A HMM approach to residual estimation for high resolution voice conversion

Winston Percybrooks¹², Elliot Moore¹

¹School of Electrical and Computer Engineering, Georgia Institute of Technology, Savannah, GA, USA
²Department of Electrical and Electronics Engineering, Universidad del Norte, Barranquilla, Colombia

wpercybrooks@gatech.edu, em80@mail.gatech.edu

Abstract

Voice conversion systems aim to process speech from a source speaker so it would be perceived as spoken by a target speaker. This paper presents a procedure to improve high resolution voice conversion by modifying the algorithm used for residual estimation. The proposed residual estimation algorithm exploits the temporal dependencies between residuals in consecutive speech frames using a hidden Markov model. A previous residual estimation technique based on Gaussian mixtures is used as comparison. Both algorithms are subjected to tests to measure perceived identity conversion and converted speech quality. It was found that the proposed algorithm generates converted speech with significantly better quality without degraded identity conversion performance with respect to the baseline, working particularly well for female target speakers and cross-gender conversions.

Index Terms: Voice conversion, residual estimation, HMM, MOS test, ABX test

1. Introduction

Voice conversion (VC) systems aim to transform segments of speech from a given source speaker so that it can be identified as spoken by a different target speaker. Short-time speech features have been determined to be very important on speaker identification by humans [1]. In [2] they found short-time vocal tract representations to be closely related to the perception of speaker identity. In particular, there is evidence suggesting that line spectral frequencies (LSFs) produce higher-quality converted speech [3, 4], so they are frequently used in the literature (e.g. [5, 6]). Results in [3, 4] showed that in addition to the short-time vocal tract representation the source signal (e.g. linear prediction residual) also contains valuable information, not only for speaker identification but also for speech naturalness. Systems that include processing of the residual signal have been called ‘high resolution VC’ [3] because they include greater spectral details on the converted speech than vocal tract-only VC systems. Additional work (e.g. [7, 8]) found that, although high resolution VC algorithms showed improvement in identity conversion over systems that do not transform residual information, there is still a need for better converted speech quality, as evidenced by subjective tests showing lower performance in this area. Recent work has attempted to improve on [3, 4] by exploring new ways to map source to target vocal tract features. In [9] the Gaussian mixture model (GMM) mapping from [3, 4] is modified using global variance constrained minimization to enhance the global variance of the spectral features. In [10] the GMM mapping is completely replaced by a vocal tract mapping based on trellis structured vector quantization. In [11] the GMM mapping is also replaced, this time by a kernel transformation that uses partial least squares regression to find a conversion function between source and target vocal tract features. Although all these recent works show better speech quality with similar or better identity conversion compared to [3, 4], there is still a significant degradation in quality with respect to unconverted speech [9, 10, 11] giving further indication that there is still work to do in this area.

While [9, 10, 11] focused on modifying or replacing the vocal tract mapping used in [3, 4], the work presented here focuses instead on changing the residual estimation part by exploiting temporal relationships between residuals in consecutive frames in addition to the relationship between residual and vocal tract features. This is done by using a Hidden Markov Model (HMM). This article compares the performance of the proposed (HMM) residual estimation technique with the GMM-based algorithm for residual estimation shown in [3, 4]. The proposed algorithm was developed by the authors and is a modification of their earlier work first described in [12]. The same training and testing protocols are applied to both, GMM-based and HMM-based systems, so any difference on the evaluated VC performance will be due to the residual estimation techniques and not to any other part of the VC system. As evaluation metrics, the results from extended ABX and Mean Opinion Score (MOS) tests are used to measure identity conversion effectiveness and speech quality, respectively. The objective of the proposed HMM-based residual estimation procedure is to generate converted speech with higher quality than the baseline GMM-based system with at least the same level of identity conversion performance.
2. Vocal tract feature mapping

For the work presented here, the GMM transformation method detailed in [3] is used to map the LSF features of the source speaker into the target speaker. First, LSF vectors are extracted pitch synchronously from target and source speakers using frames two pitch periods long with a 50% overlap. The resulting LSF sequences are then time aligned using DTW, obtaining two aligned vector sequences \(x = [x_1, x_2, \ldots, x_n]\) and \(y = [y_1, y_2, \ldots, y_n]\) for source and target speakers respectively. Next, the joint pdf \(p(z) = p(x, y)\), where \(z = [x^T, y^T]^T\), is estimated through a GMM. Estimation of the GMM elements is done from the training data using the expectation-maximization (EM) algorithm, which iteratively recomputes the parameters until a convergence criterion is reached. Once the GMM for \(p(z)\) has been obtained, mean-square optimization shows that for every source LSF vector \(x_k\) a corresponding target LSF vector \(\hat{y}_k\) can be optimally estimated as the expected value of \(y\) given \(x_k\), i.e. \(E(y|x_k)\) [3].

