Text-to-Speech Voice Adaptation from Sparse Training Data

Alexander Kain (kain@cse.ogi.edu)
Michael Macon (macon@ece.ogi.edu)

Center for Spoken Language Understanding (CSLU)
Oregon Graduate Institute of Science and Technology
P.O. Box 91000, Portland, OR 97291-1000, USA

ABSTRACT

Voice adaptation describes the process of converting the output of a text-to-speech synthesizer voice to sound like a different voice after a training process in which only a small amount of the desired target speaker's speech is seen. We employ a locally linear conversion function based on Gaussian mixture models to map bark-scaled line spectral frequencies. We compare performance for three different estimation methods while varying the number of mixture components and the amount of data used for training. An objective evaluation revealed that all three methods yield similar test results. In perceptual tests, listeners judged the converted speech quality as acceptable and fairly successful in adapting to the target speaker.

1. INTRODUCTION

Voice conversion systems aim to modify a source speaker's speech so that it is perceived to be spoken by a different target speaker. Integrating voice conversion technologies into a concatenative speech synthesizer allows for the production of additional voices from a single source-speaker database. When this system is used to "personalize" a synthesizer to speak with any desired voice, we refer to the process as "voice adaptation".

As an extension of our previous work [2], this paper explores issues related to the goal of performing voice adaptation using only a small amount of adaptation data. This is desirable because users want to adapt a new voice quickly and with as little speech as possible. This paucity of the data limits the scope of adaptation algorithms to segmental properties only, such as pitch and spectral characteristics related to vocal tract size and shape. The general approach is to find a regression mapping between features in the source and target spaces. The generalization of this mapping to unseen cases is critical to our application. The mapping function is a probabilistic, locally-linear function based on a Gaussian mixture model (GMM) estimated from source and target feature densities. We will discuss the advantages of modeling the joint density rather than using a least-squares solution approach published recently.

This choice of features and mapping function can be used to gain insight into the relationship and dimensionality of spectral differences between two speakers. This information can be used to constrain the mapping technique to be robust to sparse training data. In particular, because line spectral frequency (LSF) features are related to formant frequencies, constraining the GMM components to have constant diagonals approximates frequency warping – another common approach to spectral voice conversion. The next section introduces the fundamentals of the system, while Sections 3 and 4 discuss the setup and evaluation of an experiment to compare performance with different estimation methods and number of mixture components, and with different amounts of training data.

2. VOICE CONVERSION SYSTEM

2.1. Features

The sparseness of the training data limits the scope of the adaptation algorithm to segmental properties only, specifically to average pitch and spectral characteristics related to vocal tract size and shape. Bark-scaled LSFs were chosen as spectral features because of the following properties:

- Errors are localized in frequency: a badly predicted vector component effects only a portion of the frequency spectrum adversely.
- LSFs have good linear interpolation characteristics [5]. This is essential because the conversion function linearly combines vectors.
- LSFs relate well to formant location and bandwidth, which have been shown to be perceptually relevant for speaker identity.
- The training cost function employs a mean squared error measure; hence a bark scaling weights prediction errors in accordance with the frequency sensitivity of human hearing (more sensitive to frequency changes at lower frequencies).

2.2. Spectral Mapping

Let $x = [x_1\ x_2 \cdots x_N]$ be the sequence of features characterizing a succession of speech sounds produced by the source speaker and $y = [y_1\ y_2 \cdots y_N]$ be features describing those same sounds as produced by the target speaker.

A GMM allows the probability distribution of x to be written as the sum of Q multivariate Gaussian functions,

$$p(x)=\sum_{i=1}^{Q}\alpha_i N(x;\mu_i,\Sigma_i),\quad \sum_{i=1}^{Q}\alpha_i=1,\quad \alpha_i\geq 0, \tag{1}$$

where $N(x;\mu,\Sigma)$ denotes a normal distribution with mean vector μ and covariance matrix Σ, and α_i denotes the prior probability of class i. The parameters of the model (α,μ,Σ) can be estimated using the well-known expectation maximization (EM) algorithm.

