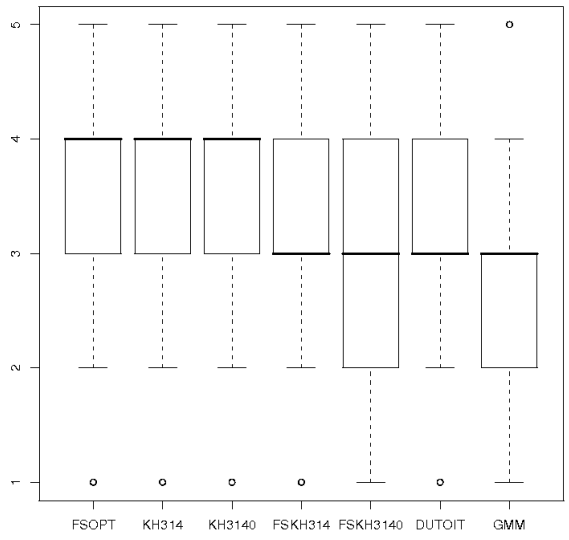

Figure 4: MOS of similarity to target voice

5. Conclusions

In this paper we presented a voice conversion algorithm based on a novel approach using a non-gaussian statistical transformation function. A second proposed method also incorporates a transformation based on frame selection.

Subjective experiments show that the method based on a non-gaussian statistical transformation has a better trade-off of similarity and quality than the other systems under evaluation, including our second proposed method that uses frame selection.

The quantization introduced in the LSF parameters to estimate the histograms and to transform source coefficients into target coefficients did not show an impact in the MOS.

Once that we have proved that k-histograms is a very good alternative to transform LSF coefficients in voice conversion, both for their direct use or for selecting target frames, we will extend the system with state-of-the-art methods to include excitation, so that the quality of the complete voice conversion system makes it usable.

6. References

[1] Childers, D.G. and Yegnanarayana, B. and Ke Wu, "Voice conversion: Factors responsible for quality", in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 748-751, 1985.

[2] Kain, A.,"High resolution voice transformation", in PhD thesis, OGI School of Science and Engineering, 2001.

[3] Stylianou, Y. and Cappe, O. and Moulines E.,"Continuous probabilistic transform for voice conversion", in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 6(2) : 131-142, 1998.

[4] Chen, Y. and Chu, M. and Chang, E. and Liu, J. and Liu, R., "Voice conversion with smoothed GMM and MAP adaptation", in Proceedings of the European Conference on Speech Communications and Technology, 2413-2416, 2003.

[5] Toda, T. and Saruwatari, H. and Shikano, K., "Voice conversion algorithm based on gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum", in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 841-844, 2001.

[6] Sundermann, D. and Hoge, H. and Bonafonte, A. and Duxans, H., "Residual prediction", in Proceedings of the IEEE Symposium on Signal Processing and Information Technology, 512-516, 2005.

[7] Dutoit, T. and Holzapfel, A. and Jottrand, M. and Moinet, A. and Perez, J. and Stylianou, Y., "Towards a Voice Conversion System Based on Frame Selection", in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 513-516, 2007.

[8] Salor, O. and Demirekler, M., "Voice transformation using principle component analysis based LSF quantization and dynamic programming approach", in Proceedings of Interspeech 2005, 1889-1892, 2005.

[9] Sundermann, D. and Hoge, H. and Bonafonte, A. and Ney, H. and Black, A. and Narayanan, S., "Text-Independent Voice Conversion Based on Unit Selection", in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 81-84, 2006.

[10] He, Z. and Xu, X. and Deng, S. and Dong, B., "K-Histograms: An Efficient Clustering Algorithm for Categorical Dataset", 2005.

[11] Moulines, E. and Charpentier, F.,"Pitch Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones", in Speech Communication 9, 453-467, 1990.

[12] Black, A. and Campbell, N., "Optimising selection of unit from speech databases for concatenative synthesis", in Proceedings Eurospeech, 581-584, 1995.

[13] Tukey, J.W., "Exploratory Data Analysis, Addison-Wesley",1970.

[14] Reynolds, D.A., "Speaker Identification and verification using Gaussian mixture speaker models", in Speech Communication 17, 91-108, 1995.