16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

EFFICIENT MODEL RE-ESTIMATION IN VOICE CONVERSION

Jilei Tian[1], Victor Popa[1] and Jani Nurminen[2]

[1]Media Laboratory, Nokia Research Center, Finland
[2]Nokia Devices R&D, Finland
phone: + (358) 0504835396, fax: + (358) 718035395
email: {jilei.tian, ext-victor.popa, jani.k.nurminen}@.nokia.com

ABSTRACT

Voice conversion systems aim at converting an utterance spoken by one speaker to sound as speech uttered by a second speaker. Over the last few years, the interest towards voice conversion has risen immensely. Gaussian mixture model (GMM) based techniques have been found to be efficient in the transformation of features represented as scalars or vectors. However, reasonably large amount of aligned training data is needed to achieve good results. To solve this problem, this paper presents an efficient model re-estimation scheme. The proposed technique is based on adjusting an existing well-trained conversion model for a new target speaker with only a very small amount of training data. The experimental results provided in the paper demonstrate the efficiency of the re-estimation approach in line spectral frequency conversion and show that the proposed approach can reach good performance while using only a very limited amount of adaptation data

1. INTRODUCTION

In voice conversion, the aim is to convert speech from one speaker to sound as if it was spoken by another speaker without changing the meaning or the content of speech. Usually, a set of training material is recorded from the source and the target speakers, and one or more conversion models are trained. Alternatively the source speaker may also be a TTS voice. The research on the topic of voice conversion has received an increased amount of interest during the recent years. One of the reasons is the attractive idea to use intra-lingual voice conversion in cost-effective individualization of text-to-speech (TTS) systems. In addition, it is becoming realistic to apply input speaker's voice characteristics to the output speech in the speech-to-speech translation using cross-lingual voice conversion. Without voice conversion, new voices have to be created in a time-consuming and expensive way using extensive recordings and manual annotations.

Many voice conversion approaches have been proposed in the literature, and the results have been quite promising. From the technical point of view, typical approaches presented in the literature include Gaussian mixture modeling (GMM) based conversion [1][4], neural network based conversion [7][12], hidden Markov model (HMM) based conversion [5], linear transformation based conversion [10][14], and codebook based conversion [2]. Research results found in the literature have shown that the GMM based approach can be used successfully in voice conversion. However, the good quality of the trained GMM requires a certain number of aligned training data which may prevent wide use of the technique in practical applications where the end users would have to spend quite some time recording their speech for the training data.

GMM models in the voice conversion task are commonly trained from scratch using a relatively large amount of aligned training data. The training data can be either parallel, meaning both the source and target speakers read the same text, or unparallel. Using a fairly large amount of data improves the quality of the models, but this approach has several inherent drawbacks from different aspects:

- Usability: there is a need to record plenty of training data;
- Memory: requirement of having more memory available for storing the training data;
- Complexity: the computational load caused by the GMM training with large training data may be rather high.

Related adaptation techniques for GMM conversion with reduced data are described in [3][6] and [11]. Maximum A Posteriori (MAP) based adaptation [3][6] trains a GMM on a large source corpus and limited amount of speech from the target speaker to estimate a joint GMM robustly. The Eigenvoice Conversion [11] uses multiple parallel sets of the same source speaker but several different target speakers to build a so called EV-GMM. Basically a GMM distribution is represented as a function of a weight vector. We can derive conversion functions to any target speaker using maximum likelihood techniques to estimate the weight vector from reduced (and unparallel) target data.

The re-estimation approach proposed in this paper is applicable in a voice conversion framework where the source and target acoustic spaces are jointly modelled as a GMM. It introduces a very efficient scheme for adapting a well-trained GMM conversion model to a completely new target speaker with only limited speech data from the new target speaker. It is readily suitable for embedded implementations and it does not require a large amount of training data or data from many speakers. The proposed approach broadens the variety of potential use cases for voice conversion especially in practical embedded applications.