The goal is to compute a conversion function F that minimizes the mean squared error

$$\varepsilon_{mse}=E\left[\|y-F(x)\|^2\right], \tag{2}$$

where E denotes expectation. The conversion function is chosen to be a probabilistic, locally linear mapping function

$$F(x)=\sum_{i=1}^{Q}h_i(x)\left[v_i+\Gamma_i\Sigma_i^{-1}(x-\mu_i)\right], \tag{3}$$

where $h_i(x)$ is the posterior probability that the i^{th} Gaussian component generated x, calculated by application of Bayes theorem

$$h_i(x)=\frac{\alpha_i N(x;\mu_i,\Sigma_i)}{\sum_{j=1}^{Q}\alpha_j N(x;\mu_j,\Sigma_j)}. \tag{4}$$

A simpler form of the conversion function can be obtained by rewriting (3) as

$$F(x)=\sum_{i=1}^{Q}h_i(x)\left[W_i x+b_i\right], \tag{5}$$

where $W_i=\Gamma_i\Sigma_i^{-1}$ and $b_i=v_i-W_i\mu_i$.

In one approach [6], the parameters (α,μ,Σ) of a GMM are estimated to model the distribution of x. Then the unknowns (v,Γ) are computed by solving normal equations for a least squares problem based on the correspondence between the source and target. We will call this the least squares (LS) estimation method.

Another approach, used in our previous work [2], vertically joins the source vectors with the target vectors to form

$$z=\begin{bmatrix}x\\ y\end{bmatrix}. \tag{6}$$

GMM parameters (α,μ,Σ) are estimated for the density $p(z)$, which is the joint density $p(x,y)$ [3]. The conversion function that minimizes the mean squared error between converted source and target vectors is the regression

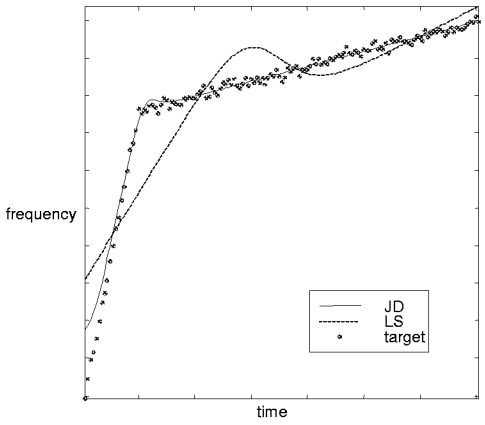

Figure 1: One-dimensional example demonstrating results from two different conversion function estimation methods.

$$\begin{aligned}F(x)&=E[y\mid x]=\int dy\; y\,p(y\mid x)\\&=\sum_{i=1}^{Q}h_i(x)\left[\mu_i^y+\Sigma_i^{yx}(\Sigma_i^{xx})^{-1}\left(x-\mu_i^x\right)\right]\end{aligned} \tag{7}$$

where

$$h_i(x)=\frac{\alpha_i N\left(x;\mu_i^x,\Sigma_i^{xx}\right)}{\sum_{j=1}^{Q}\alpha_j N\left(x;\mu_j^x,\Sigma_j^{xx}\right)}, \tag{8}$$

with $\Sigma_i=\begin{bmatrix}\Sigma_i^{xx} & \Sigma_i^{xy}\\ \Sigma_i^{yx} & \Sigma_i^{yy}\end{bmatrix}$ and $\mu_i=\begin{bmatrix}\mu_i^x\\ \mu_i^y\end{bmatrix}$.

The joint density (JD) method estimates mixture components based on observations of **both** the source and the target vectors, and makes no assumptions about the target distributions, whereas the LS method clusters are based on the source vector distributions only.

Modeling the joint density rather than only the source density can lead to a more judicious allocation of mixtures components. This is demonstrated in Figure 1 with the aid of a simplified, one-dimensional problem. Suppose we needed to map from a linear source trajectory with fixed slope to a more complex target trajectory shown. Training a conversion function with two mixture components results in a fairly accurate match in the JD case, while the LS case has large deviations from the target. This is due to the fact that LS constructed clusters without taking into account the target distributions.

