Probabilistic Principal Component Analysis Applied To Voice Conversion

Mark M. Wilde and Andrew B. Martinez
Electrical Engineering and Computer Science
Tulane University
New Orleans, Louisiana 70118–5674
Email: mwilde@tulane.edu

***Abstract*— In our model for voice conversion, we represent the joint probabilistic acoustic space of the source and target speakers with a mixture of Probabilistic Principal Component Analyzers (PPCAs). We present a finer resolution of options to the user of the voice conversion system than traditional Gaussian Mixture Model based conversion. Objective experiments demonstrate that the dimension of the PPCA directly impacts resulting objective performance but saves both time and memory complexity. Subjective tests imply that incremental removal of information does not affect the listener perceptually. Thus, the end user can select with more freedom how well the system should perform.**

I. Introduction

In voice conversion, we map the acoustic features of a *source* speaker to those of a *target* speaker. We collect speech in a parallel training corpus from both the source and target speaker for use in training the model. After training is complete, we predict what a target speaker sounds like using the information from the new speech of the source speaker.

We have chosen the *line spectrum frequencies* (LSFs) to represent the vocal tract features for a short frame of speech because of their desirable properties outlined in [1]. In order to have high quality conversion, we must convert both the excitation at the vocal cords and the spectral properties of the vocal tract. Early research used just the excitation of the source speaker with converted spectral features, but a study by Kain and Macon [1] determined that the excitation played an important role in identifying a speaker.

For mapping the excitation, we follow Kain's method of residual prediction in which we predict the excitation from the LP envelope [2].

II. Mixture Modeling for Spectral Conversion

Past researchers model the high dimensional probabilistic acoustic space of the just the source speaker or the joint density of the source and target speaker to determine the mapping for voice conversion. Stylianou modeled the acoustic probability space of the source speaker with a Gaussian Mixture Model (GMM) in [3]. He then found the cross-covariance of the target speaker with source speaker and the mean of the target speaker using least squares optimization of an overdetermined set of linear equations. In his work, he demonstrated the theoretical superiority of the GMM to codebook methods by showing that codebook methods are a special case of the GMM in which only the mean of a cluster is mapped. Kain extended Stylianou's work by modeling the joint probability density of both the source and target speakers [1], [2], [4]. Although this method increases the complexity during EM training, it obviates the need to perform the least squares optimization as with Stylianou's method. Modeling the joint probability density allows the system to capture all possible correlations between the source and target speaker's spectrum.

III. Proposed Method

We extend the spectral mapping aspect of voice conversion by modeling the joint probability space of both speakers with a mixture of Probabilistic Principal Component Analyzers (PPCAs). Previous methods that used the GMM to model the space are constrained to only two possible selections for representing covariance structure — diagonal and full covariance matrices. With diagonal structure, the training time is quick but conversion performance is sacrificed. With full covariances, we can model the underlying second order statistics with improved conversion performance but incur the penalty of longer training time.

By modeling covariance structure with a mixture of PPCAs, we provide an entire range of covariance structure that incrementally includes more covariance information. As we incrementally include more information, results indicate that the objective performance of the system incrementally improves. Subjective listening tests also indicate that the quality of conversion for incremental amounts is not perceptually noticeable. Thus, we present a wider array of options for voice conversion to the end user; and the user determines the tradeoff to fit the needs for the application.

IV. Spectral Conversion with a GMM

We first discuss spectral conversion with a GMM to provide the foundation which we extend upon in this paper. In order to estimate the parameters of the GMM, we use the classic expectation-maximization (EM) algorithm [5]. The limiting operation computationally in EM is the re-estimation of the j^{th} sample covariance matrix weighted by the j^{th} posterior component probability. The EM algorithm's complexity with fully populated covariance matrices is $\mathcal{O}(NMd^2)$ for each iteration where N is the amount of training data, M is the

number of components in the mixture, and d is the total dimensionality of the source and target LSFs $\boldsymbol{x}$ and $\boldsymbol{y}$.

