Personalizing a Speech Synthesizer by Voice Adaptation

Alexander Kain (kain@cse.ogi.edu)
Mike 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

A voice adaptation system enables users to quickly create new voices for a text-to-speech system, allowing for the personalization of the synthesis output. The system adapts to the pitch and spectrum of the target speaker, using a probabilistic, locally linear conversion function based on a Gaussian Mixture Model. Numerical and perceptual evaluations reveal insights into the correlation between adaptation quality and the amount of training data, the number of free parameters. A new joint density estimation algorithm is compared to a previous approach. Numerical errors are studied on the basis of broad phonetic categories. A data augmentation method for training data with incomplete phonetic coverage is investigated and found to maintain high speech quality while partially adapting to the target voice.

1. INTRODUCTION

Voice conversion systems intend to alter a source speaker's speech so that it is perceived to be spoken by another target speaker. Integrating voice conversion technologies into a concatenative text-to-speech (TTS) synthesizer makes it possible to produce additional voices from a single source-speaker database quickly and automatically. The process of "personalizing" a synthesizer to speak with any desired voice is referred to as "voice adaptation".

Generally, creating a new voice for a TTS system is a tedious process, requiring hours of speech recorded in a studio followed by more or less automatic processing, resulting in large databases. A voice adaptation system enables the ordinary user to create a new voice with standard computer equipment within minutes, requiring a fraction of the storage of a speech database.

Possible applications of voice adaptation are numerous; for example, email can be synthesized in the sender's voice or information systems with dynamic prompts can have distinct voice identities.

The voice adaptation system is implemented as part of the OGIresLPC module [5] within the Festival text-to-speech synthesis system [1] which is distributed via the CSLU Toolkit [9]. Section 2 describes the voice adaptation system in more detail.

Section 3 addresses the question of how the amount of training data and the number of free parameters of the mapping correlate with voice adaptation performance. It also compares a new joint density training algorithm with a previously published approach.

Section 4 goes into further detail by studying prediction errors on the basis of phonetic classes. In addition, a new data augmentation method is investigated that promises to allow for an iterative improvement of the voice adaptation system, while maintaining high speech intelligibility at all times. Using this method, a small training set with highly incomplete phonetic coverage is sufficient to begin adapting the voice identity of the system towards the target speaker, while adding more target speech improves the adaptation quality.

Finally, we conclude with a summary of our findings and point to future directions.

2. VOICE ADAPTATION SYSTEM

2.1 Speech Material and Alignment

Upon establishment of a speech corpus to be spoken by the target speaker, her or his speech is recorded at 16kHz/16bit to match the TTS engine's output audio format. Data for the source speaker are generated by the synthesizer using a source-speaker database. Next, phonetic labeling is accomplished with the force-alignment package in the CSLU Toolkit. For the purposes of this work, phoneme boundaries were checked by hand and adjusted when necessary to assure maximum accuracy. Because speech segments between source and target speakers are of different lengths, features for the shorter segment are linearly interpolated to match the longer segment in the number of vectors. Finally, features for selected phonemes are collected in sequence to assure the same original phonetic context and stored as labeled source/target pairs.

2.2 Features

The current system changes only segmental properties, specifically spectral envelope and average pitch. Bark-scaled line spectral frequencies (LSF) were selected as spectral features because of the following properties:

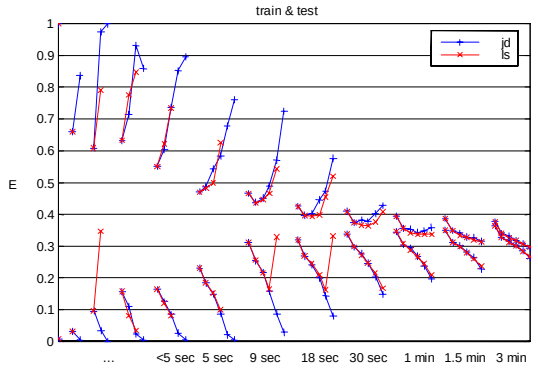

Figure 1: Training (lower curves) and test (upper curves) errors for the joint density (JD) and least-squares (LS) estimation method. From left to right within one set, the number of mixture components is 1, 2, 4, 8, and 16.

