RESIDUAL CONVERSION VERSUS PREDICTION ON VOICE MORPHING SYSTEMS

Helenca Duxans and Antonio Bonafonte

Department of Signal Theory and Communication, TALP Research Center
Technical University of Catalonia (UPC), Barcelona, Spain

ABSTRACT

Many of the research efforts in Voice Morphing, or also called Voice Conversion (VC), has been carried out in the field of vocal tract mapping. It has been studied that in the vocal tract parameters there is the most relevant part of the information about speaker identity. However, to achieve an effective personality change it is also needed to modify the glottal flow characteristics of the source speaker. In this paper two strategies of transformation of LPC residual signals for a voice morphing system based in LSF mapping are compared: conversion of the source residual by codebook mapping and prediction of the target residual from LSF vectors. Experimental results demonstrates that the relationship between LSF parameters and their residual signals is higher that the relationship between LPC residual signals of two different aligned speakers.

1. INTRODUCTION

Voice Morphing or Voice Conversion systems modify a speaker voice (*source speaker*) to be perceived as if another speaker (*target speaker*) had uttered it. Until recently, many of previous published VC approaches have been centered on vocal tract mapping, whose features are parametrized by some related LPC parameters. The reason to focus on vocal tract is the main part of the information related to speaker identity is modeled by these features. However, it has already been reported that to achieve an effective voice morphing some kind of transformation has to be applied to the residual signals [1]. Although this statement has been tested many times, there are no many studies about residual signal transformation.

Previous works on residual signal transformation can be grouped in two categories. The first group tries to modify the source speaker residual signal to match with the target one [2, 3]. This is an idea similar to the vocal tract conversion, and we will call them *residual signal conversion systems*. The second group of systems are based on the prediction of the converted residual signal from the vocal tract parameter vector obtained with the vocal tract mapping system [1, 4, 5]. We will call this second group of approaches *residual signal prediction systems*.

The main goal of this paper is to compare both strategies and to study the relationship between LPC residual signals of one speaker with another speaker, and the relationship between LPC residual signals of one speaker with their vocal tract parameters. Also, a new strategy to store the residual signals and to reconstruct the transformed speech avoiding large noises due to phase discontinuities is presented. The method used to map the vocal tract parameters is based on a decision tree classification, which produces better results than the state-of-the-art GMM VC system [6].

The outline of this paper is as follows. We first review a vocal tract mapping system based on decision trees in section 2. Then, in section 3 and section 4 different possibilities for residual signal conversion and prediction systems are studied. The method used to reconstruct the transformed speech is explained in section 5. In section 6 perceptual results are presented and finally in section 7 there are the conclusions of this study.

2. VOCAL TRACT MAPPING SYSTEM BASED ON DECISION TREES

The vocal tract mapping system used in this study was designed to work as a post-processing block of a TTS system, where there is available other kinds of information (pitch marks, phonetic transcription ...) more than only speech samples.

In the training step, source and target speech training data were segmented in frames of two pitch period length. LSF parameters were estimated for each frame, which were inverse filtered to obtain the residual signal frames. For unvoiced regions an artificial fix *pitch* period length was used.

A good alignment between source and target training data is required to train the system. We have used lineal alignment using phoneme boundaries as anchor points. Both source and target frame repetitions were allowed. Then, the proper vocal tract mapping function was trained.

To estimate the vocal tract mapping function a CART decision tree has been used. It is based on the idea that the acoustic space of both speakers is organized in acoustic classes, and a conversion function can be estimated for each class. With the decision tree approach not only spectral informa-

This work has been partially sponsored by the European Union under grant FP6-506738 (TC-STAR project, http://www.tc-star.org) and the Spanish Government under grant TIC2002-04447-C02 (ALIADO project, http://gps-tsc.upc.es/veu/aliado).

tion is used to identify the classes, but also phonetic information. The tree extracts, at each splitting step, overlapping regions of the acoustic space that can be represented by only one acoustic class, modeled by a joint Gaussian probability function. The questions used to split the tree nodes are phonetic (voiceness, point of articulation, manner and a vowel/consonant flag) and the criteria to decided the better split is a spectral measure, in particular the accuracy of the regression:

$$\Delta(t,q) = E(t) - \frac{(E(t_L,q)|t_L|) + (E(t_R,q)|t_R|)}{(|t_L| + |t_R|)} \quad (1)$$

where $E(t,q)$ is the error index of the t node for the question q, $|t|$ indicates the number of spectral vectors of the training set belonging to the t node, and t_L and t_R are the child nodes of t. More details about the tree growing can be found in [6].