Having estimated the parameters of the GMM, we can now estimate the target speaker's LSFs $\boldsymbol{y}$ from the source speaker's LSFs $\boldsymbol{x}$. The joint covariance matrix $\boldsymbol{\Sigma}_j$ for the j^{th} Gaussian component is partitioned as follows.

$$\boldsymbol{\Sigma}_j = \begin{bmatrix} \boldsymbol{\Sigma}_j^{xx} & \boldsymbol{\Sigma}_j^{xy} \\ \boldsymbol{\Sigma}_j^{yx} & \boldsymbol{\Sigma}_j^{yy} \end{bmatrix} \tag{1}$$

In the case of one component, a single Gaussian, the expectation $E[\boldsymbol{y}|\boldsymbol{x}]$ is the conditional mean of a joint Gaussian given by

$$\begin{aligned} E[\boldsymbol{y} \,|\, \boldsymbol{x}] &= \int \boldsymbol{y}\, p(\boldsymbol{y}|\boldsymbol{x})\, d\boldsymbol{y} && (2) \\ &= \boldsymbol{\mu}_y + \boldsymbol{\Sigma}_{yx}(\boldsymbol{\Sigma}_{xx})^{-1}\left(\boldsymbol{x} - \boldsymbol{\mu}_x\right) && (3) \end{aligned}$$

Extending to the mixture case as previously done in [3], the expectation is

$$E[\boldsymbol{y} \,|\, \boldsymbol{x}] = \sum_{j=1}^{M} P(j|\boldsymbol{x}) \left[\boldsymbol{\mu}_j^y + \boldsymbol{\Sigma}_j^{yx}(\boldsymbol{\Sigma}_j^{xx})^{-1}\left(\boldsymbol{x} - \boldsymbol{\mu}_j^x\right)\right] \tag{4}$$

where we weight the j^{th} conditional mean by the probability of component j given the information from the source vector $\boldsymbol{x}$.

A. Decorrelation Assumption

Assuming that the individual random variables of the source and target vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ respectively are decorrelated is a constraint that previous researchers [3], [6] have placed on $\boldsymbol{x}$ and $\boldsymbol{y}$ to speed up training and conversion time. In order to benefit from these savings, researchers have elected either to impose the structure of each submatrix of $\boldsymbol{\Sigma}_j$ or of the entire covariance matrix $\boldsymbol{\Sigma}_j$ to be diagonal. When assuming this decorrelation between the various components of the feature vectors, the computational time for training with EM is significantly reduced to $O(NMd)$.

B. Weakness of the Decorrelation Assumption

Although the above diagonalizing methods can significantly reduce computational time, we should note that this restriction is inappropriate because the feature vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are not completely decorrelated. This decorrelation assumption implies that in the d-dimensional feature space, all of the covariance structures are aligned with the feature space axes. Another way to view the diagonal constraint is that it is a primitive method of reducing the dimensionality of covariance structure from d^2 to d by imposing its structure to be diagonal.

Since only these two extremes, diagonal or full covariance matrices, are available, a need exists for a method which can fill in the "spectrum" of options between the extremities. With only these two extremes, the end user must make a difficult tradeoff between the time that it takes to train the system and the resulting performance desired.

V. Voice Conversion with a Mixture of Probabilistic Principal Component Analyzers

Although the GMM has become quite popular recently for modeling complex probability densities, one of its shortcomings is that an increase in the dimensionality of the feature space increases the complexity of the model. Each covariance matrix $\boldsymbol{\Sigma}_j$ in the mixture becomes excessively large, and estimation of each sample covariance matrix and its inverse is less tractable to compute as dimensionality increases.

Probabilistic Principal Component Analysis (PPCA), a method developed by Tipping and Bishop [7], solves the inflexibility of GMMs by performing a pseudo *local Principal Component Analysis* on each component of the mixture.

