



New Algorithm for LPC Residual Estimation from LSF Vectors for a Voice Conversion System

Winston S. Percybrooks^{1,2}, Elliot Moore II¹

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

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

elliott.moore@gtsav.gatech.edu, wpercybrooks@gatech.edu

Abstract

Voice conversion involves transforming segments of speech from a *source* speaker to make them to be perceived as if spoken by a *target* speaker. Generally, this process involves the estimation of vocal tract parameters and an excitation signal that match the *target* speaker. The work presented here proposes an algorithm for estimating the excitation residuals of the *target* speaker using a weighted combination of clustered residuals. The algorithm is subjected to objective and subjective comparisons to other basic types of residual estimation techniques for voice conversion. Tests were carried on 2 male and 2 female *target* speakers in an ideal setting. The overall goal of this work is to create an improved algorithm for estimating excitation residuals during voice conversion that maintain speaker recognizability and high synthesis quality.

Index Terms: LPC residual, LSF parameters, GMM, voice conversion

1. Introduction

One objective of voice conversion systems is to transform segments of speech spoken by a *source* speaker into segments that are similar perceptually to a *target* speaker. Traditionally, voice conversion has been based on the source-filter model that synthesizes speech utilizing a source model (i.e., a model of the excitation or glottal waveform) and filter coefficients to represent the vocal tract [1], [2]. The vocal tract configuration has been closely linked to the perceptual identification of a particular speaker's voice [3], [4], [5]. However, the excitation model has also provided useful cues in identifying specific voices [6], [7], [8]. The inherent interaction involved in natural speech makes it necessary to modify the configuration of parameters for the vocal tract and excitation signal in order to achieve high quality voice conversion.

This paper will focus on a new approach for creating the excitation signal for a target speaker in a voice conversion system that maintains speaker recognizability and high synthesis quality. The algorithm presented here involves the following primary components: 1) Estimation and clustering of linear prediction residuals from a target speaker; 2) Training of transition probabilities between the residual clusters; and 3) Use of Gaussian mixture models (GMM) for characterizing the dependencies between the residual clusters and vocal tract parameters (i.e., Linear Spectral Frequencies).

The rest of the paper is organized in the following way: Section 2 shows relevant alternative estimation techniques that have been used in previous work; Section 3 describes the proposed method; Section 4 presents the comparative results of objective and subjective tests; Section 5 contains conclusions and planned extensions to this work.

2. Previous work

Initial attempts to estimate the excitation signal for voice conversion were focused on building transformation functions between the *source* and *target* excitations. For example, in [9] the *source* excitation is transformed to match the *target* excitation by using weighted combinations of codeword filters derived from the average *source* and *target* excitation spectra. However, speech synthesis for voice conversion systems is inherently speaker dependent. It has been shown [1] that there is some correlation between the vocal tract filter and excitation signal that can be exploited for estimating the *target* excitation from its vocal tract parameters without the need to transform the *source* excitation. Algorithms that have generated excitation signals based on a relationship to the *target* vocal tract parameters have generally produced better quality speech than those based on transforming the excitation of the *source* [1], [5], [8]. Therefore, this paper will focus on techniques that utilize vocal tract information in the estimation of the excitation signal.

Current literature focuses on probabilistic and non-probabilistic methods for estimating the excitation signal from vocal tract information in voice conversion systems. Both methods have a common analysis phase that extracts vocal tract parameters in the form of linear prediction coefficients (LPC) or linear spectral frequencies (LSF) from training samples of a *target* speaker. Additionally, both methods generally operate only on voiced frames. The main difference in these algorithms lies in their treatment of the excitation residual obtained through inverse filtering. The primary components of these two methods are outlined in the following subsections.

2.1. Non-probabilistic estimation

Non-probabilistic estimation, as the one described in [2], follows these general steps:

- LPC/LSF training vectors are clustered into K different classes.

- A single representative residual vector for each cluster is then selected. For example in [2] the residuals corresponding to the LSF vectors nearest to each cluster's centroid are chosen.
- A codebook is built containing the cluster's centroids and corresponding representative residuals.
- In the estimation phase, when a new LSF vector is produced, its distance to every centroid in the codebook is computed and the residual corresponding to the closest one is used as excitation.

We will refer to this method as the Non-probabilistic LPC (NP-LPC) algorithm in the rest of this paper.

2.2. Probabilistic estimation

A typical example of probabilistic estimation is the one presented in [1] involving the following steps:

- The training LSF vectors are used to fit a Gaussian Mixture Model (GMM) with K mixtures.
- Training LSF vectors are then soft-classified between the mixtures, assigning degrees of class-membership to them according to the pdf of the GMM.
- Those class-membership values are then used as weights for combining the corresponding residual vectors to form a representative residual for each mixture, forming a residual codebook.
- In the estimation phase, when a new LSF vector is produced, the GMM is used to generate its corresponding class-membership values for each mixture. Such values are then used as weights for combining the residuals in the codebook, forming the estimated residual to be used in synthesis.

