Hybrid Nearest-Neighbor/Cluster Adaptive Training for Rapid Speaker Adaptation in Statistical Speech Synthesis Systems

Amir Mohammadi and Cenk Demiroglu *
Ozyegin University, Istanbul, Turkey
amir.mohammadi@ozu.edu.tr, cenk.demiroglu@ozyegin.edu.tr

Abstract

Statistical speech synthesis (SSS) approach has become one of the most popular methods in the speech synthesis field. An advantage of the SSS approach is the ability to adapt to a target speaker with a couple of minutes of adaptation data. However, many applications, especially in consumer electronics, require adaptation with only a few seconds of data which can be done using eigenvoice adaptation techniques. Although such techniques work well in speech recognition, they are also known to generate perceptual artifacts in statistical speech synthesis. Here, we propose two methods to both alleviate those quality problems and improve the speaker similarity obtained with the baseline eigenvoice adaptation algorithm. Our first method is based on using a Bayesian approach for constraining the eigenvoice adaptation algorithm to move in realistic directions in the speaker space. Our second method is based on finding a reference speaker that is close to the target speaker, and using that reference speaker as the seed model in a second eigenvoice adaptation step. Both techniques performed significantly better than the baseline eigenvoice method in the subjective quality and similarity tests.

Index Terms: statistical speech synthesis, speaker adaptation, cluster adaptive training, eigenvoice adaptation

1. Introduction

The statistical speech synthesis (SSS) approach has become one of the most popular methods in the speech synthesis field [1]. One of the advantages of the SSS systems is the ability to adapt to a new speaker’s voice with a couple of minutes of data. Thousands of voices have been generated with SSS using speech databases prepared for speaker-independent speech recognition systems [2].

Although linear-regression based speaker adaptation in the SSS systems have been shown to be successful with only a couple of adaptation utterances [3], the issue of adaptation with only a few seconds of data have not been investigated as much. High performance speaker adaptation with such minimal data is very useful and sometimes critical in embedded applications of SSS where there may not be enough resources to store the utterances and/or users are not willing to train the system even with a couple of utterances. A short discussion of the prior work on adaptation with minimal data and our contributions is given below.

Eigenvoice techniques have been shown to be promising for rapid adaptation [4] and can be performed using the Cluster Adaptive Training (CAT) [5] technique. CAT approach has been investigated for SSS [5] [6] [7]. The eigenvoices generated with the CAT algorithm can capture the most important and most common variations in speech. However, the characteristics of the target speaker’s voice that are not captured by the eigenvoices are also important in speaker similarity after adaptation. Moreover, synthesized speech is known to have perceptual artifacts after adaptation using eigenvoice approach with minimal data [4].

Interpolation between different speakers and styles have been used to generate voices with the desired voice quality, style, and emotion [8] [9]. In that approach, weights of pretrained voices/styles are adjusted to make the interpolated voice sound close to target. This is similar to the eigenvoice approach except adaptation is done by interpolating the speaker-adapted voices, as opposed to eigenvoices, and weights are set manually.

We propose two methods to both alleviate the speech quality problems of the eigenvoice technique and improve the speaker similarity after adaptation with minimal data. Cluster adaptive training (CAT) is used as the baseline eigenvoice approach. The first proposed method is based on using a Bayesian approach to estimate the weights of the CAT algorithm with the goal of forcing the adaptation algorithm to move in realistic directions in the speaker space. Pretraining many speakers and using their models to define a proper prior distribution in weight estimation allows the system to create models with significantly less artifacts.

The second proposed method is based on finding a nearest-neighbor (NN) from a set of pretrained models after the first adaptation step. The NN model is then used as the seed model in a second eigenvoice adaptation step where the CAT model parameters are set depending on the selected NN. This approach allows the adaptation algorithm to be more flexible in diverging from the speaker-independent model and achieve higher performance.

Subjective experiment results show that both of the algorithms outperform the baseline CAT adaptation algorithm in speaker similarity. Moreover, synthesized speech after the proposed adaptation techniques is found to have significantly higher speech quality compared to the baseline system in the AB quality tests.