- A badly predicted vector component affects only a portion of the frequency spectrum adversely.
- LSFs have good linear interpolation characteristics [6], which is essential for a locally linear conversion function as used by the voice adaptation system.
- LSFs relate well to formant location and bandwidth, which are perceptually relevant for speaker identity.
- Because the training cost function minimizes the mean squared error a bark scaling weights prediction errors in accordance with the frequency sensitivity of human hearing, which is more sensitive to frequency changes at lower frequencies.

2.3 Training

Given a sequence of aligned source and target feature vectors x and y we want to compute a conversion function F that minimizes the mean squared error

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

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

$$F(x) = \sum_{i=1}^{Q} h_i(x)\left[\mu_i^y + \Sigma_i^{yx}(\Sigma_i^{xx})^{-1}\left(x - \mu_i^x\right)\right], \tag{2}$$

where (α, μ, Σ) are Gaussian mixture model (GMM) parameters estimated with Q mixture components on the joint probability density $p(x, y)$ by the estimation maximization (EM) algorithm and $h_i(x)$ is the posterior probability that the i^{th} Gaussian component generated x [2,4,7]. The generated conversion function is stored as a new voice in the TTS system.

2.4 Conversion

The conversion function F is applied to spectral vectors drawn from the source speaker's database during synthesis to yield predicted target vectors. The pitch of the source speaker's residual is adjusted to match the target speaker's pitch in average value and variance. Fine details of the LPC residual are left unchanged.

3. EXPERIMENT 1

3.1 Motivation and Setup

In this experiment, the correlation between the amount of training data and the number of parameters with respect to the final quality is investigated. At the same time, the joint density (JD) training algorithm is compared to a previous least-squares (LS) solution training approach [7]. The final system output is evaluated numerically and perceptually.

The speech material for source and target speakers was selected from two diphone databases. Training data sets of increasing sizes were automatically constructed by a customized vector quantization method carried out on the entire database. A GMM was constructed for every data set with the number of mixtures Q set to 1, 2, 4, 8, and 16. A random 20% of all available data are held out for each of 3 rotations to construct independent test sets.

3.2 Objective Evaluation

Prediction errors are measured by the normalized mean squared error

$$\varepsilon_{norm\,mse} = \frac{\frac{1}{N}\sum_{n=1}^{N}\left\|y_n - F(x_n)\right\|^2}{\frac{1}{N}\sum_{n=1}^{N}\left\|y_n - \mu^y\right\|^2}. \tag{3}$$

Given normal probabilities, this error measure yields 1 for a trivial system that always predicts the mean of the target vectors. The presented errors are averages over all rotations.

Figure 1 displays training and test errors produced after training on data sets equivalent to the length of speech indicated on the bottom of the graph. It can be observed that the training error always decreases with the number of mixture components. The test error increases with the use of additional mixture components 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

Test	5 sec	18 sec	53 sec	198 sec
ABX m/m	47.5%	40.0%	37.5%	52.5%
ABX m/f	92.5%	95.0%	95.0%	97.5%
MOS m/m	3.7	4.0	4.1	4.2
MOS m/f	2.4	2.4	2.1	2.7

Table 1: Results of perceptual tests.

close proximity to the training error, indicating a generalization of the mapping to the test data.

For the most part, the two estimation methods yield comparable results. However, in several cases, the LS error is significantly higher than the JD error or not displayed. This is due to problems during optimization which resulted in numerical errors in the conversion function parameters. The JD method behaves more reliably, especially for small training data sets.

Additionally, there are implementation differences between the JD and LS estimation procedures: During the EM step, JD is computationally more expensive than LS because the dimensionality of the space to be estimated is doubled. However, LS requires an extra solution step. In addition, the largest matrix in the solution step of LS is several times larger than the total storage requirement for JD. Finally, LS necessitates approximately twice the number of overall operations as JD during training.

