SPECTRAL VOICE CONVERSION FOR TEXT-TO-SPEECH SYNTHESIS

Alexander Kain and Michael W. Macon[*]

Center for Spoken Language Understanding (CSLU)
Oregon Graduate Institute of Science and Technology
P.O. Box 91000, Portland, OR 97291-1000, USA
http://cse.ogi.edu/cslu kain@cse.ogi.edu macon@ee.ogi.edu

ABSTRACT

A new voice conversion algorithm that modifies a source speaker's speech to sound as if produced by a target speaker is presented. It is applied to a residual-excited LPC text-to-speech diphone synthesizer. Spectral parameters are mapped using a locally linear transformation based on Gaussian mixture models whose parameters are trained by joint density estimation. The LPC residuals are adjusted to match the target speaker's average pitch. To study effects of the amount of training on performance, data sets of varying sizes are created by automatically selecting subsets of all available diphones by a vector quantization method. In an objective evaluation, the proposed method is found to perform more reliably for small training sets than a previous approach. In perceptual tests, it was shown that nearly optimal spectral conversion performance was achieved, even with a small amount of training data. However, speech quality improved with an increase in training set size.

1. INTRODUCTION

Voice conversion is a technique that modifies a source speaker's speech to be perceived as if a target speaker had spoken it. Integrating voice conversion technologies into concatenative speech synthesizers would allow the production of additional voices from a single-speaker database, as well as the "personalization" of the synthesizer to speak with any desired voice after an adaptation process. The goal of this work was to explore the application of a new voice conversion technique to a concatenation-based synthesizer.

The *Festival* Text-to-Speech Synthesizer [3] and a publicly available residual-excited LPC diphone synthesizer (*OGIresLPC* [5]) were chosen for this task. To perform voice conversion, the source speech spectrum is mapped on a frame by frame basis while the pitch range is modified to match the target speaker's average pitch. Currently, the LPC residual in each pitch period is left unchanged. Spectral conversion is performed by a locally linear transformation based on Gaussian mixture models (GMMs), whose parameters are calculated by a joint density estimation technique.

In this paper, we evaluate the performance of the new Gaussian mixture conversion model as the training data size and the number of trainable parameters are varied. A normalized mean squared conversion error is used as an objective measure. Perceptual tests were also conducted to assess the subjective differences of the different models and the overall effectiveness of the voice conversion algorithm.

[*]This work was supported by a grant from Intel Corporation.

2. SPECTRAL CONVERSION

Let $x = [x_1 \quad x_2 \quad \cdots \quad x_N]$ be the sequence of spectral vectors characterizing a succession of speech sounds produced by the source speaker and $y = [y_1 \quad y_2 \quad \cdots \quad y_N]$ be spectral vectors describing those same sounds as produced by the target speaker. The goal is to find a conversion function *F* that minimizes the mean squared error

$$\varepsilon_{mse} = E\left[\|y - F(x)\|^2\right], \tag{1}$$

where *E* denotes expectation.

In the literature, the conversion function *F* has been implemented using a variety of techniques, e.g. vector quantization with mapping codebooks [1], dynamic frequency warping [10], and neural networks [6]. Recently, the use of a Gaussian mixture model was proposed to estimate parameters for a piece-wise linear conversion function in a probabilistic framework [8].

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

$$p(x) = \sum_{i=1}^{Q} \alpha_i N(x; \mu_i, \Sigma_i), \quad \sum_{i=1}^{Q} \alpha_i = 1, \quad \alpha_i \geq 0, \tag{2}$$

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 [4].

2.1 GMM with least squares estimation

In a previous approach [8], the parameters (α, μ, Σ) of a GMM are estimated to model the distribution of *x*, the source speaker's spectral space. The conversion function is chosen to be a probabilistic locally linear mapping function

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

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

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

The unknowns (ν, Γ) are computed by solving normal equations for a least squares problem, based on the correspondence between the source and target data [9]. The solution of these normal equations requires inversion of a large and sometimes poorly conditioned matrix.