The paper is organized as follows. First, a general overview of the voice conversion system is briefly given in Section 2.

In Section 3, the proposed efficient re-estimation approach for GMM based voice conversion is introduced while Section 4 shows the experimental results we have achieved in objective measurements and in subjective listening tests. Section 5 discusses different aspects of the proposed approach and draws the conclusions.

2. VOICE CONVERSION OVERVIEW

2.1. Feature extraction

In our voice conversion system first introduced in [8], the source and filter parameterisation of speech relies on linear prediction (LP) to estimate a model of the vocal tract that can be represented using line spectral frequencies (LSFs). The excitation is parameterised by other parameters such as pitch, voicing, residual spectral amplitudes. In addition to the separation into the vocal tract model and the excitation model, the overall gain or energy is used as a separate parameter to simplify the processing of spectral information. Such a representation has favorable properties from the viewpoint of both speech coding and voice conversion.

2.2. Voice conversion

The training of a conventional GMM based voice conversion system requires a parallel corpus from source and target speakers. An alignment algorithm is used to align the source features to their counterparts in the target sequence to obtain the feature pair sequence $\mathbf{v}=[\mathbf{v}_1\ \mathbf{v}_2\ \ldots\ \mathbf{v}_t\ \ldots\ \mathbf{v}_w]$, where $\mathbf{v}_k=[\mathbf{x}_p^{\mathrm{T}}\ \ \mathbf{y}_q^{\mathrm{T}}]^{\mathrm{T}}$ while $\mathbf{x}_p$ and $\mathbf{y}_q$ denote aligned vectors corresponding to times p and q at source and target sides, respectively. The distribution of $\mathbf{v}$ is modeled by GMM as given by Equation (1).

$$P(\mathbf{v}) = P(\mathbf{x},\mathbf{y}) = \sum_{l=1}^{L} c_l \cdot N(\mathbf{v},\boldsymbol{\mu}_l,\boldsymbol{\Sigma}_l), \qquad (1)$$

where c_l denotes the prior probability of $\mathbf{v}$ for the component l. $N(\mathbf{v},\ \boldsymbol{\mu}_l,\ \boldsymbol{\Sigma}_l)$ denotes a Gaussian distribution with the mean vector $\boldsymbol{\mu}_l$ and the covariance matrix $\boldsymbol{\Sigma}_l$. The parameters of the GMM can be estimated using the Expectation Maximization (EM) algorithm [9][13]. The conversion function that converts the source feature $\mathbf{x}$ to the target feature $\mathbf{y}$ is given by Equation (2).

$$F(\mathbf{x}) = E(\mathbf{y} \mid \mathbf{x}) = \sum_{l=1}^{L} p_l(\mathbf{x}) \cdot \left(\boldsymbol{\mu}_l^{\mathbf{y}} + \boldsymbol{\Sigma}_l^{\mathbf{yx}} \left(\boldsymbol{\Sigma}_l^{\mathbf{xx}}\right)^{-1} \left(\mathbf{x} - \boldsymbol{\mu}_l^{\mathbf{x}}\right)\right), \qquad (2)$$

where $p_l(\mathbf{x})$ is the posterior probability that the source feature vector $\mathbf{x}$ belongs to the l-th mixture component. $\boldsymbol{\mu}_l^x$ and $\boldsymbol{\mu}_l^y$ are the source and target parts of the mean vector $\boldsymbol{\mu}_l$ and $\boldsymbol{\Sigma}_l^{yx}$ and $\boldsymbol{\Sigma}_l^{xx}$ are the target-to-source and source-to-source blocks of $\boldsymbol{\Sigma}_l$ as defined in (3).

$$\Sigma_l = \begin{bmatrix} \Sigma_l^{xx} & \Sigma_l^{yx} \\ \Sigma_l^{yx} & \Sigma_l^{yy} \end{bmatrix}, \quad \mu_l = \begin{bmatrix} \mu_l^x \\ \mu_l^y \end{bmatrix} \qquad (3)$$