3. Residual estimation

3.1. GMM-based algorithm

The baseline system compared in this work uses an established algorithm for residual estimation that is described extensively in [3]. The algorithm models the probability density function of the target speaker LSF vectors with a GMM, and then uses that model to construct a codebook to quantize the residual space of the target speaker. First, LSF vectors and the corresponding residual spectral magnitude and phase are extracted from every voiced frame in the target speaker training data. Next, a GMM with \(Q\) mixtures is trained on the extracted target LSF vectors using the EM algorithm. A residual magnitude codebook entry is calculated for each GMM component as a weighted average of all training magnitude residuals, with the weights corresponding to the normalized probability for each residual magnitude that belongs to that particular Gaussian component, thus resulting in a residual codebook with \(Q\) total magnitude entries. The corresponding residual phase entries are set to the training residual phases with the highest normalized probability for each particular Gaussian component. Based on the results presented in [3], \(Q = 64\) was selected for the subjective tests described in Section 4.

During voice conversion, LSF vectors from the source speaker are transformed into target LSF using the procedure in Section 2. The resulting transformed LSF vectors are used as input for the residual estimation algorithm. For each transformed LSF vector a residual magnitude is estimated as a weighted average of the magnitude codebook entries, using as weights the posterior likelihood of that particular LSF for each GMM component. The respective residual phase is estimated as the codebook entry corresponding to the GMM component with maximum posterior likelihood for that LSF vector.

3.2. HMM-based algorithm

While the results reported in [3] showed an improvement in identity conversion with respect to earlier VC systems, the quality of the converted speech was low according to the MOS scale, especially for conversions between speakers of different gender (i.e., cross-gender conversion). Such findings leave space for searching alternative methods to perform the residual estimation. In particular, the algorithm in [3, 4] assumes that residuals in consecutive frames are independent from each other. However, since the human phonation system cannot change abruptly in the time span between two short-term analysis frames it is reasonable to assume that there exists a dependency between speech features from consecutive frames. That temporal relationship is not exploited by the algorithm in [3, 4]. Since a HMM is a well-known probabilistic strategy to exploit temporal relationships between features, a method for performing residual estimation using HMM was considered as a viable candidate to improve on the results of [3, 4]. The proposed algorithm uses LSF vectors from the target speaker as HMM observations, \(O_i\), while the hidden states, \(S_i\), represent residual classes and the transition probabilities, \(a_{ij}\), represent the probability of switching from one residual class to another in consecutive speech frames. For each class, the probability density function of the observations, \(b_i(O_k)\), is modeled using a GMM. The resulting HMM is illustrated in Figure 1 for a simple 3-state model. The training procedure begins by extracting a sequence of LSF vectors and its corresponding sequence of residual vectors from the target training database. Contrary to what was done in [12], where residuals were computed in the time-domain using inverse filtering, here residual magnitude and phase vectors are obtained by subtracting the LSF cepstral magnitude and phase from the original speech frames. Then, instead of using K-means to classify the residuals as in [12], the LSF sequence is used directly to train a M-state HMM using Baum-Welch formulas, with the LSF probability density function \(b_i\) associated to each state being a N-mixture GMM. Next the LSF vectors and their associated residuals are classified into HMM states using the Viterbi algorithm. Each state is then represented by the residual magnitude and phase corresponding to the LSF with the highest probability of belonging to that state.

During voice conversion, LSF vectors from the source speaker are transformed into target LSF using the procedure in Section 2. The resulting transformed LSF are then used as input for the HMM residual estimation algorithm. Instead of using weighted averages to estimate the residuals as in [3, 4, 12], here the Viterbi algorithm is used to compute the most likely sequence of HMM states given the model previously trained and
the incoming sequence of transformed LSF vectors. The sequence of estimated magnitude and phase residuals is then obtained by representing each state in the resulting Viterbi sequence by the corresponding magnitude and phase residual vectors chosen during training, as illustrated in Figure 1. It should be noticed that because of the greater number of parameters needed to train the HMM (including training M different GMM, moreover considering the fact that the proposed HMM is ergodic) the proposed algorithm is more computationally complex than the GMM-based approach. The greater number of model parameters to estimate could also lead to the requirement of more training data than the GMM-based algorithm. Based on preliminary tests, in the work presented here the HMM was trained with M = 8 states and N = 16 mixtures per GMM.

4. Tests and discussions

Three tests were performed in total using speech from the VOICES database [3], which is an English language database designed for VC applications. First, an ABX test (Test 1) was done with 20 English-speaking listeners to measure how well each system performed identity conversion. Later, the ABX test was repeated (Test 2) using 20 Non-English-speaking listeners (i.e., this second group of listeners neither spoke or understood English). In this way the listeners were unaffected by the content of the recordings and could only focus on the speaker identity. During each test listeners had to rate 10 examples of conversions for each conversion case (i.e., Male-to-Male (M-M), Female-to-Female (F-F), Male-to-Female (M-F), Female-to-Male (F-M)) and residual estimation system (i.e., GMM-based or HMM-based) based on the ABX scale (1, no conversion, to 5, perfect conversion into target). Table 1 summarizes the average results for both systems. Statistical analysis shows that, except for the F-M case, the proposed system gets statistically significant higher ABX scores than the GMM system. Given that both systems are using the same vocal tract mapping procedure on the same training and test data, this result suggests that the proposed residual estimation algorithm tends to capture and estimate better the speaker identity related information contained in the target residual than the GMM algorithm. The proposed HMM system obtained particularly better ABX scores than the GMM system for the cases where the target is a female, suggesting that temporal relationship between residuals in consecutive frames is particularly important to synthesize female voices.