GMMs with diagonal covariances were also investigated [3]. In this configuration, the conversion function approximates frequency warping. Even though LSFs are not statistically independent, this method achieves accuracy comparable to the JD method.

3.3 Subjective Evaluation

To investigate conversion performance perceptually, a forced-choice (ABX) experiment and a mean opinion score (MOS) test were carried out. Short, phonetically balanced sentences were taken from the Harvard sentences database as test material.

In the 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.

In the MOS experiment, carried out in accordance with [8], 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. Calibration examples were provided at the beginning of the test.

In both experiments, converted speech was produced by models trained on different amounts of training data. The number of mixtures was chosen so that the test error was minimized in each case. The results of the perceptual tests on 20 subjects are shown in Table 1.

In the ABX experiment, listeners were mostly undecided when presented with the male-to-male conversion. Subsequent interviews revealed that many listeners had the impression that a "third" male speaker was created, with similarities 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 confirms the important role that pitch plays during speaker identification.

Results of the MOS experiment show a steady improvement with the increase of the amount of training data for the male-to-male conversion. Generally, the speech quality was judged as "good", in contrast to the male-to-female conversion which was judged below "fair". This can be attributed to problems when synthesizing a female voice with a residual derived from a male voice.

Another ABX experiment was conducted in which **X** was a "perfect mapping" which was constructed using the target speaker's original spectral vectors and the source speaker's pitch-modified residual, representing a flawless mapping algorithm. The purpose of this was to measure the success of spectral conversion, independent of the effects of the LPC residual. The outcome of this experiment showed that at least 87% of listeners identified the converted utterance as closer to **X**, indicating a successful spectral conversion by the mapping. Comparing this result to the first ABX experiment suggests that merely changing the spectrum is not sufficient for changing speaker identity, especially when the pitch is not modified.

4. EXPERIMENT 2

4.1 Motivation and Setup

There are cases when the amount of training data for a target speaker is so small that certain parts of the acoustic space are not covered. Normally, this would lead to diminished speech quality in the resulting system output. However, if the input data is labeled by classes, then a data augmentation (DA) method can substitute the source vectors themselves instead of target vectors for unseen classes. The hypothesis is that

instead of relying on mappings that contain a few "global" transformations it is better to localize learned transformations and not transform the rest of the speaker space. This is based on the assumption that, to some degree, the source and target speakers are similar. This also satisfies the goal that at all times speech intelligibility should be maintained as high as possible while voice identity is adapted in the regions of space where training data are available.

For this experiment we developed a speech corpus that contains 32 English words which cover all phonemes (vowels once, all others twice). A female voice was recorded for a male to female conversion and four different training data sets were constructed:

Set	Phonetic content	Vectors
1	only monophthongs	1061
2	only monophthongs and nasals	1469
3	all, except plosives and affricates	3537
4	all	4148

Then, for the DA method, sets 1 through 4 are augmented with data from the source speaker for unseen classes. A random 20% are held out of the corpus for each of 3 rotations to construct independent test sets containing all classes.

Whereas we search for the number of mixture components (1, 2, 4, 8, 16 and 32) yielding the lowest test error for the normal method, we fix the number of mixture components for DA at 32.

4.2 Objective Evaluation

As in the previous section we employ the normalized mean squared error for gauging adaptation performance. When presenting errors, the number of mixture components for the normal method is chosen such that it yields the lowest average test error for any particular training set. Errors shown are averages over all rotations.

Figure 2 shows the results of comparing the normal method and the DA method over all phonetic classes. It it observed that the error for the DA method is significantly lower than for the normal case in sets 1 and 2. As more target data becomes available, the difference between the two methods will decrease, as can be seen in set 3. This suggests that the process of transforming only classes that have been seen during training is successful when the training data set is limited in phonetic scope.

Figures 3 through 6 display the same test errors, but broken down by the broad categories monophthongs (mono), diphthongs (di), plosives and affricates (plos&affr), nasals (nas), fricatives (fric), and approximants (approx).