A. Probabilistic Principal Component Analysis

PPCA's statistical model assumes that a set of q latent variables $\boldsymbol{f}$ are responsible for generating the d-dimensional data set $\boldsymbol{z}$ as given in the following equation where $q < d$.

$$\begin{aligned} \boldsymbol{z} &= \boldsymbol{W}\boldsymbol{f} + \boldsymbol{\mu} + \boldsymbol{\epsilon} && (5) \\ \boldsymbol{f} &\sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) && (6) \\ \boldsymbol{\epsilon} &\sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi}) && (7) \end{aligned}$$

We place the constraint that the latent variables $\boldsymbol{f}$ are independent and Gaussian with unit variance, and the noise variables $\boldsymbol{\epsilon}$ are independent and Gaussian with covariance $\boldsymbol{\Psi}$. In the PPCA model, we restrict $\boldsymbol{\Psi}$ to have isotropic variance $\sigma^2\boldsymbol{I}$. In addition, the factors $\boldsymbol{f}$ are independent of the noise $\boldsymbol{\epsilon}$. The $d \times q$ matrix $\boldsymbol{W}$ contains the factor loadings, and the parameter vector $\boldsymbol{\mu}$ permits the data to have non-zero mean. Under these assumptions, the observations $\boldsymbol{z}$ are Gaussian with mean $\boldsymbol{\mu}$ and model covariance $\boldsymbol{C} = \boldsymbol{W}\boldsymbol{W}^T + \sigma^2\boldsymbol{I}$.

B. Mixtures of Probabilistic Principal Component Analyzers

Because PPCA has a generative model, we can combine several of these models into a mixture of PPCAs as described in [7]. The form of the mixture is the following

$$p(\boldsymbol{z}) = \sum_{j=1}^{M} \int p(\boldsymbol{z}|\boldsymbol{f}_j) p(\boldsymbol{f}_j|j) P(j) d\boldsymbol{f} \tag{8}$$

where we represent the j^{th} component of the mixture with a PPCA model. We find certain similarities between this model and the GMM — the only difference is that with a mixture of PPCAs we represent each of the M component densities with a single PPCA model rather than with a multivariate normal.

C. EM for a Mixture of PPCAs

We can compute the parameters of the mixture of PPCAs with the two stage EM algorithm formulated in [7]. In the first stage, we ignore the latent variables $\boldsymbol{f}$ and compute the expected log likelihood; then we maximize $P(j)$ and $\boldsymbol{\mu}_j$. In the second stage, we increase the likelihood again by maximizing $\boldsymbol{W}_j$ and σ_j^2 with the following equations

$$\begin{aligned} \hat{\boldsymbol{W}}_j &= \boldsymbol{S}_j\boldsymbol{W}_j\left(\sigma_j^2\boldsymbol{I} + \boldsymbol{M}_j^{-1}\boldsymbol{W}_j^T\boldsymbol{S}_j\boldsymbol{W}_j\right)^{-1} && (9) \\ \hat{\sigma}_j^2 &= \frac{1}{d}\,\mathrm{tr}\left\{\boldsymbol{S}_j - \boldsymbol{S}_j\boldsymbol{W}_j\boldsymbol{M}_j^{-1}\hat{\boldsymbol{W}}_j\right\} && (10) \end{aligned}$$

where $\boldsymbol{S}_j$ is the weighted sample covariance matrix for the j^{th} component and $\boldsymbol{M}_j = (\boldsymbol{W}_j^T \boldsymbol{W}_j + \sigma^2 \boldsymbol{I})$. Although it is convenient to include $\boldsymbol{S}_j$ in the above equation, its computation, an $\mathcal{O}(NMd^2)$ operation, is not explicitly necessary. It is only necessary to evaluate the trace of the j^{th} weighted sample covariance matrix, an $\mathcal{O}(NMd)$ operation; and we evaluate $\boldsymbol{S}_j \boldsymbol{W}_j$ as

$$\boldsymbol{S}_j \boldsymbol{W}_j = \frac{1}{\hat{\pi}_j N} \sum_{n=1}^{N} R_{nj} [\boldsymbol{z}_n - \boldsymbol{\mu}_j] \left\{ [\boldsymbol{z}_n - \boldsymbol{\mu}_j]^T \boldsymbol{W}_j \right\} \quad (11)$$

where R_{nj} is the *responsibility* of the j^{th} component and $\hat{\pi}_j$ is the j^{th} mixing coefficient as described in [7]. Equation 11 is the limiting computation in the two stage EM formulation with a complexity of $\mathcal{O}(NMdq)$. With this formulation, we can use q, the amount of information we are willing to keep, to vary the training complexity.