2.2 GMM with joint density estimation

In our approach, the combination of source and target vectors $z = [x^T y^T]^T$ is used to estimate GMM parameters (α, μ, Σ) for the joint density $p(x, y)$. The conversion function that minimizes the mean squared error between converted source and target vectors is the regression

$$F(x) = E[y \mid x] = \int dy \; y \, p(y \mid x) = \sum_{i=1}^{Q} h_i(x) \left[\mu_i^y + \Sigma_i^{yx} \Sigma_i^{xx-1} (x - \mu_i^x) \right], \quad (5)$$

where

$$h_i(x) = \frac{\alpha_i N(x; \mu_i^x, \Sigma_i^{xx})}{\sum_{j=1}^{Q} \alpha_j N(x; \mu_j^x, \Sigma_j^{xx})}, \quad (6)$$

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 estimation (JDE) method makes no assumptions about the target distributions: clustering takes place on observations of both the source and the target vectors. This is in contrast to the least squares estimation (LSE) method above, where clustering is based on the source vectors only. In theory, modeling the joint density rather than the source density should lead to a more judicious allocation of mixtures for the regression problem. However, JDE is computationally more expensive during the EM step than LSE, since the dimensionality of the space to be estimated doubles.

3. VOICE CONVERSION SYSTEM

3.1 Data sets and training

Two male and one female speaker were selected from the OGI diphone databases (1665 diphones, sampled at 16kHz) [5] to perform one male-to-male and one male-to-female conversion. To study the effects of the amount of training data on conversion performance, we automatically constructed training data sets with vectors derived from a varying number of diphones from both the source and the target speaker. Diphones to be included in a data set were chosen as follows: First, a binary split vector quantization (VQ) was performed on all vectors in the entire database of the source speaker. Diphones whose spectral vectors were closest to one or more codewords of the VQ procedure were identified, and all vectors of these diphones were included in the training set. Table 1 contains more information on the data sets. Additionally, one data set contained the entire source and target databases.

set	diphones	vectors	time (s)
1	2	34	0.3
2	4	68	0.6
3	8	123	1.1
4	16	249	2.2
5	32	470	4.5
6	63	935	9.2
7	123	1822	17.5
8	231	3397	30.6
9	409	5980	53.6
10	725	10462	96.5
ALL	1665	23308	197.6

Table 1: Data sets with different amounts of training data for one speaker.

The VQ method is one way to automatically create a subset of vectors (corresponding to diphones) that carry information from the entire speaker space. It is not optimal, since some speech sounds are perhaps more speaker-dependent than others and should be given preference in the inclusion into the data set. However, it produces reasonable results. For example, the list of diphones (using Worldbet symbols) included in set 4 is {*I-j, I-pau, E-v, &-aI, u-A, U-pau, >i-j, p-u, tS->i, tS-D, m-E, f-k, T-z, S->i, S-s, z-p*}.

Source and target vectors from corresponding diphones were aligned using a dynamic time warping algorithm and collected into the variables *x* and *y*, respectively. A GMM was constructed for every data set with the number of mixtures *Q* set to 1, 2, 4, 8, and 16. Models were considered trained when the EM algorithm indicated an average change of less than 10^{-6} in the estimated vectors μ.

3.2 Features

Bark-scaled, 16th order line spectral frequencies (LSFs) were used as spectral features for the following reasons:

1. Errors are localized in frequency: a badly predicted component adversely effects only a portion of the frequency spectrum.
2. LSFs have been shown to possess very good linear interpolation characteristics [7]. This is important because we use a conversion function that linearly combines vectors.
3. LSFs relate well to formant location and bandwidth, which have been shown to be perceptually relevant for speaker identity.
4. Since the training cost function is the mean squared error, a bark scaling weights errors in accordance with the frequency sensitivity of human hearing.

LSF features were recently applied to voice conversion in [2] as well.

3.3 Conversion

To convert an utterance, the speech synthesis signal processing engine is modified as follows: spectral vectors drawn from the source speaker's database are converted using the conversion function *F* with parameters from the trained GMM. The pitch of the source speaker's residual is adjusted to match the target speaker's pitch in average value and variance. The

