Efficient Gaussian Mixture Model Evaluation in Voice Conversion

Jilei Tian, Jani Nurminen and Victor Popa

Multimedia Technologies Laboratory
Nokia Research Center, Tampere, Finland
{jilei.tian, jani.k.nurminen, ext-victor.popa}@nokia.com

Abstract

Voice conversion refers to the adaptation of the characteristics of a source speaker's voice to those of a target speaker. Gaussian mixture models (GMM) have been found to be efficient in the voice conversion task. The GMM parameters are estimated from a training set with the goal to minimize the mean squared error (MSE) between the transformed and target vectors. Obviously, the quality of the GMM model plays an important role in achieving better voice conversion quality. This paper presents a very efficient approach for the evaluation of GMM models directly from the model parameters without using any test data, facilitating the improvement of the transformation performance especially in the case of embedded implementations. Though the proposed approach can be used in any application that utilizes GMM based transformation, we take voice conversion as an example application throughout the paper. The proposed approach is experimented with in this context and evaluated against an MSE based evaluation method. The results show that the proposed method is in line with all subjective observations and MSE results.

Index Terms: voice conversion, speech subjective evaluation, Gaussian mixture model

1. Introduction

Voice conversion technology enables to transform one speaker's speech pattern into another speaker’s pattern with distinct characteristics, giving it a new identity, while preserving the original content or meaning. Speech and signal processing techniques are used for the modification of the speech of a source speaker to sound as if it was spoken by a target speaker. Though commercial usage of voice conversion techniques has not been very popular yet, the interest has risen immensely over the last few years. One of the reasons is the attractive idea to use voice conversion in cost-effective individualization of text-to-speech (TTS) systems. Without voice conversion, new voices have to be created in a time-consuming and expensive way using extensive recordings and manual annotations. Voice conversion can be also used to make a synthetic voice speak in languages that the original voice talent cannot speak. Other applications for voice conversion include security related usage to hide the identity of the speaker and entertainment applications, etc.

The research on voice conversion has received an increasing amount of attention, and the different voice conversion approaches have been proposed in the literature. From the technical point of view, typical approaches presented in the literature include Gaussian mixture modeling (GMM) based conversion [2], neural network based conversion [6], hidden Markov model (HMM) based conversion [3], linear transformation based conversion [7], and codebook based conversion. Among those techniques, the vast majority of the current voice conversion systems focus on data-driven GMM-based transformation on the spectral aspects of conversion, including instantaneous pitch. We have also used the GMM based approach in a parametric framework that also allows very efficient speech compression.

Research results found in the literature have shown that the GMM based approach can be used successfully in voice conversion. In the GMM based transformation, the combination of source and target vectors is used to estimate the GMM parameters for the joint density. The GMM-based conversion function is used to minimize the mean squared error between the transformed and target vectors. Apparently, the quality of the trained GMM has a tremendous influence on the performance. Therefore, efficient objective evaluation of GMM models is becoming very important when going towards a better conversion quality.

The existing conventional objective approaches for GMM quality evaluation based on distance measures such as mean squared error require test or validation data and are rather heavy from the viewpoint of embedded implementations, which may prevent such quality evaluations in embedded applications. These kinds of approaches have several inherent drawbacks:

1. Memory and cost: need to obtain and store the validation data;
2. Consistency: test results are dependent on the selection of the test data, different results may be achieved using different validation sets;
3. Real-time feedback: difficult to integrate this kind of measurement into the model training process;
4. Complexity: computational load caused by the evaluation is rather high;

Thus more efficient objective GMM evaluation schemes that could avoid these problems should be investigated

The approach presented in this paper introduces a very efficient approach for objectively evaluating GMM quality that is readily suitable also for embedded implementations. The main idea in the proposed approach is to measure the quality of the model directly from the model parameters without using any test data. The approach makes it possible to generate better GMM models especially in practical embedded applications.

The paper is organized as follows. In the next section, the target application using the GMM based transformation approach is introduced. In Section 3, the efficient approach for the evaluation of GMM models is presented. Then, the promising experimental results are given and analyzed in Section 4. Finally, the conclusions are drawn in Section 5.

2. GMM based voice conversion system

In GMM based transformation, multiple mixtures of Gaussian distributions are trained using joint aligned feature vectors combined from source and target. As an example application that uses this technique, we introduce our voice conversion system very briefly in this section. The introduction covers the three main areas: feature extraction, alignment and GMM training.

2.1. Feature extraction

The features presented in this paper are based on a parametric speech model inspired by the successful usage of a similar model in a low-bit-rate speech coding application [5]. The parametric model contains favorable properties from the viewpoint of both voice conversion and speech coding, and allows a seamless combination of these two aspects.

