VOICE CONVERSION WITH LINEAR PREDICTION RESIDUAL ESTIMATON

Winston S. Percybrooks[1,2], Elliot Moore II[1]

[1] Department of Electrical and Computer Engineering, Georgia Institute of Technology, Savannah, Georgia, USA

[2] Department of Electrical and Electronics Engineering, Universidad del Norte, Barranquilla, Colombia

ABSTRACT

The work presented here shows a comparison between a voice conversion system based on converting only the vocal tract representation of the *source* speaker and an augmented system that adds an algorithm for estimating the *target* excitation signal. The estimation algorithm uses a stochastic model for relating the excitation signal to the vocal tract features. The two systems were subjected to objective and subjective tests for assessing the effectiveness of the perceived identity conversion and the overall quality of the synthesized speech. Male-to-male and female-to-female conversion cases were tested. The main objective of this work is to improve the recognizability of the converted speech while maintaining a high synthesis quality.

Index Terms— Voice conversion, LP residual estimation, GMM, Linear spectral frequencies

1. INTRODUCTION

Voice conversion is the technique of transforming speech spoken by a source speaker to speech that is perceptually similar to a target speaker [1], [2]. There are several applications of voice conversion (VC) including movie dubbing for accurate portrayals of the original actors in foreign translations, restoration of old audio tapes, custom text-to-speech systems, and foreign language learning. Research in the area has been largely based on the classic source-filter model of speech production [1], [3], which views speech as the output of a time-variant filter (a model of the vocal tract) excited by a source signal (that represents the glottal excitation). Many research efforts have found the vocal tract filter to be more closely related to the perceptual identification of speakers [4], [5]. Accordingly, most previous work in the area is related to different ways of transforming the vocal tract representation from source to target speaker [1], [3], [5]. However, the excitation signal also has provided useful information for identifying specific voices [5], [6], [7]. Therefore, it is expected that for achieving high quality voice conversion adequate processing of both elements from the speech production model would be necessary.

This paper will show how a baseline VC system, which relies on transforming only the filter parameters from *source* to *target* speaker, can be improved by adding a new algorithm for estimating the corresponding excitation signal from the transformed vocal tract parameters [6]. During training, this estimation algorithm first classifies *target* excitations into several classes, finds the probability of transition between such classes in the training sequence and finally builds a relationship between the vocal tract parameters and each excitation class using gaussian mixture models (GMM). The three resulting objects of the training phase (a codebook composed of representative excitations from each class, the matrix of probabilities of transition between classes, and the vocal tract parameter's GMMs) are then used during estimation for creating the *target* excitations to use for synthesizing the converted speech.

The rest of the paper is organized in the following way: Section 2 shows some related previous work; Section 3 describes the baseline VC system used as reference; Section 4 presents the estimation method proposed for enhancing the baseline VC system; Section 5 shows comparative results of objective and subjective tests on the baseline and augmented VC system; Section 6 contains conclusions and planned extensions to this work.

2. PREVIOUS WORK

Currently, most VC systems use some type of linear prediction (LP) representation for modeling vocal tract filter parameters [7], [8], [9]. Line spectral frequencies (LSF) are the most common choice, mainly because of the ability to capture the resonances of the vocal tract using fewer coefficients and with better interpolation properties than alternatives like cepstral coefficients [8]. Several techniques have been proposed for converting the vocal tract representation, as segmental codebooks [10] and artificial neural networks [11], but the most common one is to use a linear transformation based on a stochastic model (e.g. GMMs) [3].

With respect to the excitation signal, earlier work was focused on transforming *source* excitations to *target* ones using techniques similar to that used for the vocal tract [10]. However, more recently the emphasis has shifted to trying to estimate the *target* excitation from its own vocal tract model features. Non-probabilistic estimation methods based in

codebooks have been proposed [7] as well as probabilistic ones based in GMMs [3]. Such estimation algorithms have been found to obtain better conversion results [7], [9]. In this work a different algorithm was used for estimating the excitation as described in Section 4.

3. BASELINE CONVERSION SYSTEM

For this work, a baseline VC system was built. It models the vocal tract filter by computing the LSF parameters for every pitch-synchronous input speech frame. The system works in two main modes, training and transformation, which operate as described below.

Training mode:

1) *LSF Extraction:* Speech samples with the same phonetic content from both *source* and *target* speaker are analyzed using LP for obtaining its corresponding LSF parameters.
2) *Feature alignment:* The LSF vectors obtained before are time-aligned using dynamic time warping (DTW) in order to compensate for any difference in duration between *source* and *target* utterances.
3) *Estimation of the transformation function:* The aligned LSF vectors are then used to train a joint GMM whose parameters then build a stochastic transformation function. Such a function basically computes the most likely *target* LSF vector given the input *source* LSF vector and the GMM as presented in [3].