We will refer to this method as the GMM based - LPC clustering (GMM-LPC) algorithm in the rest of this paper.

3. Proposed approach

The algorithm being presented here for estimating the excitation signal is similar to the approach in [1] in that the excitation is based on the residual signals obtained from the LPC analysis of a training speech corpus from a single (target) speaker and it is used only for voiced frames. However, there are three main differences:

- The initial clustering is performed on the LPC residuals not on the LSF vectors.
- Several GMMs are constructed, one per residual cluster, instead of a single one.
- Information about the temporal transitions between clusters are now taken into account.

The estimation procedure presented here is based on the following set of hypotheses:

- For a single speaker it is possible to classify the LPC residuals in a finite number of sets, and each one can be represented by a different 'characteristic' residual.
- For a single speaker it is also possible to find a relationship between the occurrence of those sets of residuals and the occurrence of different sets of LPC/LSF parameters.

- As, in general, the status of the phonatory system doesn't change abruptly from one speech frame to the following, it should be possible to find a correlation between the residual signals in two successive frames.

3.1. Training phase

For training, two sequences of features are extracted from the voiced segments of speech in a *target* training corpus. The first sequence, \mathbf{L} , is composed of the LSF vectors obtained from LPC analysis, while the second one, \mathbf{R} , is composed of the corresponding LPC residuals from inverse filtering the original speech frames. Once the two training sequences are obtained the algorithm proceeds as:

- 1) K-means clustering is run on \mathbf{R} to obtain K clusters of residual vectors, each one represented by its centroid ($Cres_i \equiv$ residual centroid of cluster i). Those centroids are the 'characteristic' residuals mentioned in the first hypothesis.
- 2) Now that the residual vectors have been effectively divided among K sets, the corresponding LSF vectors in \mathbf{L} are used to build a Gaussian Mixture Model (GMM) for each cluster. Those K GMMs are used to describe the relationship between residual and LSF vectors in the second hypothesis.
- 3) Finally, the training sequence \mathbf{R} is used to estimate $P(r_i \in Cres_a | r_{i-1} \in Cres_b)$, i.e. the conditional probability of residual r_i belonging to cluster $Cres_a$ given that the previous residual, r_{i-1} , belongs to cluster $Cres_b$. These transition probabilities are trained simply by counting the number of times that a residual in cluster $Cres_b$ is followed by one in cluster $Cres_a$ in the training sequence \mathbf{R} , then dividing it by the total number of transitions from residuals in cluster $Cres_b$.

Summarizing, at the end of the training phase the following modeling elements are available: A codebook of K residual vectors; for each entry in these codebook, an associated GMM fitted to the probability distribution of LSFs for that residual cluster; a matrix containing the transition probabilities between two successive residual clusters.

3.2. Estimation phase

The estimation phase requires obtaining the *target* LSF vector and excitation signal to be used in the voice conversion. The *target* LSF vector (l_i) is obtained by transforming a *source* LSF vector using the GMM based transformation function proposed in [1]. Once the *target* LSF vector has been obtained, the corresponding residual vector (res_i) is estimated as: For each *target* LSF vector l_i , which in a voice conversion system would be obtained by an independent transformation function from the *source* ones [1], [2], [8], the corresponding residual vector res_i is estimated as:

$$res_i = \sum_{n=1}^k w_{n,i} Cres_n \quad (1)$$

where $Cres_n$ is the n-th residual in the codebook and $w_{n,i}$ is a variable weight computed as:

$$w_{n,i} = \frac{a_{n,i} p(l_i | \Theta_n)}{\sum_{m=1}^k a_{m,i} p(l_i | \Theta_m)} \quad (2)$$

where $a_{n,i}$ is the probability for res_i belonging to cluster n given the cluster to which res_{i-1} belongs, while $p(l_i | \Theta_n)$ is the value of the pdf of the n-th GMM evaluated at l_i . Thus the

estimated residual is a linear combination of the residual centroids with weights determined by the GMMs and the transition matrix.

4. Results and discussion

The algorithm presented here was evaluated in comparison to the alternative approaches described in section 2 (i.e., non-probabilistic LPC clustering (NP-LPC) and GMM-LPC clustering). An ideal voice conversion scenario was created in which the *source* and *target* speaker were identical. In this way, the inherent speaker dependencies of voice conversion could be slightly mitigated and each algorithm could be evaluated on their ability to generate high perceptual quality in the synthesized speech. Only voiced segments of speech were considered for evaluating the residual estimation. The synthesis was conducted using a pitch-synchronous LPC analysis and overlap-add method.