Once the tree has been grown, a conversion function is estimated for each leave with the form of Eq. 2, where i indicates the leave identifier and x/y the source/target speaker.

$$\hat{y}_i = \mu_i^{\mathbf{y}} + \mathbf{\Sigma}_i^{yx} \mathbf{\Sigma}_i^{xx^{-1}} (\mathbf{x} - \mu_i^x) \quad (2)$$

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 transformation function belonging to its leaf. In [6] it was reported that listeners preferred CART conversion over GMM conversion speech. Moreover, CART systems do not need any parameter tuning, like the number of GMM components.

3. RESIDUAL SIGNAL CONVERSION

To study the degree of relationship between source residual signals and target residual signals a system based on mapping codebooks has been used, similar to one of the first methods published in the voice conversion field [7], also applied to the residual problem in [3]. The main difference of the current paper is that we are interested in comparing conversion with prediction, so we are not concern about errors due to the quantization of the residual signal. This is the reason why in this study all the data available will be part of the mapping codebooks. It means that instead of reducing the dimensionality of the training data by building a codebook, we will keep each residual signal vector as a codeword. So, mapping codebooks are built with all the available pairs of source-target training residual signals.

To convert a new residual signal we must find the most similar source one, and replaced it for its aligned target residual. The similarity measure used is spectral distortion:

$$SD(r_1, r_2) = 20log\Big(\frac{1}{N}\sum_{n=0}^{N-1}\sqrt{(S_{r1}(n) - S_{r2}(n))^2}\Big) \quad (3)$$

where $S_1(n)$ and $S_2(n)$ have been power normalized.

Residual conversion has been applied only to voiced frames. Source residual signal has been kept for unvoiced frames, because it doesn't contain many glottal characteristics of the speaker.

4. RESIDUAL SIGNAL PREDICTION

In this section LPC residual signals are predicted from the LSF spectral envelopes, in contrast of converting source residual signals. This idea was first introduced by [1], and have been applied with different methodologies in [4] and [5].

In order to predict the residuals from their LPC parameters, the assumption that the residual is completely incorrelated with the spectral envelope is broken. For a particular speaker it is expected that the residuals corresponding to phones of acoustically similar classes are similar and predictable. The LPC residual signal contains all the effects that the spectral envelope estimation has not been able to model, such as nasalizations, phase errors due to the minimum phase assumption or details of the glottal pulse shape.

As in the residual conversion, we are not concerned about errors due to the quantization of the residual signal or errors due to the modelization of the search space into regions less populated. So, our prediction approach uses all the available information and the acoustic space is not hard or soft parted into classes. The strategy followed was to build parallel codebooks for LSF and their LPC residual signals associated, with as many codewords as vectors available in training.

The prediction procedure is similar to the explained in the previous section. Converted LSF are compared with all the codewords and the residual signal associated to the codeword more similar is chosen. The similarity measure for LSF is the Inverse Harmonic Mean Distance [8] calculated as:

$$D(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{p=1}^{P} c(p)(x(p) - y(p))^2} \quad (4)$$

$$c(p) = \frac{1}{w(p) - w(p-1)} + \frac{1}{w(p+1) - w(p)} \quad (5)$$

with $w(p) = argmax\{c(p)|_{w(p)=x(p)}, c(p)|_{w(p)=y(p)}\}$, $w(0) = 0$ and $w(P+1) = \pi$ (p is the vector dimension). Using this distance measurement we weight more the mismatch in spectral picks than the mismatch in spectral valleys.

As in residual conversion, residual prediction has been applied only to voiced frames.