D. Spectral Conversion with PPCAs

In order to convert the spectrum in the PPCA case, we calculate the expectation of the target vector $\boldsymbol{y}$ given the source vector $\boldsymbol{x}$ for the single PPCA model and then extend the result to a mixture of PPCAs. To find this expectation, we partition $\boldsymbol{z}$ so that its joint multivariate density is

$$\mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{bmatrix}, \begin{bmatrix} \boldsymbol{W}_x \boldsymbol{W}_x^T + \sigma^2 \boldsymbol{I} & \boldsymbol{W}_x \boldsymbol{W}_y^T \\ \boldsymbol{W}_y \boldsymbol{W}_x^T & \boldsymbol{W}_y \boldsymbol{W}_y^T + \sigma^2 \boldsymbol{I} \end{bmatrix} \right) \quad (12)$$

where we partition the mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{C}$. Then, using Equation 3, the conditional expectation of a joint Gaussian, we find that

$$E[\boldsymbol{y}|\boldsymbol{x}] = \boldsymbol{W}_y \boldsymbol{W}_x^T \left(\boldsymbol{W}_x \boldsymbol{W}_x^T + \sigma^2 \boldsymbol{I} \right)^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_x) + \boldsymbol{\mu}_y \quad (13)$$

Extending the result from above to the case of a mixture of PPCAs, our statistical mapping is

$$\sum_{j=1}^{M} P(j|\boldsymbol{x}) \boldsymbol{W}_y \boldsymbol{W}_x^T \left(\boldsymbol{W}_x \boldsymbol{W}_x^T + \sigma^2 \boldsymbol{I} \right)^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_x) + \boldsymbol{\mu}_y \quad (14)$$

An equivalent expression for the expectation from Equation 13 is

$$E[\boldsymbol{y}|\boldsymbol{x}] = \boldsymbol{W}_y \left(\boldsymbol{W}_x^T \boldsymbol{W}_x + \sigma^2 \boldsymbol{I} \right)^{-1} \boldsymbol{W}_x^T (\boldsymbol{x} - \boldsymbol{\mu}_x) + \boldsymbol{\mu}_y \quad (15)$$

as derived in [8] by using the Matrix Inversion Lemma. The above formulation can be extended to the mixture case as well.

The two above results in Equations 13 and 15 are similar to the mapping found for GMMs except that with the PPCA we are now free to remove dimensionality. When choosing which expression to use for spectral conversion, we should determine first whether $q < \frac{d}{2}$ for computational purposes. If it is, then we should use Equation 15 for converting. Instead of computing the inverse of the $\frac{d}{2} \times \frac{d}{2}$ matrix $\boldsymbol{W}_x \boldsymbol{W}_x^T + \sigma^2 \boldsymbol{I}$, we compute the inverse of the $q \times q$ matrix $\boldsymbol{W}_x^T \boldsymbol{W}_x + \sigma^2 \boldsymbol{I}$ with a total complexity of $\mathcal{O}(Mq^3)$ for all components in the mixture. This decrease in complexity may not be that significant since we only compute it once with our method for voice conversion; but, if we update the model with an online EM algorithm by including novelty test data as done in [9], using this method reduces complexity of conversion. If $q > \frac{d}{2}$, then we should use Equation 14 for conversion.

E. PPCA as a General Case of Spectral Conversion with a GMM

We now demonstrate with a simple proof how conversion with the mixture of PPCAs is a general case of GMM conversion in which we can take away dimensionality. Because the covariance matrix $\boldsymbol{\Sigma}$ is symmetric, we can factor it with an eigen-decomposition.

$$\boldsymbol{\Sigma} = \boldsymbol{U} \boldsymbol{\Lambda} \boldsymbol{U}^T = (\boldsymbol{U} \boldsymbol{\Lambda}^{\frac{1}{2}})(\boldsymbol{U} \boldsymbol{\Lambda}^{\frac{1}{2}})^T \quad (16)$$

This decomposition is equivalent to the case when we retain all principal components with $\sigma^2 \rightarrow 0$. So, setting $\boldsymbol{W} = \boldsymbol{U} \boldsymbol{\Lambda}^{\frac{1}{2}}$, we can formulate $\boldsymbol{\Sigma}$ as $\boldsymbol{W} \boldsymbol{W}^T$ when retaining all components.[1] Partitioning $\boldsymbol{W}$ into $\boldsymbol{W}_x$ and $\boldsymbol{W}_y$ gives the following.