3. EFFICIENT GMM RE-ESTIMATION

In the GMM based transformation, multiple mixtures of Gaussian distributions are trained using joint feature vectors combined from the feature vectors estimated from the source and target sides. The main idea of the proposed re-estimation approach is to utilize an existing well-trained GMM model from the source speaker X to the target speaker Y (trained using an adequate amount of speech data), and to adapt it to be a GMM model from the source speaker X to a new target speaker Z with only a very limited amount of training data.

The proposed technique is very fast in adapting the voice conversion model to the new source and target speaker pair. It does not require much training data as the parameter estimation is done directly on the well-trained model. One possible use case is individualization of the text-to-speech functionality. With only a small amount of training data, TTS can start using any new voice provided by the user. The performance of the proposed approach is demonstrated using experimental results in Section 4.

3.1. GMM training for source and target speakers: *X, Y*

First, let us assume that we have an adequate amount of training data from the source speaker X and the target speaker Y. In the case of the TTS application, it is rather easy to collect this speech data since a lot of speech is automatically available in the databases of the speech synthesis system. Also in other applications, the recordings of the two voices (X and Y) can be done quite easily e.g. by the developer of the application. After we have the data available, we can train a model that converts speech from the speaker X to sound like the speaker Y just like in the case of conventional voice conversion.

The GMM for the random variable $\mathbf{v}$ can be estimated from a time sequence of $\mathbf{v}$ samples $[\mathbf{v}_1\ \mathbf{v}_2\ \ldots\ \mathbf{v}_t\ \ldots\ \mathbf{v}_w]$, when $\mathbf{v}_k=[\mathbf{x}_p^{\mathrm{T}}\ \mathbf{y}_q^{\mathrm{T}}]^{\mathrm{T}}$ is a joint variable and $\mathbf{x}_p$, $\mathbf{y}_q$ would denote aligned features from the source X and target Y speaker respectively. The distribution of $\mathbf{v}$ is modeled by GMM as in Equation (1). The weighting terms are chosen to be the conditional probabilities that the feature vector $\mathbf{x}_t$ (at time t) belongs to the different components.

3.2. Re-estimation for source and target speakers: *X, Z*

As mentioned above, we have trained a GMM model, $\boldsymbol{\lambda}$, for source speaker X and target speaker Y, $\{c_l, N(\mathbf{v}, \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)\}$. For the new conversion pair from source speaker X to target speaker Z, given a limited training data of the joint variable $\mathbf{v}_k=[\mathbf{x}_p^{\mathrm{T}}\ \ \mathbf{z}_r^{\mathrm{T}}]^{\mathrm{T}}$, the GMM model can be adapted into $\hat{\boldsymbol{\lambda}}$ $\{\hat{c}_l, N(\hat{\mathbf{v}}, \hat{\boldsymbol{\mu}}_l, \hat{\boldsymbol{\Sigma}}_l)\}$ for mapping between the source and the target speakers X, Z, based on the well-trained model $\boldsymbol{\lambda}$.

Since c_l is the prior information to measure the probability that the training data falls into the cluster or the mixture, for the re-estimated GMM model $\hat{\boldsymbol{\lambda}}$, the new target data does not change the data distribution for the source side. Thus it is reasonable to assume that the clusters have not changed for the source data. The outcome of having new target speaker only causes the cluster shifting along the target space as illustrated in Figure 1.

With the GMM model re-estimation scenario, if we assume that the model can be re-estimated only with the mean, while keeping the prior probability unchanged, we have the model re-estimation algorithm shown in Equation (4-5).