5. SIGNAL RECONSTRUCTION STRATEGY

The reconstruction strategy applied to the converted LSF and transformed residual signals is the inverse filtering and TD-PSOLA method. A straightforward application of TD-PSOLA over the speech frames obtained from a complete voice morphing system results in noise concatenation errors, mainly due to no similarity criteria over neighbour pitch period residual signals is imposed. Previous published studies dealt with this particular class of noises through phase interpolation [1], phase prediction [4] or smoothing of residual signals over different periods [5].

In this paper a strategy to both alleviate concatenation noises and memory load requirements is presented. In order to avoid synthesis artifacts due to residual phase discontinuities from frame to frame a harmonic model for each of the transformed residuals is estimated [9]. The number of harmonic frequencies is the integer $\leq f_{sampling}/2f0$.

$$s(n) = \sum_{k=1}^{K} a_k cos(2\pi f_k n + \phi_k) \tag{6}$$

Then, the harmonic phases ϕ_k are modified so that continuity is assured in every voiced region. It means that the initial phase of the harmonic k of the frame i is calculated as:

$$\phi_k^i = 2\pi \frac{f_k^{i-1} + f_k^i}{2} N_k \tag{7}$$

Where N_k is the number of points between the center of the $(i-1)th$ frame and the ith frame. When a birth of a harmonic frequency occurs it is assigned a random initial phase.

Duration, loudness and speech rate of the source speaker are kept. The mean value of the pitch (μ) and its variance (σ) is estimated for source and target speakers from the training data. When the conversion is applied, the pitch of the utterance is modified to adjust its mean and variance according to Eq. 8.

$$f0_{converted} = \mu_{target} + \frac{\sigma_{target}}{\sigma_{source}}(f0_{source} - \mu_{source}) \tag{8}$$

6. EXPERIMENTS

The experiments were carried out with four speakers, two males and two females. Speech and laringograph signals were recorded in an acoustically isolated room. The corpus has been segmented into phonemes, and manually supervised for one male and one female speakers. We have used 30 sentences for training and 20 for test.

First, we have informally tested that keeping the source residual signal for unvoiced frames does not degrade the final quality of the target speech. So, we have focused on dealing with voiced frames.

To evaluate the performance of the proposed systems we have carried out three sets of experiments, each one with F0 modification according to Eq. 8:

1. Source LPC residual signal filtered by converted LSF.
2. Converted LPC residual signal filtered by converted LSF.
3. Predicted LPC residual signal filtered by converted LSF.

Three different tests have been evaluated. The first test consisted in ABX questions, where A is source or target speaker, B is the other one and X is one of the evaluation files. The three utterances are different from one another, and also from each ABX question. Listeners rated every question according to the following rule: 1-Very close to A; 2-Close to A; 3-Neither A nor B; 4-Close to B; 5-Very close to B.

Although ABX test is a direct way of evaluating conversion systems and very useful to compare different works because its use is very spread, it contains information that can help the listener to decide to rate the system in a better way. So, a similarity test was carried out too. A set of pairs of speech files were presented and listeners were asked to rate their similarity from 1 (different speakers) to 5 (the same speaker). One of the files of every pair was either the source or target speaker, and the other one corresponded to the converted voice. Finally, a MOS test has been carried out to evaluate the quality of the proposed systems. Ten listeners completed the evaluation.

Table 1 resumes the results of the ABX test. Each column corresponds to the proportion of times that listeners has rated the converted signal as source (very close or close), target speaker or neither of them. According to this results, it is determinant to include some kind of residual signal treatment to the converted voice to achieve the personality modification. Slightly better results were obtained with Predicted Residuals than with Converted Residuals. To be more precise, table 3 contains the numerical results of the ABX test in a scale where 1 is the source speaker identity and 5 the target speaker identity.

Experiment	Source	Neither	Target
1	12	32	56
2	2	26	72
3	4	20	76

Table 1. ABX test results

Results for the similarity test are displayed in Table 2. Although the similarity test was the most difficult task for the listeners, the similarity between source and converted speech files was less rated than the similarity between target and converted files for all three experiments. This result confirms that the VC systems proposed achieve the goal of moving away the individuality of voices from a source speaker to a target speaker. The only remark is about the experiment 2, Converted Residuals, where rates are low in both questions. This result can be explained with the results of the MOS test, because experiment 2 presents the lowest quality.

