Comparing linear and non-linear transformation of speech

Larbi Mesbahi, Vincent Barreaud and Olivier Boeffard
IRISA / ENSSAT - University of Rennes 1
6, rue de Kerampont, Lannion, France
{lmesbahi, vincent.barreaud, olivier.boeffard}@irisa.fr

Abstract: This paper aims to study voice conversion using linear and non-linear transform systems, based on respectively the Gaussian Mixture Models, GMM, and the Radial Basis function, RBF. We compare on an identical speech database both proposed approaches. We insist in particular on the objective measures of the transformation, in the case that we have not enough data recorded for the target speaker. We show for databases containing only one and two speech sentences, that the non-linear transform (RBF) gives weaker distortion scores than the GMM.

Key–Words: Voice conversion, Gaussian Mixture Model, Radial Basis Function.

1 Introduction

Automatic voice conversion consists in transforming a source speaker’s voice parameters in order to make it sounds as the voice of a target speaker. In many applications, particularly in biometric domain, it is necessary to do a conversion voice task with a few data uttered by the target speaker. Due to the lack of learning data, the voice conversion methods must be adapted, to ensure that the transform process still offers a good quality of conversion.

Usually, voice transformation of speech calls upon vector quantization [4], Gaussian Mixture Models (GMM) [1], [2], or Hidden Markov Models (HMM) [5]. Considering these different approaches, if we try to reduce the size of the learning data set, we should expect a negative effect on the transformation quality. In any case, we need to make adjustments on settings to improve the quality.

A transformation model, either linear or non-linear, is characterized by a number of free parameters. If the amount of learning data is not sufficient, the conversion could be of bad quality [6]. So a solution should be to propose a model with a minimal number of free parameters fitting a maximum amount of data.

Most studies in voice conversion research field consider a GMM-based linear transformation, thanks to their good transformation quality. The major drawback of this approach is that these models require lot of parameters to increase their accuracy, but become unstable with the amount of learning data is not sufficient. Considering this scenario, we propose to study a non-linear transformation function based on RBF in order to improve the quality of GMM when the learning data is sparse. The RBF are considered as a universal approximators [8], [7], which take into account the intrinsic characteristics of every speaker. The study made by [9] has revealed that the RBF approach gives a good objective and subjective performances.

Our contribution is inspired by the approach proposed by [10], which consists in transforming the speech by using multi-speakers interpolation, based on the RBF. The obtained results show a reduction of the spectral distortion of the order of 35% between the source and the target. This developed approach remains insufficient, because we don’t know the necessary conditions on the number of speakers and the number of sentences pronounced for each of them.

Our suggestion is based on a transformation between two speakers (source and target) under the hypothesis that we have a little training data recorded on the target. In this case, we try to compare it objectively with a GMM model. Our goal is to propose a voice conversion system applicable with a reduced amount of training data.

This paper is organized as follows. In section 2 we describe a GMM-based linear transformation. In section 3 a RBF-based non-linear transforms is detailed. Finally, in section 4, we describe the experimental methodology and the obtained results before a conclusion.

2 GMM-based linear transform

In the following we consider two sequences of $N$ $q$-dimensional acoustical vectors. The sequence corresponding to source speaker is represented by $X = [x_1, \ldots, x_N]^T$ and the target speaker by $Y = [y_1, \ldots, y_N]^T$. Given a GMM-based partitioning of a
speaker’s acoustic space, we need to estimate a piecewise function $F(\cdot)$ such that, for $n \in [1, \ldots, N]$, $F(x_n)$ will be close to $y_n$. The GMM partitioning is commonly a joint source/target estimation realized after a Dynamic Time Warping (DTW) alignment. GMM are frequently used to model a speaker’s acoustic space as it offers a continuous mapping of a vector. With such a model, the probability for a vector to be in a class is given by the weighted sum of probabilities for this vector to belong to each Gaussian component [1].