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{W}_x \\ \boldsymbol{W}_y \end{bmatrix} \begin{bmatrix} \boldsymbol{W}_x^T & \boldsymbol{W}_y^T \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy} \end{bmatrix} \quad (17)$$

We determine from above that $\boldsymbol{\Sigma}_{xx} = \boldsymbol{W}_x \boldsymbol{W}_x^T$ and $\boldsymbol{\Sigma}_{yx} = \boldsymbol{W}_y \boldsymbol{W}_x^T$. Substituting into the expression for the conditional mean of a joint Gaussian from Equation 3, we find that

$$E[\boldsymbol{y}|\boldsymbol{x}] = \boldsymbol{W}_y \boldsymbol{W}_x^T \left(\boldsymbol{W}_x \boldsymbol{W}_x^T \right)^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_x) + \boldsymbol{\mu}_y \quad (18)$$

and notice that this solution is the same as Equation 13 in the limit as $\sigma^2 \rightarrow 0$.

Thus, PPCA is a more general case of the GMM for spectral conversion. In PPCA, we have the flexibility of removing dimensions from $\boldsymbol{W}$ which in turn adds to σ^2 the average variance not captured in the projection. We notice a relationship: the more dimensions that we take away from $\boldsymbol{W}$, the farther away we get from the "true" conversion with a GMM. The question now is how far we can go away before it perceptually makes a difference to the human ear. We assess this question with both objective and subjective measurements in the next section.

VI. Objective Evaluation

In evaluating our system, we use the objective measure given by Kain [2]. This *performance index* is a ratio of two measures. The first measure, the *transspeaker distance*, is the spectral distance between the converted speech and the target speech determining how "close" the converted speech is to the target speaker's. The second, the *interspeaker distance*, measures the spectral distance between the source and target speaker. To present the performance index, let us again consider the vector of source speech for the n^{th} frame as $\boldsymbol{x}_n$

[1] $\boldsymbol{W}$ in this case is the maximum likelihood estimate for the factor loadings given in [7] without the noise term σ^2.

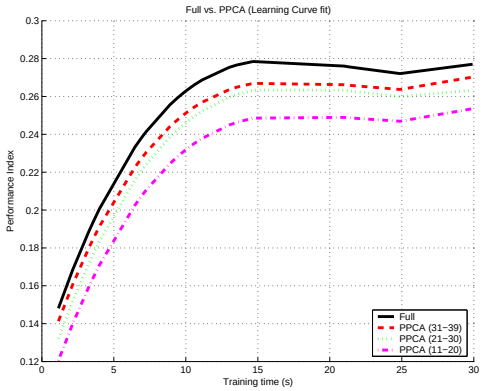

Fig. 1. Performance Index Learning Curve for a Full GMM and a mixture of PPCAs with varying dimensionality.

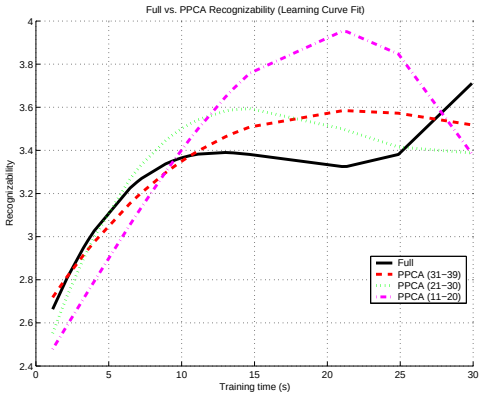

Fig. 2. Subjective Recognizability Learning Curve for a Full GMM and a mixture of PPCAs with varying dimensionality.