The results of the MOS test (scaled from 1 to 5) are showed in table 3. First of all, it can be stated that the conversion of the vocal tract only (Exp. 1) degrades the quality of the signal (natural speech is usually rated ≥ 4). When apart from vocal tract, residuals are also modified the quality drops again,

Experiment	Source-Converted	Target-Converted
1	1.48	3.35
2	1.38	2.93
3	1.48	3.48

Table 2. Similarity test results

but in different grades depending on the method used. When predicted residual signals fed the converted LPC filter MOS results moves from 2.92 to 2.22, but the reduction in quality is more severe when converted residual signal are used.

The lack of quality in converted signals comes from three different reason: errors in vocal tract mapping, mismatch between vocal tract and residuals, and frame concatenation errors. Errors due to vocal tract mapping are shared by the three systems tested and they are not the object of this paper. Frame concatenation errors are present in Exp. 2 and 3 only, as the residual signal used in Exp. 1 is the original source one. Although a harmonic model and phase modification has been used for signal reconstruction (see section 5) not all the artifacts has been avoided and future studies will be able to increase both prediction and conversion quality approaches. The key difference between converted and predicted residual systems is the mismatch between vocal tract and residuals. According to the results, the relationship between LSF parameters and their residuals results in a better morphing quality than the relationship between residuals of the source and target speakers.

	Exp. 1	Exp. 2	Exp. 3
ABX	3.62	3.90	3.92
MOS	2.92	1.74	2.22

Table 3. ABX and MOS test results

7. CONCLUSIONS

The main goal of this paper is to study the degree of relationship between LPC residual signals of one speaker with another speaker, and the degree of relationship between LPC residual signals of one speaker with their vocal tract parameters in order to compare two different strategies for residual transformation in voice morphing. As a secondary goal, a new method to store the residual signals and to reconstruct the transformed speech avoiding large noises due to phase discontinuities is presented. A CART vocal tract mapping system has been used for all the experiments.

Perceptual results have showed that voices obtained from predicted residual signals have better quality and are more effective in the speaker personality transformation task than voices obtained from converted residuals. It can be stated that the relationship between LSF parameters and their residuals is a better measure in residual mapping than the relationship between residuals of two different aligned speakers.

The harmonic model for residual signals alleviates the problem of memory load present in previous residual prediction methods [4] [5], because only amplitudes (and F0 if no pitch normalization is applied) must be stored. Also, it is straightforward to modified harmonic phases to assure continuity. Although phase continuity avoid large noises in the reconstructed signal the final quality can be improved by future studies.

8. REFERENCES

[1] A. Kain and M. W. Macon, "Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction," in *International Conference on Acoustics, Speech, and Signal Processing*, 2001.

[2] D.G. Childers, "Glottal source modeling for voice conversion," *Speech Communication*, vol. 16, 1995.

[3] K.S. Lee, D.H. Youn, and I.W. Cha, "A new voice transformation method based on both linear and nonlinear prediction analysis," in *International Conference on Spoken Language Processing*, 1996, pp. 1401–1404.

[4] Hui Ye and Steve Young, "High quality voice morphing," in *International Conference on Acoustics, Speech, and Signal Processing*, 2004.

[5] D. Sündermann, A. Bonafonte, H. Ney, and H. Höge, "A Study on Residual Prediction Techniques for Voice Conversion," in *International Conference on Acoustics, Speech, and Signal Processing*, 2005.

[6] H. Duxans, A. Bonafonte, A. Kain, and J. van Santen, "Including dynamic and phonetic information in voice conversion systems," in *International Conference on Spoken Language Processing*, 2004.

[7] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion vector quantization," in *International Conference on Acoustics, Speech, and Signal Processing*, 1988.

[8] R. Laroia, N. Phamdo, and N. Farvardin, "Robust efficient quantization of speech LSP parameters using structured vector quantizers," in *International Conference on Acoustics, Speech, and Signal Processing*, 1991, pp. 641–644.

[9] H. Depalle and T. Hélie, "Extraction of Spectral Peak Parameters using a Short-Time Fourier Transform Modeling and No Sidelobe Windows," in *ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, 1997.