The probability distribution of $x_n$ is modeled by a $M$-component GMM as in the following equation:

$$P(x_n) = \sum_{m=1}^{M} \alpha_m N(x_n, \mu_m, \Sigma_m)$$

with $\sum_{m=1}^{M} \alpha_m = 1$, $\forall m \in [1, \ldots, M]$, $\alpha_m \geq 0$, where $N(\cdot, \mu_m, \Sigma_m)$ is the normal distribution of mean $\mu_m$ and covariance matrix $\Sigma_m$. The $\alpha_m$ scalars represent prior probabilities of component $m$. The GMM parameters are estimated by EM (Expectation-Maximisation) on a learning set. The obtained GMM can be a source model [1] or a joint model [2].

Once the GMM partitioning is done, the source/target conversion function can be derived as a weighted linear regression drawn from a conditional distribution of $y_n$ with respect to $x_n$, this piecewise linear transform [2] can be expressed as follows:

$$F(x_n) = \sum_{m=1}^{M} P_m(x_n)(\mu_m + \sum_{m, xy} \Sigma_{m, xx}^{-1} (x_n - \mu_m))$$

with $P_m(x_i)$ the posterior probability of the $m$-th component given $x_n$, $\mu_m$ and $\nu_m$ are the mean vectors of the source and target. The parameters $\Sigma_{m, xy} \Sigma_{m, xx}^{-1}$ represent the covariance matrices.

### 3 RBF-based non-linear transform

The purpose of Radial Basis Function (RBF) networks is to approximate a distribution by a set of kernel functions. A kernel is defined by a center $c_k$ and a receptor field $r$ which is the distance of a vector to the center. RBF networks are capable to approximate and interpolate multi-dimensional spaces and have been successfully applied in various scientific fields. For instance, RBF networks are often used in classification and approximation problems as well as in speech recognition. In [7], the author pointed out that RBF networks can approximate any continuous multivariate function in a compact domain providing that the number of centers is sufficient. The approximation not only depends on the number of centers but also on the coordinates of the centers, the shape of receptor field and the learning process [7] proposes several learning strategies:

- randomly select centers
- unsupervised learning process of centers
- supervised learning process of centers

In the following, we describe the network architecture and learning process of an RBF applied to the voice transformation issue. We can learn each $i$-th dimension of the target vector independently of the other dimensions. The input space of this network is the set of the learning source cepstral vectors. The hidden layer corresponds to the kernels used in the interpolation. The output space is the set of corresponding target cepstral vectors. Figure 1 represents a RBF network used for speech transformation.

![RBF network for non-linear transform of the i-th dimension of an acoustic speech vector.](image)

The non-linear transformation function produced by the RBF on the $i$-th dimension can be described as follows:

$$F(x_n)^i = \hat{y}_n^i = b_i + \sum_{k=1}^{m} w_{i,k} h_k(x_n), 1 \leq i \leq q$$

such as:

$$h_k(x_n) = \phi(||x_n - c_k||), 1 \leq k \leq m$$

where $c_k$ is a center picked among vectors within the source learning set. The selection of a center can either be done with the k-means method [7] or a maximum error criterion [8]. $h_k(x_n)$ is the radial basis kernel linked with the center $c_k$ and the $x_n$ vector. Gaussian ($\phi(r) = exp(-\frac{r^2}{\sigma_k^2})$) and Thin-plate-spline ($\phi(r) = r^2 log(r)$) are favored as kernel functions.

$w_{i,k}$ is the weight of the $k$-th center ($k \in [1 \ldots m]$) for the target’s $i$-th dimension ($i \in [1 \ldots q]$). $b_i$ represents the bias of the target’s $i$-th dimension. $\sigma_k$ is the $k$-th receptor field’s variance.
The learning process of the transformation function consists in estimating the weights \( w \). This can be done either with the Orthogonal Least Square method (OLS) [8] or by a Gradient Descent [7]. We use the latest in the following, so this method permits to monitor the number of center of the hidden layer whereas the OLS method automatically sets this number when the error threshold is reached.

### 3.1 Gradient Descent approach

There are two steps in the Gradient Descent method:
- Initialization of centroids of the source acoustical space with the k-means method.
- Learning of weights \( w \).

We used gaussian functions as kernel functions since their asymptotical characteristics match the properties of radial functions. The main problem lies in the setting of the center and receiving field of these kernels. Should the kernel be small, but the estimated density would be discontinuous and characterized with sharp summits. On the opposite, if the kernel is too large, the approximation is to lax and the density does not show enough details. This behavior is illustrated on figure 2.