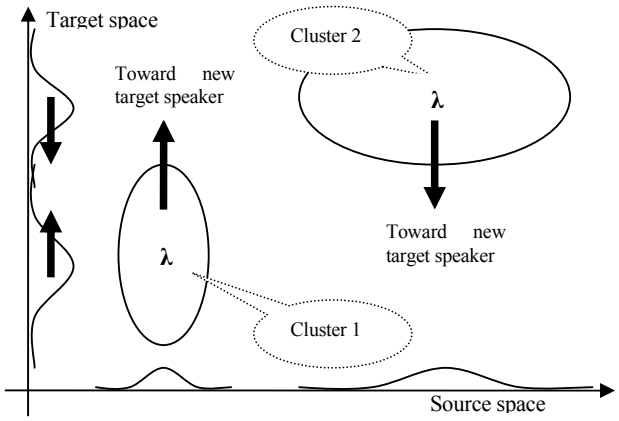

Figure 1. Diagram of GMM model re-estimation scenario.

$$\hat{c}_l = c_l$$
$$\hat{\Sigma}_l = \Sigma_l \tag{4}$$
$$\therefore \hat{p}_l(\mathbf{x}_t) = \frac{\hat{c}_l \cdot N(\mathbf{x}_t, \boldsymbol{\mu}_l^{\mathbf{x}}, \boldsymbol{\Sigma}_l^{\mathbf{xx}})}{\sum_{i=1}^{L} \hat{c}_i \cdot N(\mathbf{x}_t, \boldsymbol{\mu}_i^{\mathbf{x}}, \boldsymbol{\Sigma}_i^{\mathbf{xx}})} = p_l(\mathbf{x}_t)$$

$$\hat{\boldsymbol{\mu}}_l^{z} = \frac{\sum_{m=1}^{M} \hat{p}_l(\mathbf{x}_m) \cdot \mathbf{z}_m}{\sum_{m=1}^{M} \hat{p}_l(\mathbf{x}_m)} \tag{5}$$
$$\hat{\boldsymbol{\mu}}_l = \left[(\boldsymbol{\mu}_l^{x})^{\mathrm{T}} \; (\hat{\boldsymbol{\mu}}_l^{z})^{T} \right]^{T}$$

In Equation (5) z_m represents a feature vector of the new target speaker aligned with source speaker's vector x_m.

Depending on the size of the training data available from the target side (new target *Z*), we can also estimate the covariance matrix as shown in Equation (6).

$$\hat{\Sigma}_l = \frac{\sum_{m=1}^{M} \hat{p}_l(x_m) \cdot \left([x_m, z_m] - [\mu_l^{x}, \hat{\mu}_l^{z}] \right) \cdot \left([x_m, z_m] - [\mu_l^{x}, \hat{\mu}_l^{z}] \right)^{T}}{\sum_{m=1}^{M} \hat{p}_l(x_m)} \tag{6}$$

If the size of the training data is very small (less than roughly 5 utterances or 15-20 seconds of speech), it is reasonable to keep the covariance matrix unchanged due to the following reasons:

- The covariance matrix cannot be reliably estimated with very limited amounts of data;
- The source space has not changed at all;
- The contribution from covariance in voice conversion is very small due to the small weighting factor in $\boldsymbol{\Sigma}_l^{\mathbf{yx}} \left(\boldsymbol{\Sigma}_l^{\mathbf{xx}} \right)^{-1}$.

4. EXPERIMENTAL RESULTS

We carried out several experiments to demonstrate that a well-trained GMM converting voice *X* to voice *Y* can be effectively adapted to a new target *Z* using a very limited amount of parallel *X* to *Z* adaptation data. Both objective and subjective measures were used in the experiments.