The speech model is based on the fact that a speech signal, or alternatively a vocal tract excitation signal, can be represented as a sum of sine waves of arbitrary amplitudes, frequencies and phases. To facilitate both voice conversion and speech coding, a simplified sinusoidal model parameterized using the pitch, the voicing, the residual amplitude spectrum, and voicing information for the spectrum, is applied to the modeling of the vocal tract excitation signal.

The excitation signal is obtained using the well-known linear prediction approach. The line spectral frequency (LSF) representation of speech is extracted as vocal tract features. From the viewpoint of voice conversion, this widely-used representation is very convenient since it has a close relation to formant locations and bandwidths, and it offers favorable properties for different types of processing and guarantees the filter stability.

2.2. Alignment

The training of the GMM models utilizes aligned parametric data from the source and target voices. The alignment is achieved in two steps. First, both the source and target speech signals are segmented and then a finer-level alignment is performed within each segment. The segmentation is performed at phoneme-level using HMM models. It is also possible to utilize manually labeled phoneme boundaries if such information is available but this is not used as the only solution to avoid the requirement for any manual processing that would be time-consuming and prone to human errors.

In principle, the speech segmentation could be conducted using very simple techniques, for example by measuring spectral change without taking into account knowledge about the underlying phoneme sequence. However, to achieve better performance, we fully exploit the information about the phonetic content and perform the segmentation using HMM models. At first a sequence of feature vectors is extracted from the speech signal frame by frame. The phoneme sequence associated with the corresponding speech is assumed known. Given the phoneme sequence, a compound HMM model is built up by sequentially concatenating the phoneme HMM models. Next, the frame-based feature vectors are associated with the states of the compound HMM model using Viterbi search to find the best path. By keeping track of the states, a backtracking procedure is able to decode the maximum likelihood state sequence [4]. The phoneme boundaries in time are then recovered by following the transition change from one phoneme HMM to another.

The phoneme-level alignment obtained using the procedure above is further refined by performing frame-level alignment using dynamic time warping (DTW) [4]. DTW can be used for finding the best alignment between two acoustic patterns. This is functionally equivalent to finding the best path in a grid to map the acoustic features of one pattern to those of the other pattern. Finding the best path requires solving a minimization problem that minimizes the dissimilarity between the two speech patterns. In the paper, DTW is applied on Bark-scaled LSF vectors and the algorithm is constrained to operate within one phoneme segment at a time. Non-simultaneous silent segments are disregarded.

2.3. GMM training

The combination of aligned source and target vectors $\mathbf{z}=[\mathbf{x}^T\ \mathbf{y}^T]^T$ can be used to train a GMM based conversion models [1][2]. In the training, we have used the popular approach that makes use of the aligned data $\mathbf{z}$ to estimate the GMM parameters of the joint distribution p($\mathbf{x}$,$\mathbf{y}$), where $\mathbf{x}$ and $\mathbf{y}$ correspond to source and target feature vectors, respectively. This is accomplished iteratively through the well-known Expectation Maximization (EM) algorithm.

The PDF of a GMM distributed random variable z can be estimated from a sequence of z samples $[\mathbf{z}_1\ \mathbf{z}_2\ \ldots\ \mathbf{z}_t\ \ldots\ \mathbf{z}_p]$ provided the dataset is sufficient by means of EM algorithms. In the particular case when $\mathbf{z}=[\mathbf{x}^T\ \mathbf{y}^T]^T$ is a joint variable the distribution of $\mathbf{z}$ can serve for probabilistic mapping between the two variables. In the case of voice conversion $\mathbf{x}$ and $\mathbf{y}$ are the corresponding features from the source and target speaker, respectively.

The distribution of $\mathbf{z}$ is modeled by GMM as

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

where c_l is the prior probability of $\mathbf{z}$ for the component l ($\sum_{l=1}^{L} c_l = 1$ and $c_l \geq 0$), L denotes the number of mixtures, and $N(\mathbf{z}, \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)$ denotes Gaussian distribution with the mean $\boldsymbol{\mu}_l$ and the covariance matrix $\boldsymbol{\Sigma}_l$. The parameters of GMM can be estimated using the well-known EM algorithm.

For the actual transformation, what is desired is a function $F(.)$ such that the transformed $F(\mathbf{x}_t)$ best matches the target $\mathbf{y}_t$ for all the data in the training set. The conversion function [1] that converts source feature $\mathbf{x}_t$ to target feature $\mathbf{y}_t$ is given by Equation (2).