2. Eigenvoice Adaptation

Eigenvoice approach has been used for rapid adaptation in speech recognition and SSS [10]. The idea is to find a set of \( N \) vectors in the high-dimensional space \( \mathbb{R}^n \) (\( n \gg N \)) that can be used to approximate a set of vectors in \( \mathbb{R}^n \). One way to accomplish this is using principal components analysis (PCA) that finds the directions in \( \mathbb{R}^n \) where the data has the highest variance and the \( L_2 \) norm of the approximation error is minimum after projection. Solution with PCA are the eigenvectors of the sample covariance matrix with the highest eigenvalues.

* This work is supported by TUBITAK and European Union FP-7 IRG grants.
In the context of SSS, each eigenvector is called an eigen-

tion. The supervector for speaker $s$ can be created by $\mu_s = [\mu_1, \mu_2, \ldots, \mu_{N_{d}}]$ where $N_{d}$ is the total number of states in all

decision trees in the acoustic model. One can construct a dif-

terence between $\mu_{\text{SI}}$ and the weights for a given speaker are the interpolation factors

\[
E=E_{c}\sqrt{N_{c}}S_{c}+w_{c}^{T}S_{c}w_{c}^{-1}
\]

where $S_{c}=\sum_{i=1}^{N_{c}}x_{c}^{i}$.

In the proposed Bayesian CAT (BCAT) approach, the ob-

jective function becomes

\[
\hat{w} = \arg \max_w p(\chi_{a}|w)p(w)
\]

where $p(w)$ is the prior distribution and set to $N(0, \Sigma_{w})$ here. Removing the term that is independent of $w$ in Eq 3 from the

BCAT objective function becomes

\[
\hat{w} = \arg \max_{w} K_{1,exp}(-\frac{1}{2} \sum_{c=1}^{N_{c}} x_{c}^{i}(w-E_{c}^{T}\Sigma_{c}^{-1}x_{c}^{i})) -

2w^{T}E_{c}^{T}\Sigma_{c}^{-1}S_{x,c} + N_{c}w^{T}E_{c}^{T}\Sigma_{c}^{-1}E_{c}w + w^{T} \Sigma_{w}^{-1}w)
\]

Using some matrix manipulation

\[
\hat{w} = \arg \max_{w} K_{1,exp}(w^{T}E^{T}\Sigma^{-1}S_{w} -

\frac{1}{2}w^{T}E^{T}N\Sigma^{-1}Ew + w^{T} \Sigma_{w}^{-1}w)
\]

where the block diagonal $\Sigma^{-1} =

diag(\Sigma_{1}^{-1}, \Sigma_{2}^{-1}, \ldots, \Sigma_{N_{c}}^{-1}), S_{w} = [S_{x,1}, S_{x,2}, \ldots, S_{x,N_{c}}]$ and $N = diag(N_{1}, N_{2}, \ldots, N_{N_{c}})$.

Because Gaussian distribution is the conjugate prior of the

likelihood function in Eq 2, posterior distribution is also a

Gaussian. Therefore, by completing the squares, Eq 6 can be

written as

\[
\hat{w} = \arg \max_{w} K_{1,exp}(-\frac{1}{2}(w - \mu_{w|x})^{T}N\Sigma^{-1}(w - \mu_{w|x}))
\]

where $\Sigma_{w|x} = (E^{T}N\Sigma^{-1}Ew + \Sigma_{w}^{-1})$ and $\mu_{w|x} =

\Sigma_{w|x}E^{T}N\Sigma^{-1}S_{w}$. BCAT estimate of $w$ is the mean $\mu_{w|x}$

of the posterior distribution.

\[
\Sigma_{w}^{-1}
\]

is estimated from the data as follows. $w$ is estimated

for 120 speakers using 150 utterances per speaker. Because a

data large number of utterances are used, there is no significant
difference between ML or Bayesian estimation of weights with

CAT. $\Sigma_{w}^{-1}$ is then calculated by calculating the sample covari-

ance matrix of the weights and setting the off-diagonal elements
to 0.