10th-order LSF vectors were used in our experiment. Based on our previous experiments, the transformation of LSFs can be handled using a GMM of 8 mixture components and thus the models discussed below use 8-component GMMs. It was observed that reducing the number of components degrades the modelling. In turn, adding components only brings marginal improvement while the complexity increases and the parameter estimation becomes less reliable. Moreover, we had a parallel corpus of speakers *X*, *Y* and *Z* available where *X* and *Y* are female voices and *Z* is a male. Male speaker *Z* was selected to have a different gender than *X* and *Y* to make the situation more challenging for the new scheme. Altogether 126 utterances from each speaker grouped formed the training set while there was also a separate test set of 10 sentences available from each speaker. The database was UK-English speech with a sampling rate of 8kHz.

In the first experiment summarized in Table 1, the performance of voice conversion was evaluated between the conventional approach using full training set and re-estimation approach using very limited adaptation data, more precisely 1 or 3 utterances. No specific criterion was used for the selection of those (1 and 3) utterances but phonetically balanced utterances are preferred for optimal results. For the conventional approach, a baseline conversion model, denoted as Baseline, was trained on the full training set from source *X* to target *Z* to demonstrate the upper performance limit. For the re-estimation approach, we firstly trained a seed conversion model GMM^{XY} on the full training set from source *X* to target *Y*. A parallel subset of 1 or 3 utterances from speakers *X* and *Z* was used to adapt the seed conversion model using re-estimation algorithm mentioned above, resulting in the corresponding GMM model, denoted as Adapt. The model re-estimation was performed using only the mean parameters.

In addition, small subsets of training data (1 or 3 utterances) were also used to build models between X and Z from scratch by EM training (denoted as EM). These models can be compared with the re-estimated conversion model using the same amount of data as seen in Table 1. The table shows the mean squared error (MSE) between converted and target speech (LSF vectors represented in Hz), measuring the performance in an objective way.

Modified mean opinion score (MOS) tests were also carried out to provide a numerical indication of the perceived conversion performance of the conventional and the re-estimation based approaches in a subjective listening test. The score was expressed as a single number in the range -2 to +2, where 0 denotes identical conversion performance, and +/-2 indicates a large difference in the perceived conversion performance. The samples were evaluated on two criteria, the identity match and the converted speech quality but only one number was assigned as a subjective score of preference with these two criteria in mind.

Table 1 : MSE between the converted and the target (*Z*)

	Adapt	**EM**	**Baseline**
1 utt	21630	27491	16284
3 utts	20460	21052	16284

Table 2 : Subjective listening test between baseline and adaptation approaches using 1 utterance.

Adapt vs. EM	Adapt vs. Baseline
+1.65	-1.0

In the subjective listening tests summarized in Table 2 a set of 10 utterances was evaluated by 11 testers and the average score of this evaluation was 1.65 in the favour of Adapt against EM, and -1.0 favouring Baseline against Adapt as shown in Table 2. Only the case of 1 training utterance is evaluated in the listening test because it demonstrates the efficiency of the proposed scheme in the case when less data is available for the re-estimation. The result clearly indicates that the performance using re-estimation is not as good as Baseline trained with a large data set but it is reasonably close to it considering the fact that only one training sentence was used in the re-estimation. On the other hand, the proposed approach clearly outperforms conventional EM training with small data sets.

5. DISCUSSIONS AND CONCLUSIONS

As can be seen from both objective and subjective results presented in Section 4, the experiments demonstrated the efficiency of the re-estimation approach in LSF conversion and showed that it can reach good performance while using a very limited amount of adaptation data. However, it should be noted that voice conversion consists of both excitation and LSF conversion, and thus, strictly speaking, the objective measurements done on LSFs and the listening test results do not measure the same aspects. Nevertheless, the experiments supported each other and made it possible to point out some interesting conclusions.

- The mean re-estimation can be used effectively with extremely reduced data. The performance is not very sensitive to the amount of data and it is reasonably close to the baseline system.
- The re-estimated models perform significantly better than EM models trained on reduced data. This is explained by the difficulty of EM in finding both priors and covariance information from the limited and potentially biased data.