Transformation mode:

1) *LSF Extraction:* As in training mode, LSF vectors are computed from the input speech, but in this case only the *source* speaker's utterances are used.
2) *LSF Transformation:* The GMM-based transformation function built during training is now used for converting every *source* LSF vector into its most likely *target* equivalent [9].
3) *Synthesis:* Transformed LSF vectors are used in conjunction with the *source* LP residual (obtained by inverse filtering) to synthesize the resulting converted speech.

As stated in the last step, this system uses the unchanged *source* LP residuals as excitation for the synthesis filter. As a result, the synthetic speech still has perceptible features from the *source* speaker [5].

4. RESIDUAL ESTIMATION

A residual estimation stage was added to the baseline VC system described in the previous section. It takes as input the converted LSF vectors and outputs an estimation of the corresponding *target* residuals. Now the synthesis filter is excited by the estimated *target* residuals instead of the *source* ones. The LSF transformation function is not modified. Fig. 1 shows a simplified diagram of the final augmented VC system.

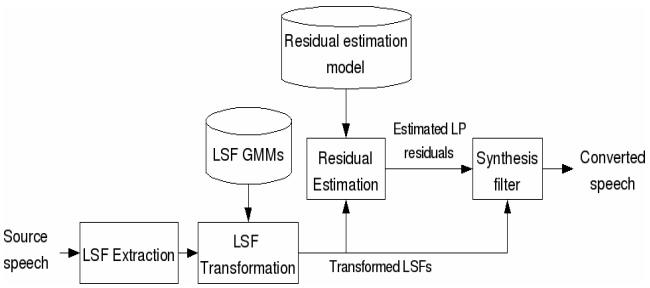


Fig.1. Simplified block diagram of the augmented VC system

For this residual estimation an algorithm first described in [6] was used. As illustrated in Fig. 2, the training procedure consists of four main steps:

1) *Feature extraction:* A pitch-synchronous LP analysis and inverse filtering is performed on training speech from the *target* speaker. The resulting features are paired vectors of LSF and LP residuals.
2) *Residual clustering:* Training LP residuals are then classified using K-means clustering with Euclidean distance, and the final centroid for every class is stored.
3) *Estimation of transition probabilities:* The training sequence of LP residuals is labeled using the classes from the K-means clustering. This sequence of labels is then used to compute the probability of transition between classes.
4) *GMM training:* The sequence of labels from the previous step is also used on the corresponding sequence of LSF vectors. Each resulting set of LSF's is then used for training a GMM, so at the end there will be a LSF's GMM for every LP residual class.

The resulting model can be seen as a hidden markov model (HMM) where the residual's classes are the states and the LSF vectors the observations. Then, the training procedure described above is equivalent to training a HMM with a known sequence of states (labeled sequence of LP residuals from clustering) and using GMMs as the probability distribution of observations within each state.

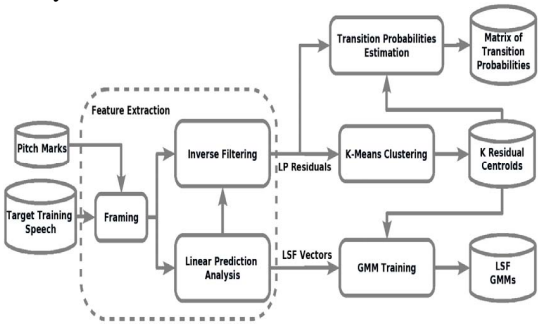


Fig.2. Outline of the training phase for the LP residual estimation algorithm.

The estimation phase is outlined in Fig. 3, where new LP residuals are computed as a linear combination of the residual centroids ($Cres_n$) obtained during training. The

weights ($w_{n,i}$) for the combination are recomputed every frame using an input LSF vector (l_i, from the LSF conversion stage) and probabilistic information stored in the GMMs ($p(l_i|\Theta_n)$) and the matrix of transition probabilities ($a_{n,i}$).

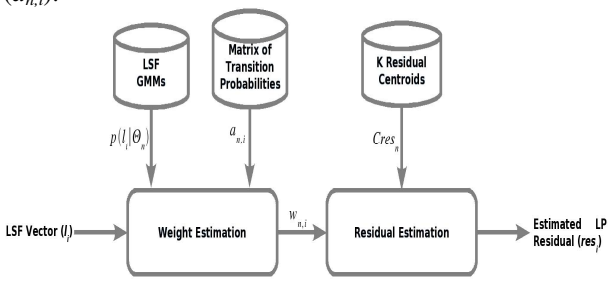


Fig.3. Outline of the estimation phase for the LP residual estimation algorithm.