3. Hybrid Nearest-Neighbor/CAT

Approach

Bayesian CAT (BCAT) constrains the estimator to move in real-
istic directions supported by empirical evidence, models of the

pretrained reference speakers, using a prior distribution. How-

ever, prior distribution also makes it more difficult to signif-

icantly move away from the SI model especially when mini-

mal adaptation data is available. Therefore, there is need for

algorithms that will allow the BCAT adaptation algorithm to be

more flexible in terms of ability to generate adapted models with

significant divergence from the SI model and producing

voices that are perceptually closer to the target speakers. Here,

we propose using a nearest-neighbor (NN) approach to increase

the flexibility of the hybrid algorithm.

In a database of many reference speakers, it is usually possi-

ble to find a reference speaker that is closer to the target speaker

than the SI model. Therefore, if such a reference speaker is

available, it can be used as the seed model for the BCAT algo-

rithm as opposed to the SI model. The reference model can be

arbitrarily far away from the SI model which makes the hybrid

algorithm more flexible compared to the BCAT algorithm.

The two step hybrid adaptation approach is shown in Fig. 2.

The BCAT algorithm is first used to generate a model, $m_{\text{BCAT}}$. 

Figure 1: Most significant eigenvalues/eigenvectors of the sample
covariance matrix of the mean supervectors belonging to

120 reference speakers.

In the context of SSS, each eigenvector is called an eigen-

value. The supervector for speaker $s$ can be created by $\mu_s = [\mu_1, \mu_2, \ldots, \mu_{N_{d}}]$ where $N_{d}$ is the total number of states in all
decision trees in the acoustic model. One can construct a differ-
tence between $\mu_{\text{SI}}$ and the weights for a given speaker are the interpolation factors
that is used to select the NN model using a distance measure. Then, the NN model is used as the seed model for another BCAT adaptation step (BCAT-NN) to generate the final model.

In addition to using the NN model as the seed, the BCAT-NN adaptation has two more significant differences from the first BCAT. The first one is that a different linear model is used depending on the selected NN. The new model becomes

$$\mu' = \mu_{nn} + E_{nn}w_{nn} + \epsilon_{nn}$$

where $\mu_{nn}$ is the mean supervector of the NN model, $w_{nn}$ is the weight vector computed in the BCAT-NN step, $\epsilon_{nn}$ is the error, $E_{nn}$ is the $E$ matrix trained with the CAT training algorithm using $NN$ as the seed model and training data from all reference speakers. Thus, depending on which NN is found to be closest to the target speaker, a different linear model is used in the second BCAT step. Model parameters for each NN are trained offline.

The second difference is that, as opposed to using $\Sigma_{w}$ in the prior distribution, $\sigma^2I$ is used. The goal in this approach is to take into account the information that the target model is expected to be close to the NN model. $\Sigma_{w}$ considers all available reference speakers while we only need to consider the speakers that are close to NN in BCAT-NN. Because reference speaker database is not large enough to estimate all diagonal elements separately, the covariance matrix is assumed to be isotropic and requires estimation of only one parameter. The parameter $\sigma^2$ is estimated by minimizing the root-mean-square error (RMSE) of parameters generated with the target and adapted models. Adaptation is done for all reference speakers by removing each reference speaker from the pool and using it as the target model. Aggregate RMSE of all reference speakers is computed for a given $\sigma^2$. Finally, $\sigma^2$ with the best aggregate RMSE performance is used for all reference speakers. Thus, $\sigma^2$ is not speaker dependent.

4. Experiments

4.1. Experiment Setup

All systems in the experiments were trained with 78 dimensional vectors consisting of 24 Mel-Generalized Cepstrum Coefficients (MGCs), 1 log-energy, 1 log-F0 coefficient, and their delta and delta-delta parameters. 20 msec analysis window with 5msec frame rate is used for feature extraction. Phonemes are modelled with 5 state HSMMs.

Wall Street Journal (WSJ) database is used to train the average voice and the speaker-adapted voices. Four male speakers with 1250 utterances for each of them are used for training the average voice. For the proposed system, 120 male reference speakers from the WSJ database are trained using 150 utterances per speaker with CSMAPLR adaptation and an additional MAP adaptation. HTS 2.2 training and synthesis tools are used to generate the samples for the baseline systems. Speaker adaptive training (SAT) is used during training the SI model and the reference models. For CAT, rank of the $E$ matrix is set to 5 for 1 sec, 10 for 2 sec, 30 for 3 and 4 sec adaptation data. Rank of the matrix is tuned based on the performance with objective measures. For CAT-based algorithms, rank is set to 30 since overfitting did not occur with the Bayesian approach.