During the EM step, JD is computationally more expensive than LS because the dimensionality of the space to be estimated has doubled. However, no extra solution step is required. In addition, the largest matrix in the solution step of LS requires several times more memory than is required for

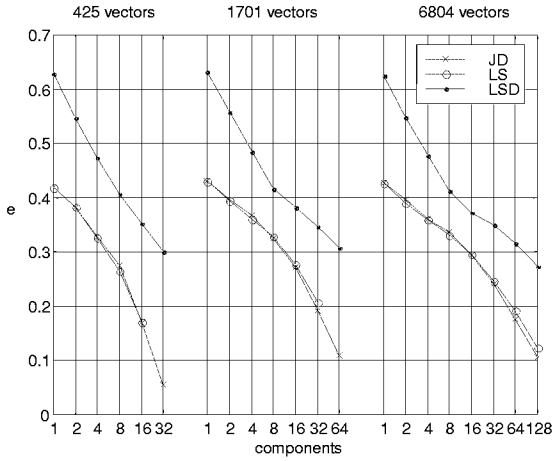

Figure 2: Training errors produced by different estimation methods. The three set of lines represents an increasing amount of data seen during training. Within each set, the number of mixture components is varied as indicated by the axis labels.

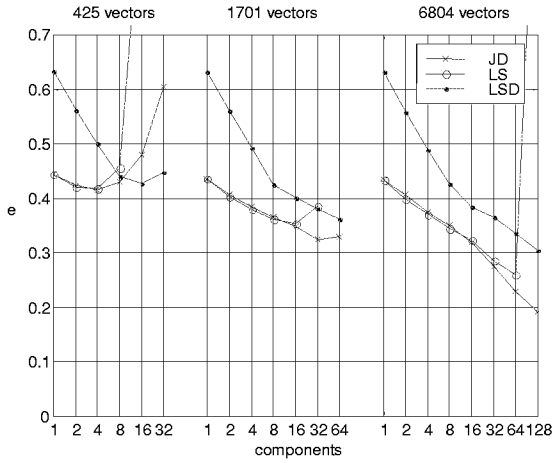

Figure 3: Test errors. The test set size is fixed throughout.

JD. Finally, LS necessitates approximately twice the number of operations as JD during training.

3. EXPERIMENT

3.1. Speech Material

It is our goal to convert a text-to-speech synthesizer's voice to a new voice. Therefore the source speaker speech is the output of a synthesizer, while the target speaker is recorded. For simplicity, we assume "cooperative" training, where it is possible to obtain any desired target speaker utterance. We have used the Festival Text-to-Speech Synthesis System [1] with the OGIresLPC module [4], both freely available for research purposes, and another commercial product as speech synthesizer sources.

The training corpus in our experiment consisted of 31 short words and 8 sentences, yielding about one minute of speech. After acquisition of the source and the target utterances the speech was force-aligned phonetically. Subsequently, the phonetic boundaries were checked and corrected by hand. LSFs were extracted with a fixed frame rate of 10ms from the pre-emphasized speech. Finally, the data were assigned to training sets of three different sizes.

Compared to our previous work [2] in which we used studio recorded diphone databases, these datasets are more realistic in that the target speech is continuous and has been recorded in an office environment. Furthermore, the spectral estimates are noisy due to the pitch-asynchronous feature extraction.

3.2. Training

We compared three different methods of training: JD and LS with full covariance matrices and a special case of LS where the covariance and mapping matrices are diagonal (LSD) [7]. This last case approximates a frequency warping function when used in conjunction with LSFs. The EM algorithm was run for 15 iterations. To prevent singularities, a small value was added to the diagonal elements of the covariance matrices after each iteration. For each training set size, the number of mixture components was varied as a power of 2 between 1 and up to 128.

3.3. Conversion

To obtain a converted utterance, spectral features are extracted exactly as during training and then mapped to new features by the conversion function whose parameters were estimated during the training process. The pitch of the source speaker's residual is scaled to match, on average, the target speaker's pitch. The modified residual and the new spectral parameters are re-convolved to render the final converted speech.

4. EVALUATION

4.1 Objective Evaluation

To objectively evaluate the spectral conversion function performance at various operating points we measure the normalized mean squared error

$$\varepsilon_{norm\,mse} = \frac{\frac{1}{N}\sum_{n=1}^{N}\left\|y_n - F(x_n)\right\|^2}{\frac{1}{N}\sum_{n=1}^{N}\left\|y_n - \mu^y\right\|^2}. \qquad (9)$$

We measured errors on both the training and a test set, which was obtained by holding out 20% of the vectors of the total available dataset. The errors presented are averages over three rotations with different data held out each time.