The tests were performed using four different speakers, two males and two females, from the VOICES [1] database. The training phase used 35 sentences from each speaker, each one spoken 3 times; while for testing the remaining 15 sentences, each one also spoken 3 times, from each speaker were used.

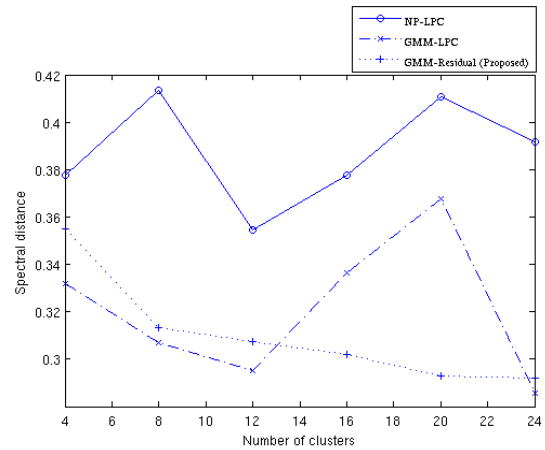
Objective and subjective evaluations were performed on the synthesized sentences. Objective comparisons of the three methods were conducted using a spectral distance measure defined as

$$SD = \frac{1}{M} \sum_{p=0}^{M-1} \left[\frac{1}{N} \sum_{k=0}^{N-1} \left[|S_{org}(p, w_k)| - |S_{con}(p, w_k)| \right]^2 \right]^{\frac{1}{2}} \quad (3)$$

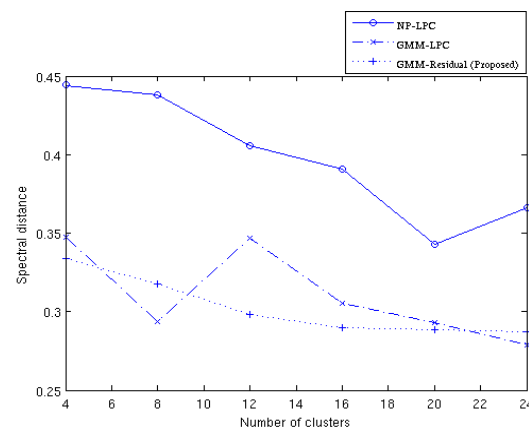
where $S_{org}(p, w_k)$ and $S_{con}(p, w_k)$ are the short time Fourier transform of the p-th original and converted voiced frames respectively; N is the number of DFT points (1024 in this paper) and M is the number of frames. This metric was used to determine the spectral distance between the converted samples of speech and the samples for the original speaker for all three algorithms. Each algorithm was run based on cluster sizes of $K = \{4, 8, 12, 16, 20, 24\}$. Figure 1 shows an example of the spectral distance measures across the various cluster sizes for a male and female speaker. For all speakers, the highest spectral distance overall was obtained for the NP-LPC based algorithm indicating that such approach was introducing the highest levels of distortion.

Specific comparisons for a particular cluster size between the GMM-LPC and the proposed GMM-Residual algorithms were dependent on the speaker. Table 1 shows the minimum spectral distance achieved by each algorithm for each speaker with the corresponding cluster size (K). The NP-LPC algorithm resulted in higher spectral distance measures for all of the speakers regardless of cluster size when compared to the GMM-LPC and the proposed GMM-Residual method. Additionally, the proposed GMM-Residual method produced the lowest overall spectral distances between the converted and original speech. The cluster sizes that resulted in the lowest spectral distance varied for each speaker which supports the understanding that voice conversion is largely speaker dependent. As a result, the cluster sizes must be trained and adjusted to each *target* speaker.

Subjective comparisons of the algorithm involved evaluating the overall perceived quality of the resulting synthesis. A Mean Opinion Score (MOS) test was conducted



a): Male Speaker.



b): Female Speaker.

Figure 1: Spectral distance vs cluster size for two different speakers.

Table 1. Minimum spectral distance for each speaker.