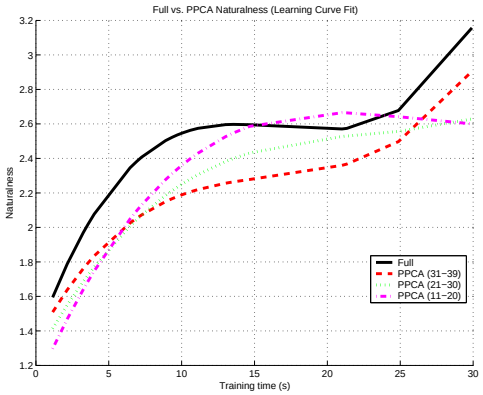

Fig. 3. Subjective Naturalness Learning Curve for Full GMMs and a mixture of PPCAs with varying dimensionality.

and the target speaker's n^{th} vector as $\boldsymbol{y}_n$. We compute the performance index p with the following equation

$$p = 1 - \sum_{n=1}^{N} \frac{d(\boldsymbol{y}_n, \hat{\boldsymbol{y}}_n)}{d(\boldsymbol{y}_n, \boldsymbol{x}_n)} \tag{19}$$

where $\hat{\boldsymbol{y}}_n$ is the converted target vector and the distance d is Euclidean.

We used a database of three speakers to assess conversion performance (corresponding to six conversions). In Figure 1, we fit each set of objective data points with its appropriate "learning curve" benchmarking the mixture of PPCAs against GMM-based conversion. As the training time increases, the performance of each system exponentially increases until it plateaus around 14 seconds of training data. We conclude that 14 seconds of training data is enough for the system to give a reasonable *average* performance between 0.24 and 0.28. We expect the average performance to increase slightly with more than 30 seconds of training data.[2] We show the mixture of PPCAs for a range of values for q — the number of dimensions we keep. Note that q varies from one dimension to 39 dimensions because we set the LP model order for each speaker to 20; thus, the joint space of both speakers has dimension 40.

The interesting observation is that for each reduction of q, the performance slightly decreases as we predicted previously.[3] The only question is whether this incremental decrease in information is perceptually noticeable to a listener.

VII. Subjective Listening Tests

In order to determine if the human ear notices the gradual loss of information, we conducted subjective listening tests with seven listeners and three speakers. Before starting the test, we trained each listener on the three speakers' distinct voice qualities by playing several of their speech files. After this initial training, we quizzed each listener until they identified each speaker correctly for ten consecutive trials.

During the test, we played converted speech in a randomized order varying both the PPCA dimensionality and the speaker. We played the same sentence so that listeners could focus intently on the quality of each recording.

First, the listener identified the speaker; overall, the accuracy was 79.2%. After determining the speaker, the listener provided a subjective answer for three categories. *Recognizability* indicates how easy it is for the listener to identify the speaker. *Naturalness* describes how much like a human the speech sounds. *Quality* is a subjective measure of how clean the speech signal is. The listeners chose a subjective score between one and five for each of the above categories.

A. Subjective Test Results

In Figure 2, we plot the mean recognizability learning curve for each model versus the training time. Note that recognizability of each system improves as the training time increases; but from the listeners' responses, we conclude that it is difficult for them to distinguish the difference between full covariances and reduced dimension covariances.

We plot the mean naturalness learning curve in Figure 3. It is apparent that the full GMM method performs mostly better

[2]Kain's *best* performance index was 0.31 [2], and Gillett's was 0.36 [10] with 120 seconds of training data.

[3]Note that this general trend occurs for each unit decrease of q, but we only show the average decrease for a range of values for q.

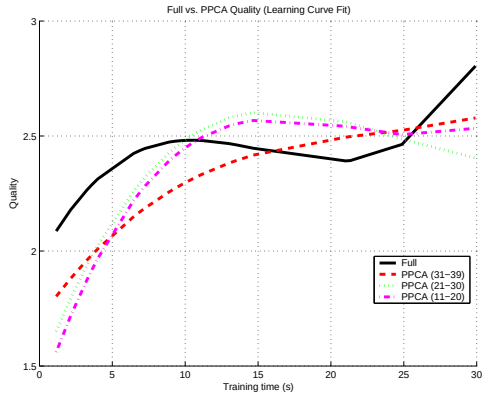

Fig. 4. Subjective Quality Learning Curve for a Full GMM and a mixture of PPCAs with varying dimensionality.

than PPCA though its performance is equivalent for some time to the PPCAs when q is between 11 and 20. The fact that lower dimensions perform better than higher dimensions indicate that perhaps a little more testing is needed for this case.