Figures 2 and 3 summarize the training and test errors for the three training methods. Each cluster of lines represents results for three increasing training set sizes. We observe that, for the most part, LS and JD perform comparatively. This behavior seems to indicate that the target distributions are similar to the source distributions in respect to their variance. Parameters for LS are not estimated reliably when a large number of mixture components is used due to numerical difficulties in the solution of normal equations. In terms of the total number of parameters used, all three methods perform very similarly (the number of mixture components needs to be 2 to 3 times higher when diagonal covariances are used as compared to full covariance matrices).

The first two test sets have minima, indicating the "best case" in terms of number of components. When even more components are used, overtraining occurs as is indicated by rising test errors. Naturally, the test set errors decrease for mappings that have been trained on more data.

In the first test set, the best case LSD yields about the same accuracy as the best case LS/JD. This is interesting because it suggests that a frequency warping, which modifies the spectral feature components individually, is as effective as a general affine operation on the entire feature vector.

Inspecting the parameters of a conversion function when the number of mixture components is set to one, and only diagonal elements are estimated, provides information about the average linear operation on each of the bark-scaled LSF components. In other words, the parameters give an overall offset and a scaling factor for every LSF between the source and the target spectra. For example, parameters for a male to female conversion yield a bias of exclusively positive numbers, which indicates that all formant positions will be higher in the converted utterance. Further, the relative offsets between the lower and higher value of a line frequency pair show an increase in formant bandwidth. Both higher formants and increased bandwidth are typical of female speech.

4.2 Subjective Evaluation

A systematic, yet informal, perceptual test was carried out to assess the resulting quality. Five listeners were asked to compare the intelligibility and quality of speech converted with different methods and amounts of training data. In each case, the number of mixture components was chosen to yield the lowest test error. Overall, the speech quality was found to be acceptable, though not as high as the original synthesis voice. Participants reported only slight differences in quality between all the methods.

When asked about how close the converted speech was to "sounding" like the target speaker, opinions varied depending on the particular source/target pair. In general, male to male conversions were perceived as more successful than male to female conversions.

To hear audio examples of the voice adaptation and other systems, please visit the web site at http://cse.ogi.edu/cslu/tts.

CONCLUSIONS

Although the JD method has theoretical advantages, the LS method yields almost the same accuracy. However, JD has the implementation advantages of requiring less operations and less memory.

Even though LSFs are not statistically independent, the LSD method achieves accuracy comparable to the JD method. Future work will involve deeper analysis of the relationship between frequency warping and the LSD method.

Informal perceptual tests reveal that the subjective quality is acceptable, even though the speaker identity of the target has only been partially adapted to.

ACKNOWLEDGEMENTS

This work was supported by a grant from Intel Corporation. The authors thank the member companies of CSLU and Fluent Speech Technologies.

REFERENCES

1. A. W. Black and P. Taylor, "The Festival speech synthesis system: System documentation," Tech. Rep. HCRC/TR-83, Human Communication Research Centre, University of Edinburgh, Scotland, UK, January 1997.

2. A. Kain and M. Macon, "Spectral voice conversion for text-to-speech synthesis," Proceedings of ICASSP, pp. 285-288, May 1998.

3. N. Kambhatla, *Local Models and Gaussian Mixture Models for Statistical Data Processing*, Ph.D. thesis, Oregon Graduate Institute, Portland, OR, January 1996.

4. M. Macon, A. Cronk, J. Wouters, and A. Kain, "OGIresLPC: Diphone synthesizer using residual-excited linear prediction," Tech. Rep. CSE-97-007, Department of Computer Science, Oregon Graduate Institute of Science and Technology, Portland, OR, September 1997.

5. K. K. Paliwal, "Interpolation properties of linear prediction parametric representations," Proceedings of EUROSPEECH, pp. 1029-32, September 1995.

6. Y. Stylianou, O. Cappé, and E. Moulines, "Statistical methods for voice quality transformation," Proceedings of EUROSPEECH, pp. 447-450, September 1995.

7. Y. Stylianou, *Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification*, Ph.D. thesis, Ecole Nationale Supérieure des Télécommunications, Paris, France, January 1996.