5. RESULTS AND DISCUSSION

A comparative evaluation of the baseline VC system versus the augmented system was conducted to compare overall performance. Two different conversion scenarios were tested: male-to-male and female-to-female, using data from the VOICES [3] database. For each speaker, the available recordings were randomly divided in two non-overlapping sets: a training corpus with 35 sentences, and a testing corpus with 15 sentences. In all cases, two pitch period long frames with overlapping were used for pitch-synchronous speech analysis and synthesis. As the LP residual for unvoiced speech sections has been found to be of little relevance for perceptual speaker identification [3], [5], the augmented system only used LP residual estimation for voiced frames. For unvoiced sections the *source* residual was used.

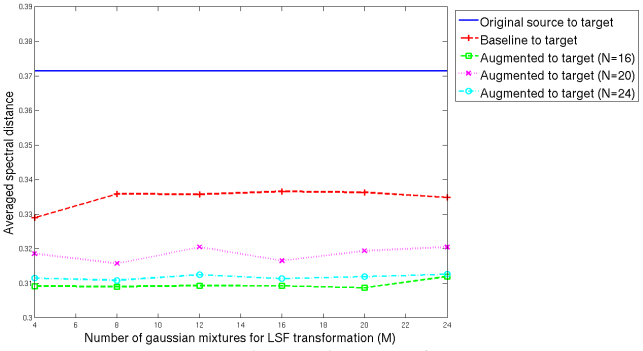
First, an objective comparison was done using a spectral distance measure defined as [7]

$$SD=\frac{1}{L}\sum_{p=0}^{L-1}\left[\frac{1}{D}\sum_{k=0}^{D-1}\left(\left|S_{org}(p,w_k)\right|-\left|S_{con}(p,w_k)\right|\right)^2\right]^{\frac{1}{2}} \quad (1)$$

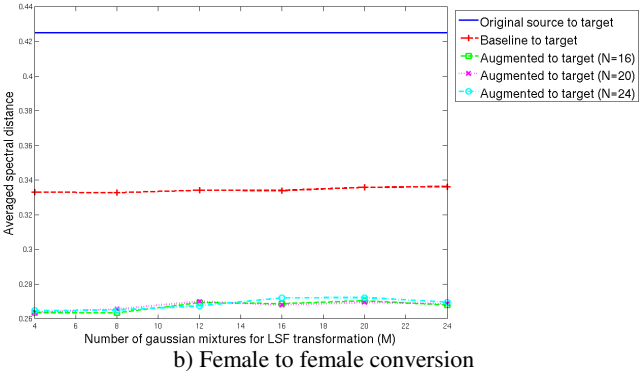
Where $S_{org}(p,w_k)$ and $S_{con}(p,w_k)$ are respectively the short time Fourier transform of the p-th original and converted voiced frames; D is the number of DFT points (1024 in this case) and L is the number of frames. Fig. 4 summarizes the results obtained for this objective measure. Different numbers of residual's clusters (i.e. number of states in the HMM analogy. N = 16, 20, 24), as well as several numbers of gaussian mixtures for the LSF conversion (i.e. number of mixtures per state. M = 4, 8, 12, 16, 20, 24) were tested. The original *source* to *target* spectral distance is also included as a reference.

The augmented system always gave smaller distances than the baseline system. Moreover, according to Fig. 4 using a higher number of residual's clusters did not necessarily imply a reduction on spectral distance. More testing will be needed for determining the optimal number of clusters to use, so the augmented systems performs consistently well with respect to spectral distance in different conversion scenarios.

Objective measures were complemented with two different subjective tests. Eight untrained, normal hearing listeners participated in listening tests. For generating the converted speech each VC system was configured with the number of gaussian mixtures (M) and residual's clusters (N) that gave the smaller spectral distance on the objective measures. The first listening test was a mean opinion score (MOS) for contrasting the perceived quality of synthesized speech from both VC systems. Eight sentences from the testing set were selected, and the listeners were presented with three different versions of each one in random order: one original *target* recording, one converted by the baseline system and one converted by the augmented system. Listeners were asked to assess the quality of every sentence using a numeric scale from 1 (very poor) to 5 (excellent). All evaluations were averaged together and Table 1 presents the consolidated results.



a) Male to male conversion



b) Female to female conversion

Fig.4. Comparison of spectral distance results for different configurations of the baseline and augmented VC systems.