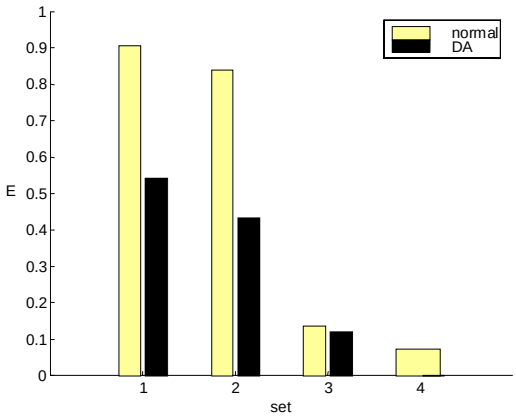

Figure 2: Normal method and DA method compared.

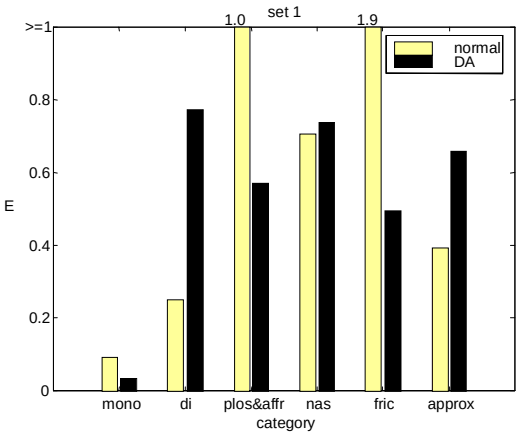

Figure 3: Error distribution for set 1, containing monophthongs.

The distribution of errors of set 1 are seen in Figure 3. Because the training set contains monophthongs, the error on that category is low for both methods as compared to other categories, with the DA method achieving half the error rate of the normal method. This is because when training data is scarce, the normal method performs best with a low number of mixture components. Therefore, one component must cover a wide range of the speaker space, whereas the DA method with its high number of mixture components can allocate components to just one class and learn its specific transformation more accurately.

The DA method also performs significantly better for plosives, affricates and fricatives, but worse for diphthongs and approximants. In the first case we see the strength of DA: Phonemes that are quite different from classes seen during training are, instead of being mapped by inappropriate transformations, left unchanged, thus preserving speech

identity and intelligibility. The second case reveals possible shortcomings of the method: Because diphthongs and approximants are quite similar to monophthongs, as can be seen by relatively low errors of the normal method as compared to other categories, the additional source diphthong and approximant data seem to "confuse" the resulting conversion function. It would be better to augment only data that is sufficiently distant in the speaker space from data that is already encountered.

Figure 4 shows the breakdown for set 2, which contains monophthongs and nasals. The situation is similar to the figure for set 1, except errors for nasals are now low, since they were included in the training set.

The errors for set 3, which contains all the phonemes but plosives and affricates are displayed in Figure 5. In this case, errors for plosives and affricates are high, since they were not included in the training set. For all other categories, the two methods perform similarly, as the difference between them shrinks as the target data covers more and more of the speaker space.

Finally, Figure 6 presents errors for set 4, which contains all phonemes. These errors indicate that plosives, affricates, and fricatives are the most difficult categories to predict. This might be due to their high variance in speech realization; that is, the same speaker will produce a phoneme in this category very distinctly each time.

4.3 Subjective Evaluation

Informal listening tests (2 subjects) were carried out to test the perceptual performance of the normal method versus the DA method. Output created by the system using the DA method yielded clearly higher quality speech than the normal method, which resulted in occasional distortions due to extremely narrow bandwidths and other prediction errors. However, because a male source-speaker database is adapted to a female voice, the DA method produced speech with speaker identity flipping back and forth between source and target for certain phonemes.

CONCLUSIONS

A voice adaptation system enables a user to personalize a TTS system with a new voice quickly and with ease. The system transforms source speaker spectra represented by bark-scaled LSFs by means of a probabilistic, locally linear conversion function based on a GMM. Pitch is adjusted to match the target speaker's pitch in average value and variance.