Lastly, in Figure 4, we illustrate the subjective quality learning curve. Again, nothing indicates that one system's quality outperforms the others. All learning curves are relatively close so we can conclude with some degree of confidence that listeners do not notice the difference when we reduce dimensionality.

Based on these tests of the three subjective categories, it is apparent that each category's performance increases with the amount of training data. Although the performance increases with more training data, incrementally removing "less relevant" information does not affect the end listener perceptually. Thus, in using PPCA, we have the benefit of reduced time and memory complexity for both training and conversion without affecting the end listener's perceptual experience.

VIII. Concluding Remarks and Future Opportunities

In summary, we have applied the mixture of PPCAs model to voice conversion. Using a mixture of PPCAs reduces the time complexity for training from $\mathcal{O}(MNd^2)$ to $\mathcal{O}(MNdq)$ and conversion time from $\mathcal{O}(M\left(\frac{d}{2}\right)^3)$ to $\mathcal{O}(Mq^3)$ if $q < \frac{d}{2}$ with the penalty of a slight decrease in the objective quality of conversion. Although the objective measure can detect that we have reduced the amount of information, the human ear has difficulty perceiving this reduction. Therefore, we recommend that the user of the system select an appropriate value of q to suit performance needs of the application. Obviously we always want the system to perform with highest quality; but, this high quality only comes with the price of expensive computations. For voice conversion training and execution, the mixture of PPCAs model provides a flexible range of tradeoffs to select from.

We have several ways that this system can be improved upon by incorporating the recent advances of variational Bayesian modeling, independent component analysis, and an on line EM algorithm.

By incorporating variational Bayesian techniques with PPCA as described in [11], the system can automatically estimate the appropriate model order for q, the amount of information to retain of each component in the mixture, and M, the number of components in the mixture.

Using a Gaussian for each component in the mixture is controversial because we do not have infinite training data. We can solve this problem by using a mixture of independent component analyzers [12] where each component's distribution is non-Gaussian and thus we could describe the density of each component in the mixture more properly.

In our model, we only estimate the probability density from a fixed set of training data. Including novelty test data in the model is expensive with our current method because it requires a complete re-estimation via EM. However, by updating the model on line with the testing data of the source speaker, we could improve performance significantly. Additionally, by using this model, we could realize computational savings for conversion with Equation 15 if $q < \frac{d}{2}$.

IX. Acknowledgements

The authors are indebted to Ben Gillett and Simon King from the University of Edinburgh for providing a base system for voice conversion which helped tremendously in developing and benchmarking our methods.

References

[1] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 1998, pp. 285–288.

[2] A. Kain, "High resolution voice transformation," Ph.D. dissertation, OGI School of Science and Engineering at Oregon Health and Science University, 2001.

[3] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, March 1998.

[4] A. Kain and Y. Stylianou, "Stochastic modeling of spectral adjustment for high quality pitch modification," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 2000.

[5] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of Royal Statistical Society*, vol. 39, pp. 1–38, 1977.

[6] A. Mouchtaris, S. S. Narayanan, and C. Kyriakakis, "Maximum likelihood constrained adaptation for multichannel audio synthesis," in *36th IEEE Asilomar Conference on Signals, Systems, and Computers*. Pacific Grove, CA: IEEE, November 2002.

[7] M. E. Tipping and C. M. Bishop, "Mixtures of probabilistic principal component analysers," *Neural Computation*, vol. 11, no. 2, pp. 443–482, 1999.

[8] M. M. Wilde, "Controlling performance in voice conversion with probabilistic principal component analysis," Master's thesis, Tulane University, May 2004.

[9] H. Duxans and A. Bonafonte, "Estimation of gmm in voice conversion including unaligned data," in *Eurospeech*. Geneva, Switzerland: 8^{th} European Conference on Speech Communication and Technology, September 2003.

[10] B. Gillett, "Transforming voice quality and intonation," Master's thesis, University of Edinburgh, The Centre for Speech Technology Research, January 2003.

[11] C. M. Bishop, "Variational principal components," in *Proceedings of the 9^{th} International Conference on Artificial Neural Networks*, vol. 1, 1999, pp. 509–514.

[12] S. J. Roberts and W. D. Penny, "Mixtures of independent component analysers," in *International Conference on Artificial Neural Networks*, 2001, pp. 527–534.