The second subjective test was an extended ABX designed to judge if the converted speech was perceived as the *target* voice or not. Listeners again were presented with three different versions of 8 testing sentences: original *source* and *target* recordings (A and B in random order), and converted speech (X). Then, they were asked to decide to which voice (A or B) X was closer to according to the

following numeric scale (for the purposes of the numbers presented here, A is the *source* speaker and B the *target*, during the actual test this order was randomly changed but the software kept track of the assignment so the final numeric result was consistent with this assumption):

1. X sounds like A for sure
2. X sounds closer to A but I'm not quite sure is the same person
3. X sounds like neither A nor B
4. X sounds closer to B but I'm not quite sure is the same person
5. X sounds like B for sure

In other words, scores closer to 5 indicated better quality of conversion. Baseline and augmented systems were tested independently for both male-to-male and female-to-female conversions, the final averaged results are shown in Table 2.

Table 1. MOS results.

Speech source	Averaged MOS	Confidence interval (99%)
Original *target* recordings	4.84	[4.68, 5.00]
Baseline VC system	3.97	[3.54, 4.40]
Augmented VC system	3.72	[3.39, 4.05]

Table 2. ABX test results.

Type of conversion	VC System tested	Avg. ABX result	Confidence interval (99%)
Male-to-male	Baseline	3.89	[3.57, 4.21]
	Augmented	4.34	[3.94, 4.74]
Female-to-female	Baseline	3.48	[3.02, 3.94]
	Augmented	4.05	[3.69, 4.41]

By contrasting MOS and ABX results, the inclusion of the residual estimation stage in the augmented VC system resulted in converted speech that is significantly closer perceptually to the *target* voice than the speech synthesized from the baseline system; but at the same time it introduced a higher distortion that lowered the MOS results. We believe the main reasons for this behavior are:

- The residual estimation strategy contributed useful information for the perceptual identification of the speakers.
- The signal representation used for the residuals resulted in phase discontinuities in the synthesized speech (while not present in the baseline system because of the use of natural residuals) that are perceived as a larger amount of artifacts during the MOS tests.

6. CONCLUSION

This paper presented how a VC system based only on the conversion of vocal tract features (LSF vectors) can be enhanced by adding a new algorithm that estimates the excitation signal of the *target* speaker. The estimated excitation signal was obtained through an HMM-like training model, involving *target* LP residual clustering, GMMs and transition probabilities. Objective tests measuring spectral distortion, and subjective tests involving perceived synthesis quality and identity conversion, were conducted on same gender conversion scenarios. The augmented system showed a significantly better performance than the baseline system in making the converted speech to sound closer to the *target* speaker. However, at the same time the residual estimation strategy slightly decreased the quality of converted speech. Future directions of this work will be focused on alternative ways of representing the LP residual to enhance the quality of the synthetic speech.

7. REFERENCES

[1] D. Sunderman, "Voice Conversion: State-of-the-Art and Future Work", in *Proc. DAGA*, 2005.

[2] O. Turk, L. Arslan, "Subband Based Voice Conversion", in *Proc. ICSLP*, 2002.

[3] A. Kain, "High Resolution Voice Transformation", *Ph.D. Dissertation*, OGI School of Science and Engineering, Oregon Health and Science University, 2001.

[4] M.D. Plumpe, T.F. Quatieri, D.A. Reynolds, "Modeling Glottal Flow Derivative Waveform with Application to Speaker ID", in *IEEE Trans. Speech and Audio Processing*, vol.7, pp. 569-586, 1999.

[5] K. Itoh, "Perceptual Analysis of Speaker Identity", in *Speech Science and Technology*, S. Saito, Ed. IOS press, pp. 133-145, 1992.

[6] W. Percybrooks, E. Moore II, "New Algorithm for LPC Residual Estimation from LSF Vectors for a VC system", in *Proc. INTERSPEECH*, pp. 1977-1980, 2007.

[7] J. Sun, B. Dai, J. Zhang, Y. Xie, "Modeling Glottal Source for High Quality Voice Conversion", in *Proc. 6th World Congress on Intelligent Control and Automation*, pp. 9459-9462, 2006.

[8] H. Ye, S. Young, "High Quality Voice Morphing", in *Proc. ICASSP*, pp. 9-12, 2004.

[9] A. Kain, M. Macon, "Spectral Voice Conversion for Text-to-Speech Synthesis", in *Proc. ICASSP*, pp. 285-288, 1998.

[10] L. Arslan, "Speaker Transformation Algorithm Using Segmental Codebooks (STASC)", in *Speech Communication*, vol. 28, pp. 211-226, 1999.

[11] M. Narendranath, H. Murthy, S. Rajendran, B. Yegnanarayana, "Transformation of formants for voice conversion using artificial neural networks", in Speech Communication, vol. 16, pp. 207-216, 1995.