Numerical and perceptual evaluations show that the voice adaptation system can adapt to a new voice with moderate success after training on approximately one minute of speech.

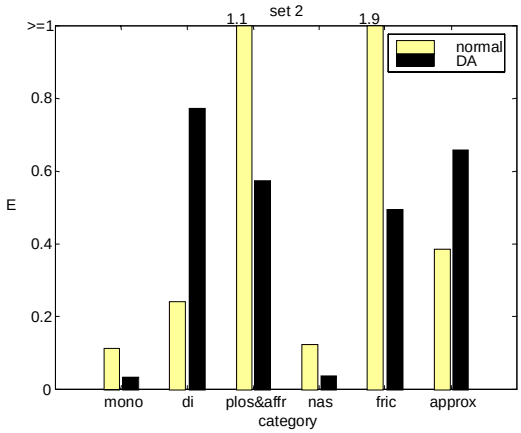

Figure 4: Error distribution for set 2, containing monophthongs and nasals.

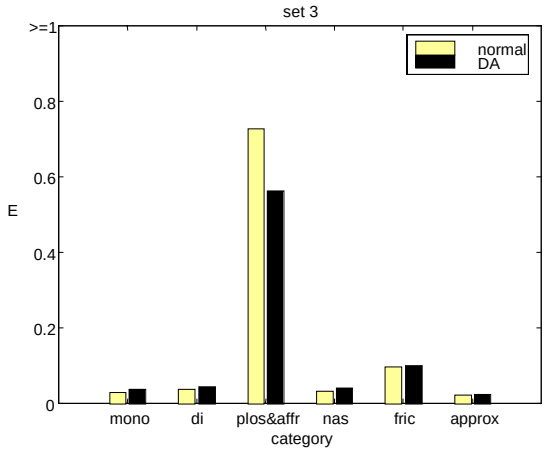

Figure 5: Error distribution for set 3, containing all phonemes except plosives and affricates.

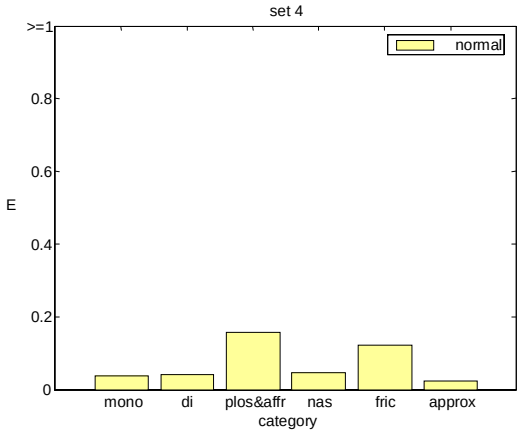

Figure 6: Error distribution for set 4, containing all phonemes.

However, voice identity and speech intelligibility improve as the amount of training data is increased.

When the training data is limited to a few phonetic classes, a data augmentation method can help in maintaining high speech intelligibility while partially adapting to the target speaker. Although this method works well, especially for very small amounts of training data, we are currently investigating methods that will use available data even more efficiently.

Many aspects of the voice adaptation system are still under development and can benefit from further investigation. Most importantly, the development of a technique for converting the LPC residual waveform is necessary to improve overall voice conversion, because fine details of the glottal pulse, as reflected in the residual, contribute to speaker identity.

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

ACKNOWLEDGEMENTS

This work was supported by a grant from Intel Corporation. The authors thank Cathy Heslin, 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. A. Kain and M. Macon, “Text-to-Speech voice adaptation from sparse training data,” to appear in Proceedings of ICSLP, November 1998.

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. K. K. Paliwal, "Interpolation Properties of Linear Prediction Parametric Representations," Proceedings of EUROSPEECH, pp. 1029-32, 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.

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

9. The CSLU Toolkit is distributed free-of-charge to non-profit researchers at http://cse.ogi.edu/cslu or commercially by Fluent Speech Technologies at http://fluent-speech.com.