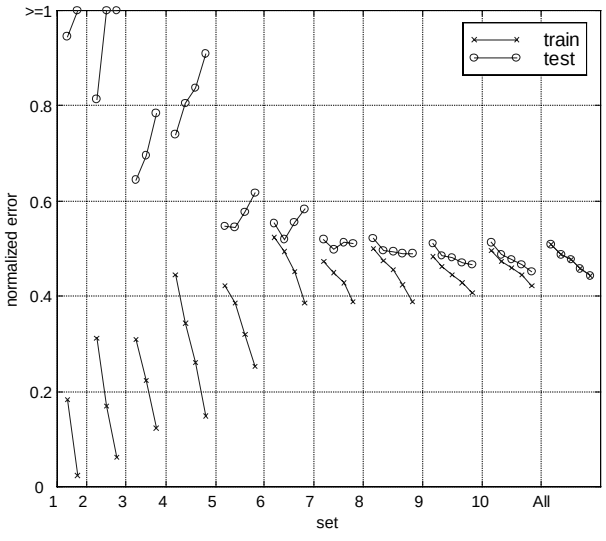

Figure 1: Conversion training and testing errors produced by JDE after training on different data sets. From left to right within one set, the number of mixtures is 1, 2, 4, 8, and 16.

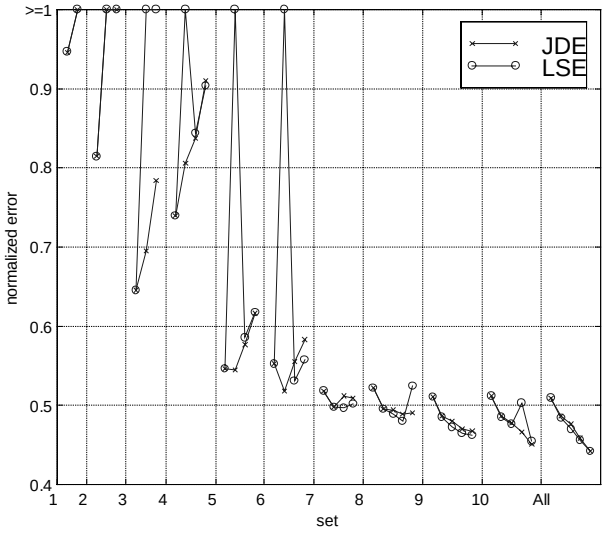

Figure 2: Conversion test errors produced by JDE and LSE after training on different data sets. From left to right within one set, the number of mixtures is 1, 2, 4, 8, and 16.

modified residual and the modified spectral parameters are convolved to render the converted speech.

4. EVALUATION

4.1 Objective Evaluation

To objectively gauge spectral conversion performance, we employ the normalized mean squared error

$$\varepsilon_{norm\ mse} = \frac{\frac{1}{N}\sum_{n=1}^{N}\left\|y_n - F(x_n)\right\|^2}{\frac{1}{N}\sum_{n=1}^{N}\left\|y_n - \mu^y\right\|^2} \quad (7)$$

This error measure yields 1 for a trivial system that always predicts the mean of the target vectors (given normal probabilities).

In this diphone synthesis application, the goal is to predict the entire spectral database after training on all or part of it. Thus, testing is carried out on the set containing all diphones, even though this contains the training set itself.

Figure 1 shows the conversion training and testing errors for the male-to-male conversion. As expected, the training error always decreases with the number of mixtures. The test error increases with use of additional mixtures for sets 1 through 4, and then contains a minimum for sets 5, 6, and 7, due to overfitting of the training data. For larger sets, the test error steadily decreases and is in close proximity to the training error, indicating that the mapping generalized to the rest of the test data. Graphs for the male-to-female conversion are very similar to Figure 1, and not shown here.

In Figure 2 conversion test errors produced by LSE and JDE are directly compared. For the most part, the two estimation methods generate very similar results. However, in several cases, the LSE error is much higher than the JDE error. This is because of problems during optimization which resulted in numerical errors in the conversion function parameters. The JDE method seems to behave more reliably, especially for small training data sets.

4.2 Subjective Evaluation

To subjectively investigate conversion performance, two forced-choice (ABX) experiments and one mean opinion score (MOS) test were carried out. The sentences in the test material were taken from the Harvard sentences database, which contains short, phonetically balanced sentences.

In the first ABX experiment, we presented 16 stimuli **A**, **B**, and **X**, and then asked, "is **X** perceptually closer to **A** or to **B** in terms of speaker identity?" **A** and **B** were speech utterances produced by the speech synthesizer using the source and target speaker databases (the order of assignment was randomized). **X** was the result of taking the source speaker's utterance as input to the conversion system that was trained to convert from the source speaker to the target speaker.