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

The weighting terms in Equation (2) are chosen to be the conditional probabilities that the feature vector $\mathbf{x}_t$ belongs to the different components, as shown in Equation (3).

$$p_i(\mathbf{x}_t) = \frac{c_i \cdot N(\mathbf{x}_t, \boldsymbol{\mu}_i^{\mathrm{x}}, \boldsymbol{\Sigma}_i^{\mathrm{xx}})}{\sum_{l=1}^{L} c_l \cdot N(\mathbf{x}_t, \boldsymbol{\mu}_l^{\mathrm{x}}, \boldsymbol{\Sigma}_l^{\mathrm{xx}})} \tag{3}$$

3. GMM model evaluation

The main idea presented in this paper is to evaluate the quality of the GMM model directly based on the model parameters without using any testing data. More precisely, the measure utilizes the trace of target parts of the covariance matrices in the transform function to approximately evaluate the performance of GMM model in the transformation task. The proposed measure is very efficient to compute and it does not require any test data as the measurement is done directly from the model itself.

The evaluation method is derived by considering the properties of the GMM based transformation approach. The objective in the optimization of the GMM parameters in the conversion function is to minimize the average squared conversion error (*D*) for the training dataset.

$$D = \frac{1}{n} \cdot \sum_{t=1}^{n} \left\| \mathbf{y}_t - F(\mathbf{x}_t) \right\|^2 \tag{4}$$

The mean squared error is usually computed also on a validation dataset to assess the GMM quality. Lower *D* scores indicate that trained GMM models perform better in the voice conversion task than the model having larger *D*. Another approach for estimating the conversion error can be derived from data statistics (i.e., model parameters) using the variance of the distribution of **y** given **x**, i.e. $\varepsilon(\mathbf{x}) = \mathrm{var}(\mathbf{y} \mid \mathbf{x})$. ε(x) can be treated as a measure of the uncertainty of the conversion. The smaller ε(x) is, the more accurate the conversion performs. The proposed method originates from equation (4) and can be applied as an efficient measure for model assessment.

In theory the quality of the GMM can be measured using:

$$Q = \int \varepsilon(\mathbf{x}) \cdot p(\mathbf{x}) \cdot d\mathbf{x} \tag{5}$$

To be able to estimate the quality from the model itself in practice, the different mixtures have to be taken into account in the computation. Moreover, to make the computational complexity lower, the following approximation is proposed, instead.

$$Q \approx \sum_{l=1}^{L} w_l \cdot tr\left(\mathbf{\Sigma}_l^{yy}\right) \tag{6}$$

where *tr(.)* denotes the trace of the matrix and w_l is the weight for the *l*th component. The value *Q*, also called trace measure and defined in Equation (5)-(6), is proposed to be used for evaluation of GMM performance.

We have applied the GMM on the features in discrete cosine transform (DCT) domain. The decorrelation tendency of DCT-ed features ensures almost diagonal covariance matrix. In this way the trace can better approximate the variance of the data in multiple dimensions. Therefore equation (6) becomes more accurate. The GMM model performs better when *Q* value decreases. The proposed measure can be computed very efficiently and the measurement can be done directly on the model itself without any validation data. This measure can be used, for example of guiding the training of the transformation system towards better modeling. As very efficient implementations can be designed for the proposed scheme, it is particularly suitable for embedded applications. Nevertheless, the technique still has benefits in other applications as well, as there is no need to have any evaluation data and the results are always consistent.

4. Experiments

In order to verify the theoretical reasoning described in the Section 3, we carried out some experiments using voice conversion data. In these experiments, pitch and LSF parameters were studied mainly because of their importance in speech perception. Parallel utterances for two speakers were used for training (90 sentences) and testing (99 sentences). The models were trained on combined 20 dimensions LSFs and two dimensions pitch features from source and target speakers using the EM algorithm, respectively.

4.1. Trace measure vs. number of mixtures

A preliminary test was firstly carried out to verify that the proposed measure can meaningfully evaluate different models having different number of mixtures. Perceptual observations have indicated that the suitable number of mixtures for the conversion of LSFs and pitch features is 16 and 8, respectively, giving the best tradeoff between conversion quality and computational load. The trace measures for the corresponding models with different number of mixtures (as seen in Figure 1 and Figure 2) are completely in line with perceptual observations.

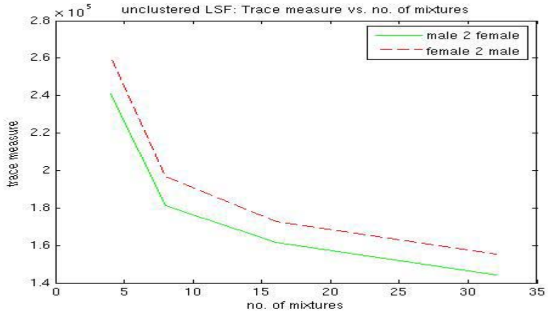

Figure 1. Trace measures vs. number of mixtures (LSF).