Gains in identity conversion are usually accompanied by degradation in converted speech quality. To check if this was the case with the proposed algorithm, a MOS test was carried out to evaluate converted speech quality from both systems using the 20 English-speaking listeners. In this case a MOS test with the Non-English-speaking listeners was not performed, since we considered the ability to understand what was being said to be an important part of accurately evaluating speech quality. Table 2 summarizes the average results for each system and each type of conversion. For this test, listeners were presented 10 sentences for each conversion case and each residual estimation algorithm plus 10 original, unconverted target speaker sentences. As a reference, the average MOS score for the original recordings without conversion was found to be 4.84 (upper level of the "Good" range on the MOS scale which goes from 1.0 to 5.0), confirming that the database contains recordings with good quality. MOS scores indicate that both systems generate converted speech with perceivable lower quality than the original sentences. In particular, listeners usually referred to a muffled quality on the converted speech and some distortion noise. The muffled sound suggests that although both systems are performing high resolution VC, there is still some level of over-smoothness of the spectral details contained in the original recordings. On the other hand, sudden changes in both vocal tract spectra and excitation signal at the frame boundaries may explain other distortion.

Table 2 shows that the GMM-based system performs better for same-gender conversions with results in the up-

Figure 1: HMM diagram for the proposed residual estimation algorithm.
Table 2: Average MOS results for every type of conversion. Male-to-Male (M-M), Female-to-Female (F-F), Male-to-Female (M-F), Female-to-Male (F-M).

<table>
<thead>
<tr>
<th>Test type</th>
<th>Avg. score (GMM)</th>
<th>Avg. score (HMM)</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test 1 M-M</td>
<td>5.85</td>
<td>5.90</td>
<td>0.05</td>
</tr>
<tr>
<td>Test 1 F-F</td>
<td>3.58</td>
<td>3.92</td>
<td>0.34</td>
</tr>
<tr>
<td>Test 1 M-F</td>
<td>2.79</td>
<td>3.94</td>
<td>1.15</td>
</tr>
<tr>
<td>Test 1 F-M</td>
<td>3.08</td>
<td>3.84</td>
<td>0.76</td>
</tr>
<tr>
<td>Overall</td>
<td>3.33</td>
<td>3.90</td>
<td>0.57</td>
</tr>
</tbody>
</table>

per middle of the MOS ‘Fair’ level (3.0 to 4.0), while for cross-gender conversions the MOS levels drop to the bottom of the ‘Fair’ level and even into the ‘Poor’ level (2.0 to 3.0). These results are consistent with what is reported in [3]. Table 2 shows that the proposed HMM-based system produces higher MOS scores than the GMM-based system and shows greater consistency in performance for both same-gender and cross-gender conversions. A statistical significance test on the MOS results confirms that the proposed system produces converted speech with significantly better quality than the baseline GMM-based system, except for the M-M case where both systems are comparable. The improvement in quality performance is particularly larger for the cross-gender conversion cases, where the 95% confidence intervals show that the mean difference can be more than one whole point in the MOS scale (which would imply a jump from one quality level to another). Normally, it is expected that for cross-gender conversions the differences between the vocal tract feature spaces of source and target speakers will be larger than in same-gender conversions, possibly resulting in lower performance of the vocal tract mapping algorithm affecting the resulting speech quality [3]. The MOS results for the proposed algorithm indicates that exploiting the temporal relationship between residuals can compensate the possible quality degradation introduced because of the lower performance of the vocal tract mapping procedure.

5. Conclusions

This work presents a comparison between two residual estimation algorithms for high resolution voice conversion. The results from subjective tests indicate that the proposed HMM-based system achieves its objective of significantly improving speech quality without compromising identity conversion. In particular, identity conversion improves the most when the target speaker is female, while speech quality gains are higher for cross-gender conversions, getting more consistent results among all conversion cases instead of presenting a marked dip in quality for cross-gender conversions, as is the case with the GMM-based algorithm. It can be concluded then that: a) Information contained in the temporal relationship between residuals in consecutive frames is particularly important for identity conversion when synthesizing female voices, and b) Said temporal information is also important to achieve better speech quality, especially for the cases where there is a higher mismatch between the feature spaces of source and target speakers (typically cross-gender conversions). The drawback of the alternative algorithm lies then in its higher computational complexity during the training phase with respect to the baseline, since training the HMM involves computing many more parameters than for a single GMM. Such complexity could also have a more pronounced impact on performance compared to the GMM-based approach when less amount of training data is available. Since work by others has shown that speech quality in VC systems can also be improved by modifying the vocal tract mapping function (e.g. [9, 10, 11]) instead of the residual estimation function, future work will explore the quality and identity conversion gains that could be achieved by working on both aspects simultaneously.

Note: This material is based upon work supported by the National Science Foundation under Grant No. 1116475. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

6. References