**Figure 2** – Representation of three kind of kernels (medium (continuous line), large(dashed line) and small(dotted line)).

The learning process of the transformation function consists in setting the optimal weights that bound the parameters of the source and target vectors. The transformation function is re-written in the following form (equation 2):

\[
\tilde{y}_n^i = \sum_{k=0}^{m} w_{i,k} h_k(x_n), 1 \leq i \leq q
\]

(4)

such as : \( w_{i,0} = b_i \) et \( h_0(x_n) = 1 \)

The optimal weights are obtained when minimizing the error \( E \) as follows:

\[
E = \frac{1}{2} \sum_{n=1}^{N} \sum_{i=1}^{q} (y_n^i - \tilde{y}_n^i)^2
\]

(5)

such as \( y_n^i \) is the target vector and \( \tilde{y}_n^i \) is the corresponding transformed vector. The transformation parameters depend on the weights \( w \) that are modified during the Gradient Descent:

\[
\delta w_i = -\xi \nabla_{w_i} E = \xi \sum_{n=1}^{N} \epsilon_{i,n} h_n
\]

(6)

such as \( \xi \) is a learning coefficient and \( \epsilon_{i,n} \) is the output error which is computed as follows:

\[
\epsilon_{i,n} = y_n^i - \tilde{y}_n^i
\]

(7)

Weights are then updated as in:

\[
w_{i}^{iter} = w_{i}^{iter-1} + \xi \sum_{n=1}^{N} \epsilon_{i,n} h_n
\]

(8)

such as:

\[
h_n = [h_0(x_n)h_1(x_n) ... h_m(x_n)]^T
\]

(9)

\( w_{i}^{iter} \) corresponds to \( w_{i} \) et the \( iter \)-th step with \( w_{i} = [w_{i,0}w_{i,1} ... w_{i,m}]^T \). The process is stopped when the error threshold is reached.

### 4 Experiments

#### 4.1 Experimental methodology

Our comparative study was carried out on english database, noted bdl-jmk, which corresponds to two male speakers bdl and jmk extracted from the Arctic database [3]. The methodology is as follows, we take 70% of the corpus for the learning set and 30% for the test set. Utterances are randomly chosen. We have defined five learning corpus, each of them contains a reduced numbers of utterances (see table 1).

In order to consider reliable confidence intervals on these average scores, experiments are conducted 16 times (the complete process from the definition of a random training and test sets).

For each corpus, we follow this methodology:

1. Mel Frequency Cepstrum Coefficient, MFCC, vectors computing (sampling frequency is 16 KHz, a 30 ms Hamming window is applied, the analyzing step is 10ms). The order of the MFCC vector is set to 13 (including energy).
2. Dynamic time warping alignment between the source and target sequences using an euclidian norm on the MFCC vectors.

3. Parameter estimation of the GMM (means, covariances and weights) and RBF models.

4. Conversion of the source MFCC vectors by applying one of the conversion techniques described by the equations 1 et 2.

In this study, we aim to estimate the performances of the linear transform (GMM) and the non-linear transform (RBF), taking into account the reduction of learning corpus. The distortion score represents the mean distance between the target and the converted speaker normalized by the distance between the source and the target (Normalized Cepstral Distance):

$$e(\mathbf{\hat{y}}, \mathbf{y}) = \frac{\sum_{j=1}^{P} \sum_{i=1}^{N} (\mathbf{\hat{y}}_{ij} - \mathbf{y}_{ij})^2}{\sum_{j=1}^{P} \sum_{i=1}^{N} (\mathbf{x}_{ij} - \mathbf{y}_{ij})^2}$$  \hspace{1cm} (10)

such that : $\mathbf{\hat{y}}$ is the transformed source vector, $\mathbf{y}$ is the target vector and $\mathbf{x}$ is the source vector.

### 4.2 Results and discussion

#### 4.2.1 Variance effect

In this experiment, we have evaluated the RBF approach on the bdl-jmk-A test database, we adopted two strategies concerning the variance : Variance(1), we keep the same variance for all kernels $h_k(x_n)$ and Variance(2), we adapt, from the learning data, the variance of each kernel.