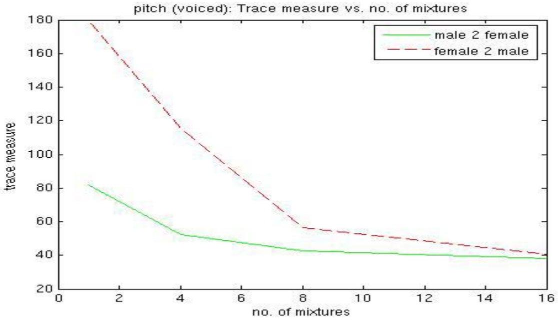

Figure 2. Trace measure vs. number of mixtures (pitch).

4.2. Comparison between trace and MSE

The second experiment included comparative tests between the trace measure and the conventional MSE approach. Again, the evaluation included pitch and LSF parameters. The training was done on normalized data. Put more specifically, the features were first normalized using scaling. DCT transform is applied to decorrelate the features. The conversion requires now normalization, DCT transform, mapping through GMM, inverse DCT transform and denormalization. It should be noted that the models were trained only on the training set, while both training and testing set were converted and analyzed separately for calculating MSE. Separate models were trained for the different directions of the conversion (from male to female and from female to male). GMM models are also trained on the voiced and unvoiced data, as denoted as model 1 and model 2. The converted data was compared to the target data in terms of MSE. The results from this experiment are given in Table 1:

Table 1. GMM models evaluated using MSE.

	GMM models	**Female to Male**	**Male to Female**
Test set	***Pitch (voiced)***	*212*	*95*
	LSF model 1	*17438*	*16515*
	LSF model 2	*18213*	*16931*
Train set	***Pitch (voiced)***	*224*	*91*
	LSF model 1	*17199*	*16234*
	LSF model 2	*18050*	*17054*

The trace measures of the same models are given in Table 2. They were computed using Equation (6).

Table 2. GMM models evaluated using trace measure.

GMM models	**Female to Male**	**Male to Female**
Pitch (voiced)	*0.785*	*0.473*
LSF model 1	*4.764*	*4.609*
LSF model 2	*5.029*	*4.886*

As can be seen, the MSE and trace measures are completely in line with each other for both the training and validation sets. Moreover, the proposed measure can again also confirm our perceptual findings on our voice conversion data: male-to-female conversion has better quality (smaller errors) than female-to-male conversion, and LSF model 1 outperforms LSF model 2.

5. Conclusions

In this paper, we focused on the model evaluation aspects in the context of Gaussian mixture modeling based transformation. More specifically, we developed a novel procedure for efficient evaluation of the GMM models without using any evaluation data. The proposed approach was experimented in the voice conversion task.

It is remarkable that the proposed trace measure is perfectly in line both with perceptual observations and MSE results (for both the training and validation sets). The use of the presented measure leads to the same conclusions with significantly less computation and without any validation data or perceptual evaluation. Thus, based on the presented practical experiments, the proposed trace measure can be regarded as an effective and efficient quality measure of the GMM model in transformation task.

The proposed GMM evaluation scheme offers several advantages when compared to the conventional MSF based evaluation technique:

1. Efficiency: very fast computation;
2. Simplicity: no validation/testing data needed for the evaluation;
3. Consistency: MSE results depend on the test data, but the trace measure always gives the same result provided the GMM is kept unchanged;
4. Easy integration: it is easy to integrate the analytical evaluation as a feedback into the model training, aiming to improve the models;

Consequently, it can be concluded that the proposed approach offers a very good solution for the evaluation of GMM model in the transformation applications. The method offers benefits in all implemented platforms, especially strong in embedded applications.

6. Acknowledgements

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

7. References

[1] Chen, Y., Chu, M., Chang, E., Liu, J. and Liu, R., “Voice conversion with smoothed GMM and MAP adaptation”, In Proceedings of 8th European Conference on Speech Communication and Technology (Eurospeech/Interspeech 2003), Geneva, Switzerland, 2003.

[2] Kain, A. and Macon, M., “Spectral voice conversion for Text-to-Speech synthesis”, In Proceedings of International Conference on Acoustics, Speech and Signal Processing, Seattle, Washington, USA, 1998.

[3] Kim, E.K., Lee, S. and Oh, Y., “Hidden Markov Model based voice conversion using dynamic characteristics of speaker”, In Proceedings of European Conference on Speech Communication and Technology, Rhodes, Greece, 1997.

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

[5] Ramo, A., Nurminen, J., Himanen, S. and Heikkinen, A., “Segmental Speech Coding Model for Storage Applications”, In Proceedings of International Conference on Spoken Language Processing (ICSLP04), Jeju Island, South Korea, 2004.

[6] Watanabe, T., Murakami, T., Namba, M., Hoya, T. and Ishida, Y., “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.

[7] Ye, H. and Young, S., “Perceptually weighted linear transformations for voice conversion”, In Proceedings of European Conference on Speech Communication and Technology, Switzerland, 2003.