The second ABX experiment was meant to compare conversion performance to a "perfect mapping" that used the target speaker's spectral vectors and the source speaker's pitch-modified residual. This represents a mapping algorithm that attains a mean-squared error of zero. The purpose of this was to measure the success of the spectral conversion, independent of the effects of the residual.

In the MOS experiment we asked subjects to rate the listening quality of 36 speech utterances, using a 5-point scale: 1-bad, 2-poor, 3-fair, 4-good, and 5-excellent. This was done in accordance with [11], which recommends methods for subjective determination of speech quality. Calibration examples were played at the beginning of the test.

In all three experiments, converted speech was produced by models trained by JDE, each having seen differing amounts of

test	set 5	set 7	set 9	set ALL
ABX1 m/m	47.5%	40.0%	37.5%	52.5%
ABX1 m/f	92.5%	95.0%	95.0%	97.5%
ABX2 m/m	87.5%	95.8%	91.7%	95.8%
ABX2 m/f	100%	100%	100%	100%
MOS m/m	3.7	4.0	4.1	4.2
MOS m/f	2.4	2.4	2.1	2.7

Table 2: Results of perceptual tests. The column headers refer to the size of the training data set, the number of mixtures used (*Q*), and the overall number of parameters present in the conversion function (*P*).

training data. The number of mixtures was chosen so that the test error was minimized in each case.

The results of the perceptual tests are shown in Table 2. In the first experiment (ABX1, 20 subjects), listeners were mostly undecided when presented with the male-to-male conversion. This seems to indicate that merely changing the spectrum is not sufficient for changing speaker identity. Interviews with test subjects indicated that many had the impression that a "third" male speaker was created, with similarity to both the source and target speaker. This is in contrast with the male-to-female conversion, which was considered successful by the majority of subjects. This affirms the important role that pitch plays during speaker identification.

In the second experiment (ABX2, 12 out of the 20 subjects in ABX1), listeners associated the converted speech with the "perfectly mapped" target speaker's speech, across all training sizes. This means that if one neglects the effects of the residual, the spectral conversion is perceived as successful. Even if only a small subset of the diphone database is seen by the training algorithm, listeners perceive the same speaker identity shift as if the mapping were perfect.

The MOS test (20 subjects from ABX1) was conducted to assess the effect of training data set size on the perceived quality of the converted signal, apart from speaker identity judgements. As expected, results of the MOS experiment indicated a steady improvement with training size for the male-to-male conversion. In general, the speech quality was judged as "good", in contrast to the male-to-female conversion which was judged below "fair". This can be traced to problems when synthesizing a female voice with an originally male residual.

To hear audio examples of the voice conversion system, please visit the web site at http://cse.ogi.edu/cslu/tts.

5. CONCLUSIONS

This paper introduces a new spectral conversion algorithm using a locally linear transformation based on Gaussian mixture models whose parameters are trained by joint density estimation. Numerically, it was found to perform roughly equivalent to a previous GMM-based approach, but was more robust for small amounts of training data. Perceptual tests confirm that the spectral conversion was perceived as successful, even after training on small training data sets. The tests also confirmed the importance of pitch in speaker identification. However, a technique for converting the residual waveform is necessary to improve overall voice conversion. An improvement of speech listening quality with training size was most observable for the male-to-male conversion.

The presented voice conversion system is capable of producing new target speaker spectra from a single-speaker database. This make it possible to synthesize speech in new voices with little training data and storage.

ACKNOWLEDGMENTS

This work was supported by a grant from Intel Corporation. The authors thank Ron Cole, Paul Hosom, Todd Leen, and all the volunteers of the perceptual tests for their assistance.

REFERENCES

[1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, pp. 655-658, IEEE, April 1988.

[2] L. M. Arslan and David Talkin, "Voice conversion by codebook mapping of line spectral frequencies and excitation spectrum," Proceedings of EUROSPEECH, pp. 1347-1350, September 1997.

[3] 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.

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

[5] 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.

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

[7] K. K. Paliwal, "Interpolation Properties of Linear Prediction Parametric Representations," Proceedings of EUROSPEECH, pp. 1029-32, September 1995.

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

[9] 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.

[10] H. Valbret, E. Moulines, and J. P. Tubach, "Voice transformation using PSOLA techniques," Speech Communication, vol. 11, pp. 175-187, 1992.

[11] International Telecommunication Union, "ITU-T Recommendation P.800: Methods for subjective determination of transmission quality," August 1996.