Our goal is to study the behaviour of the transformation under the influence of the variance. The results described in table 2 represent the distortion scores calculated for both strategies. They show that the results remain comparable in the two case, whatever the solution used to change the variance. As a consequence, we adopted the variance(1) strategy for the rest of the experiment.

#### 4.2.2 Trajectory of cepstral parameters

According to the transformation models (equations 1 and 2), We deduced that the number of parameters of the RBF and the GMM are respectively of the order of $2mq + q$ and $M(2q^2 + 2q + 1)$ (m represents the number of centroids, M represents the number of GMM components).

Considering the behavior observed on several cepstral dimensions which remain the same, we draw three trajectories of the $3^{rd}$ cepstral parameter.

The first trajectory (in red) corresponds to a GMM transform with 2 components, the second (in blue) corresponds to a RBF transform with 32 centroids(kernels) and the last (in green) corresponds to the trajectory of the target. The comparison between RBF (32 centroids) and GMM (2 components) is based on the equivalence of the number of free parameters of both approaches.

#### 4.2.3 Reduction of learning data size

The purpose of this experience is to estimate both approaches, GMM and RBF, under the influence of training data reduction. By observing the table of figure 3, we can raise two important remarks :

<table>
<thead>
<tr>
<th>#centroids</th>
<th>2</th>
<th>4</th>
<th>8</th>
<th>16</th>
<th>32</th>
<th>64</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variance(1)</td>
<td>0.61</td>
<td>0.51</td>
<td>0.48</td>
<td>0.41</td>
<td>0.41</td>
<td>0.40</td>
</tr>
<tr>
<td>Variance(2)</td>
<td>0.62</td>
<td>0.52</td>
<td>0.45</td>
<td>0.42</td>
<td>0.41</td>
<td>0.40</td>
</tr>
</tbody>
</table>

Tab. 2 – Scores of distortion obtained on bdl-jmk-A test database, evaluated for one experiment. Variance(1) : global variance and Variance(2) : local variance.
1. For the bdl-jmk-A and bdl-jmk-B databases, containing respectively 52 and 21 sentences, the linear model (GMM) gives a distortion score smaller than the non-linear model (RBF), in spite of an overfitting effect (see the results of [6]).

2. For the bdl-jmk-C base, containing 10 sentences, we observe that we cannot evaluate a joint GMM of 64 components. The same remark on bdl-jmk-D for a GMM of 8, 16 and 32 components. Besides, the distortion increases as we reduce the amount of data. This can be justified by the fact that there are fewer data with more free parameters for the transformation function. On the other hand, the RBF approach behaves better from the results obtained on the bdl-jmk-D base. This justifies that the RBF transformation function requires fewer free parameters. As a consequence, the model is more adaptable with a little amount of training data.

3. For the bdl-jmk-E base which contains only one utterance, the RBF model gives a smaller distortion score compared to GMM model. This is based on the evaluation of the RBF model on 2, 4, 8 and 16 centroids, which gives respectively $0.664 \pm 0.014$, $0.605 \pm 0.023$, $0.542 \pm 0.018$ and $0.513 \pm 0.017$ of distortion scores. In the other hand, for the GMM model, we have obtained a distortion score for only 2 components, which corresponds to $0.629 \pm 0.037$.

As a conclusion, we can say that facing a situation where there is not enough data to represent a speaker, it is recommended to use non-linear transforms which require fewer adjusted parameters. Consequently, they are more adaptable when the amount of data is not enough to train a GMM.

5 Conclusion

According to the obtained results, we can conclude that the GMM behaves better than the RBF for the learning databases containing enough data. The challenge lies in the faculty of converting speech with a small footprint of training data. In this context, with databases containing less than 10 sentences, the non-linear transform (RBF) supplies weak distortion scores than the GMM. The obtained results on learning database with only 2 utterances (on average 4 seconds of recording) can prove it. We can justify this by the fact that when the number of parameters increases, we reach an overfitting effect for the GMM earlier than for the RBF network. Whilst, the non-linear transform systems as the RBF, required less free parameters, thus more adaptable with less training data.

References:


Tab. 3 – Measures of distortion obtained on test databases (bdl-jmk-A, bdl-jmk-B, bdl-jmk-C and bdl-jmk-D) with a 95% of confidence interval. On the left, the results of a GMM approach with different components. On the right the results of a RBF approach with different centroids.