If the data set is extremely small, the covariance estimation becomes highly unreliable. It has been observed that the performance degrades if the re-estimation is done on combined mean and covariance as compared to mean-alone re-estimation. The covariance information becomes beneficial in re-estimation if the data is of reasonable size. The contribution of the covariance re-estimation may be important if the estimation is reliable.

A major strength of the proposed scheme is that it takes advantage of the re-estimation of the baseline model with only a limited amount of adaptation data. With the convention EM based solution, it is much more difficult to have reliable prior and covariance information since they cannot be reliably estimated from the reduced data. As a consequence, there are many practical advantages of the proposed approach as listed below.

- Only a very limited amount of speech data needs to be recorded. This is crucial for many practical embedded applications;
- Low complexity: less computation to adapt the model, no need to train on full training sets;
- Efficiency: fast model adaptation;
- Ideal solution especially for embedded applications;

The experimental results have demonstrated through both objective measures and subjective listening tests that the proposed approach is an efficient and practical tool for voice conversion applications.

6. ACKNOWLEDGEMENT

This work has been partially funded by the European Union under the integrated project TC-STAR - Technology and Corpora for Speech-to-Speech Translation - (IST-2002-FP6-506738, http://www.tc-star.org).

REFERENCES

[1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization", In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, New York, USA, 1988.

[2] L. Arslan, and D. Talkin, "Voice conversion by codebook mapping of line spectral frequencies and excitation spectrum", In *5th Proceedings of European Conference on Speech Communication and Technology*, Rhodes, Greece, 1997.

[3] Y. Chen et al., "Voice conversion with smoothed GMM and MAP adaptation", In *Proceedings of 8th European Conference on Speech Communication and Technology*, Geneva, Switzerland, 2003.

[4] A. Kain, and M. Macon, "Spectral voice conversion for Text-to-Speech synthesis", in *Proceedings of International conference on Acoustics, Speech and Signal Processing*, Seattle, USA, 1998.

[5] E. Kim, S. Lee, and Y. Oh, "Hidden Markov Model Based Voice Conversion Using Dynamic Characteristics of Speaker", In *5th Proceedings of European Conference on Speech Communication and Technology*, Rhodes, Greece, 1997.

[6] C.-H. Lee, C.-H. Wu, "MAP-based adaptation for speech conversion using adaptation data selection and non-parallel training", In *Proceedings of International conference on Spoken Language Processing*, Pittsburgh, USA, 2006.

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

[8] J. Nurminen, V. Popa, J. Tian, Y. Tang, and I. Kiss, "A Parametric Approach for Voice Conversion", In *Proceedings of TC-STAR Speech-to-Speech Translation Workshop*, Barcelona, Spain, 2006.

[9] L. Rabiner, and B. Juang, *Fundamentals of speech recognition*, Prentice-Hall, USA, 1993.

[10] Y. Stylianou, O. Cappe, and E. Moulines, “Continuous probabilistic transform for voice conversion”, *IEEE Transaction on Speech and Audio Processing*, vol. 6, no.2, pp.131-142, 1998.
[11] T. Toda, Y. Ohtani, K. Shikano, ”Eigenvoice conversion based on Gaussian mixture model”, In *Proceedings of International conference on Spoken Language Processing*, Pittsburgh, USA, 2006.
[12] T. Watanabe, T. Murakami, M. Namba, T. Hoya, and Y. Ishida, “Transformation of spectral envelope for voice conversion based on radial basis function networks”, In *Proceedings of International conference on Spoken Language Processing*, Denver, USA, 2002.
[13] G. Xuan, W. Zhang, and P. Chai, “EM Algorithms of Gaussian Mixture Model and Hidden Markov Model”, In *Proceedings of International Conference on Image Processing*, Thessaloniki, Greece, 2001.
[14] H. Ye, and S. Young, “Perceptually weighted linear transformations for voice conversion”, In *Proceedings of 8th European Conference on Speech Communication and Technology*, Geneva, Switzerland, 2003.