Algorithm	Speaker	Min. Spectral distance	K
NP-LPC	Male 1	0.3546	12
	Male 2	0.2824	4
	Female 1	0.3428	20
	Female 2	0.4939	4
GMM-LPC	Male 1	0.2919	24
	Male 2	0.2169	16
	Female 1	0.2875	24
	Female 2	0.4054	20
GMM-Residual (Proposed)	Male 1	0.2858	24
	Male 2	0.1842	16
	Female 1	0.2787	20
	Female 2	0.3839	8

on the synthesized speech generated from all three algorithms. A total of ten, untrained, normal hearing listeners were asked to evaluate the perceived speech quality of a series of synthesized sentences. For every *target* speaker a pool of 15 sentences was available. Three versions of each sentence

were generated using one of the three algorithms under consideration with the cluster sizes based on the minimal spectral distances determined for each speaker as seen in Table 1. Each listener was presented with all three versions of 8 randomly selected sentences from the testing pool. They were then asked to evaluate the perceived quality of the three versions of the synthesized sentence using the rankings in Table 2. It should be noted that the different versions played in random order, so the listeners could not know to which version they were listening.

Table 2. Ranking for the MOS test.

Rating	Speech quality	Level of distortion
1	Unsatisfactory	Very annoying and objectionable
2	Poor	Annoying but not objectionable
3	Fair	Perceptible and slightly annoying
4	Good	Just perceptible but not annoying
5	Excellent	Imperceptible

Averaged results for the MOS test are summarized in Table 3. The NP-LPC version of voice conversion produced the lowest MOS score when compared to the other algorithms under consideration and a very poor score overall. The GMM-LPC and GMM-Residual algorithms produced speech that was generally judged as being of "Fair" quality. However, the proposed GMM-Residual algorithm generated a MOS score that was close to a "Good" MOS rating. Some artifacts always appeared on the synthesized speech, indicating that additional work should be carried on upgrading the proposed procedure. Nonetheless, the fact that the proposed algorithm achieved the best result in both objective and subjective evaluation indicates a measurable improvement over previous algorithms. We believe this improvement is related to a better characterization of the *target's* residual space due to two main reasons:

- Clustering the residuals directly instead of through the corresponding LPC/LSF vectors would be producing a more natural and meaningful set of clusters.
- Incorporating the temporal information described by the matrix of cluster's transition probabilities would be producing a sequence of estimated residuals which more accurately resembles the general evolution of naturally occurring residuals for a given *target* speaker.

Table 3. Results of MOS test.

Algorithm	Avg. MOS result
LPC Clustering Non-probabilistic	2.20
LPC Clustering GMM based	3.20
Residual Clustering GMM based (Proposed)	3.86

5. Conclusions and future work

This paper presented a new algorithm for estimating the excitation residual of the *target* speaker in a voice conversion system. The excitation signal was obtained through a combination of LPC residual clustering, GMMs and transition probabilities. The proposed algorithm was compared to two basic types of residual estimation for voice conversion that involved a non-probabilistic approach based on clustering of LSF vectors for building a codebook of residuals, and a probabilistic approach that used a GMM for relating LSF vectors to LPC residuals. Objective tests measuring spectral distortion and subjective tests involving perceived synthesis quality were conducted on 2 male and 2 female speakers. In all cases, the proposed algorithm showed an improvement over the existing algorithms under consideration. Future directions of this work will involve evaluations involving different *source* and *target* speakers for a full voice conversion system.

6. References

- [1] Kain, A., "High Resolution Voice Transformation", PhD dissertation, OGI School of Science and Engineering, Oregon Health and Science University, 2001.
- [2] Sun, J., Dai, B., Zhang, J., and Xie, Y., "Modeling Glottal Source for High Quality Voice Conversion", Proceedings of the 6th World Congress on Intelligent Control and Automation, pp. 9459-9462, 2006.
- [3] Kain, A., and Macon, M., "Spectral Voice Conversion for Text-to-Speech Synthesis", Proceedings of ICASSP, Vol. 1, pp. 285-288, 1998.
- [4] Plumpe, M.D., Quatieri, T.F., and Reynolds, D.A., "Modeling of the Glottal Flow Derivative Waveform with Application to Speaker Identification", IEEE Trans. Speech and Audio Processing, Vol. 7, no. 5, pp. 569-586, 1999.
- [5] Kain, A., and Macon, M., "Design and Evaluation of a Voice Conversion Algorithm Based on Spectral Envelope Mapping and Residual Prediction", Proceedings of ICASSP, Vol. 2, pp. 813-816, 2001.
- [6] Itoh, K., "Perceptual Analysis of Speaker Identity", in Speech Science and Technology, S. Saito, Ed. IOS press, ch. 2.6, pp. 133-145, 1992.
- [7] Childers, D.G., "Glottal Source Modeling for Voice Conversion", Speech Communication, Vol. 16, pp. 127-138, 1995.
- [8] Ye, H., and Young, S., "High Quality Voice Morphing", Proceedings of ICASSP, Vol. 1, pp. 9-12, 2004.
- [9] Arslan, L., "Speaker Transformation Algorithm using Segmental Codebooks (STASC)", Speech Communication, Vol. 28, no. 3, pp. 211-226, 1999.