Root-mean-square-error (RMSE) is used for objectively measuring the distance between the MGC and pitch features of synthesized and original speech samples. To make meaningful comparison between them, original speech is first aligned at the state level with the average voice model. Durations obtained during alignment is used for synthesizing the samples. For each target speaker, adaptation is performed for 1, 2, 3, and 4 seconds of adaptation data, excluding any silence segments. 40 utterances are synthesized for each target speaker to measure the RMSE distance.

21 male target speakers are used in the objective tests. For each target, a speaker-dependent (SD) model is generated using CSMAPLR adaptation with an additional MAP step using 150 adaptation utterances per speaker. Those SD models are used as the upper bound in adaptation performance.

ABX tests are used to subjectively measure the similarity of synthesized speech to speaker’s original sample. AB test is done to measure the quality differences. Similar to the RMSE tests, for each target speaker, adaptation is performed with 1, 2, 3, and 4 seconds of data. For each case, one utterance is synthesized per target speaker. 10 target speakers are used. Ten listeners took the tests.

L1 norm, L2 norm, and cosine distance are compared for measuring the supervector distances to pick the NN model. L2 norm is used here since it was found to perform better than others. Delta, delta-delta, and energy features were not used in the distance computations. Variances of the features are not normalized since that degraded the performance.

4.2. Results and Discussion

ML-based CAT, BCAT, and BCAT-NN algorithms are compared objectively and results are shown in Fig. 3 and Fig. 4. CAT-based algorithms outperform the CSMAPLR algorithm for the MGC case. However, for lf0, difference is not significant. RMSE performance of CAT-based systems are not significantly different from each other.

Two sets of subjective speaker similarity tests are done. In the first set, BCAT is compared with the CAT method using the ABX speaker similarity test where the speaker is asked to prefer the sample A, sample B or none depending on the similarity to the original sample X. Results are shown in Fig. 5. BCAT significantly outperforms CAT for one and two second cases. Gap between the two systems decreases with increasing amount of data which is expected since the effect of prior distribution decreases with more adaptation data.

In the second set, BCAT-NN is compared with BCAT. Results are shown in Fig. 6. In this case, there is modest and consistent improvement with BCAT-NN over BCAT for all adaptation data sizes.

Alleviating the perceptual artifacts observed with ML-based CAT is another goal of this work. To measure the improvement in quality, AB preference test is used where the listener is asked to prefer between A, B or none based on the naturalness of speech. Results are shown in Fig. 7. The BCAT-NN
MOS tests are performed to measure the quality of CAT and BCAT methods. CAT method has a MOS score of 2.7 with a $(+/- 0.08)$ 95% confidence interval. BCAT method has a MOS score of 3 with a $(+/- 0.05)$ 95% confidence interval. These scores are lower than the speaker-dependent MOS score which is 3.3. Lower MOS scores are thought to be a result of adapting with minimal data which is not enough for reducing the variances in the acoustic model.

Objective test results show that on average CAT-based systems all have the same performance measured with RMSE. However, in listening tests, especially in the quality tests, listeners had significant preference for the proposed systems. BCAT algorithm can use more eigenvectors than CAT without overfitting. In addition, BCAT-NN algorithm starts adapting from a better seed model than BCAT. Although those advantages do not seem to make a significant difference in RMSE, perhaps because of the minimal adaptation data used, improvements are clearly audible by the listeners.

5. Conclusion
In this work, we propose a hybrid Bayesian CAT and nearest-neighbor approach to rapid speaker adaptation for SSS. Significant improvements in speaker similarity and quality are achieved in the listening tests with BCAT compared to the baseline CAT approach. Moreover, the BCAT-NN approach is shown to have higher speaker similarity performance compared to BCAT in the listening tests. BCAT algorithm successfully eliminated the perceptual artifacts that were observed with CAT. BCAT-NN could alleviate the slow convergence problem of BCAT through finding a nearest-neighbor among the pretrained speakers and using it as the seed model.
6. References


