GMM-Based Speech Transformation Systems under Data Reduction

Larbi Mesbahi, Vincent Barreaud, Olivier Boeffard

IRISA / University of Rennes 1 - ENSSAT
6 rue de Kerampont, B.P. 80518, F-22305 Lannion Cedex
France
{lmesbahi,vincent.barreaud,olivier.boeffard}@irisa.fr

Abstract

The purpose of this paper is to study the behavior of voice conversion systems based on gaussian mixture model (GMM) when reducing the size of the training data corpus. Our first objective is to locate the threshold of degradation on the training corpus from which the error of conversion becomes too important. Secondly, we seek to observe the behavior of these conversion systems with regard to this threshold, in order to establish a relation between the size of training data corpus and the complexity of each method of transformation. We observed that the threshold is beyond 50 sentences (ARCTIC corpus), whatever the conversion system. For this corpus, the conversion error of the best approach increases only by 1.77 % compared to the complete training corpus which contains 210 utterances.

Index Terms: voice conversion , GMM, learning data reduction.

1. Introduction

During these last years, many applications in speech processing as text-to-speech synthesis or biometric identification by voice called upon speech transformation techniques. A voice conversion system tries to modify the vocal characteristics of a source speaker so that it is perceived as a target speaker. This technological issue is important: a man/machine interaction service can be more acceptable by offering various TTS voices [1][2].

Seminal approaches carried out a mapping-codebook conversion [3]. The main drawback of these approaches lies in the introduction of spectral discontinuities on the transformed signal. Several solutions were proposed in order to improve quality and precision, among them neuronal approaches [4], or segmental codebooks (STASC) [5]. Gaussian Mixture Model classifiers (GMM) make it possible to improve the mapping-codebook approaches, [6] [7]. Recently, the latter have been generalized using Hidden Markov Models, HMM, in order to treat the temporal dynamic aspect of the conversion function [8]. During our study, we were interested in the transformation techniques based on GMM acoustic segmentation, seen its many advantages such as the robustness, the continuity and the precision of the conversion function. However, these techniques have some drawbacks as over-fitting and the over-smoothing [1]. For many applications, in particular in the biometric field, it is necessary to carry out a voice conversion with very few data for the target speaker. The recording duration is usually short and the kind of voice differs from a speaker to another. In such a situation of scarce data, voice conversion methods must be adapted to ensure a good conversion quality. Our objective is to study the behavior of different conversion techniques based on GMM models [6] [7] [2] [1] under few learning data conditions. For this purpose, we gradually reduced the learning corpus and we evaluated the transformation quality on a reference test corpus. We carried out successive reductions, respectively of 75%, 50%, 25%, 10% and 5% on the initial learning set. Our objective is not to search for an optimization of the content of the training corpus under some imposed speech duration constraints, but to see whether the performance of the studied approaches remain stable when reducing the training corpus size. A complementary objective is to estimate the number of necessary sentences in order to maintain a good quality of conversion. Our main goal is to establish a compromise between the size of stored data and the conversion precision. This paper is organized as follows. In section 2 the studied GMM-based voice conversion approaches are presented. In the section 3 we treat and analyze the effect of data reduction on the quality of the conversion. Section 4 describes the experimental methodology and the obtained results. The conclusion will draw some prospects of this study.

2. GMM-based voice conversion

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 the speakers' acoustic spaces, we need to estimate a piecewise function $\mathcal{F}(.)$ such that, $\forall n \in [1, \ldots, N]$, $\mathcal{F}(x_n)$ 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 offering a continuous mapping of the acoustic vector space. 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 [6].

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

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

with $\sum_{m=1}^{M} \alpha_m = 1, \quad \forall m \in [1, \ldots, M], \alpha_m \geq 0$, where $\mathcal{N}(., \mu_m, \Sigma_m)$ is the normal distribution of mean μ_m, and covariance matrix Σ_m. The α_m scalars represent prior probabilities of component m. The GMM parameters are estimated by EM (*Expectation-Maximisation*) on a learning set. The obtained GMM is a source model (see 2.1) or a joint model (see 2.2).

Once the GMM partitioning is done, the source/target conversion function can be derived as a weighted linear regression drawn from the conditional distribution of y_n with respect to

x_n (analogous to a bayesian regression). To present the studied techniques uniformly, this piecewise linear transform can be expressed with parameters A_m (matrix) and B_m (vector) as follows:

$$\mathcal{F}(x_n) = \sum_{m=1}^{M} P_m(x_n)[B_m + A_m(x_n - \mu_m)] \quad (1)$$

with $P_m(x_n)$ the posterior probability of the m-th component given x_n.

In the following section, we present the solution proposed by Stylianou, [6]. Section 2.2 describes Kain's approach[7]. Finally, in sections 2.4 and 2.4, we present two approaches which takes into account the risk of *over-smoothing*[1] and *over-fitting*.

2.1. GMM on source only

The conversion method proposed in [9] uses a GMM source model. The conversion function that produces the linear regression is given by:

$$\mathcal{F}(x_n) \quad = \sum_{m=1}^{M} P_m(x_n)[\nu_m + \Gamma_m \Sigma_m^{-1}(x_n - \mu_m)]$$

The ν_m and Γ_m parameters are estimated by a least squares minimization [6]. The covariance matrix Σ_m of the GMM model can be full or diagonal (called **stylianou-diag** in the following).

2.2. GMM on joint source-target

In this approach, [7] suggests to jointly model the target and the source by a GMM. Thus, $\forall n \in [1, \ldots, N]$ a joint vector is build, $z_n = [x_n y_n]$. We obtain the following density:

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

$$\Sigma_m = \begin{bmatrix} \Sigma_{(m,XX)} & \Sigma_{(m,YX)} \\ \Sigma_{(m,XY)} & \Sigma_{(m,YY)} \end{bmatrix} \qquad \mu_m = \begin{bmatrix} \mu_{(m,X)} \\ \mu_{(m,Y)} \end{bmatrix}$$

The conversion function becomes:

$$\mathcal{F}(x_n) = \quad (2)$$
$$\sum_{m=1}^{M} P_m(x_n)[\mu_{(m,Y)} + \Sigma_{(m,YX)}\Sigma_{(m,XX)}^{-1}(x_n - \mu_{(m,X)})]$$

In the following, this technique will be referred to as **kain**.

2.3. Conversion and Over-smoothing risk

The major flaw of these GMM-based techniques is clearly presented in [1]. The spectral characteristics of the converted voices are excessively smoothed, referred to as *over-smoothing*; the consequence is an unclear speech signal. In [2], *Chen et al.* have demonstrated that 90% of the elements of the matrix product $\Sigma_{(m,YX)}\Sigma_{(m,XX)}^{-1}$ are ≤ 0.1 and 40% are ≤ 0.01, the correlation between the source and target speakers being weak. The effect of this statistical smoothing is to reduce the influence of the second term in equation 1, that is to say the term which contains the variability of X. Toda et al., [1], preserves the quality of a GMM-based conversion while simultaneously reducing the *over-smoothing* phenomenon. The solution is to impose a minimum level on the variance of the converted speech vectors. The maximum likelihood model (ML) is proposed to overcome the *over-smoothing* effect. The conversion function correspond to the following form:

$$\mathcal{F}(x) = (W^T D_m^{-1} W)^{-1} W^T D_m^{-1} E_m \quad (3)$$

with

$$E_m = [E_1(m_{i1}), E_2(m_{i2}), \ldots, E_N(m_{iN})]$$

$$D_m^{-1} = diag[Dm_{i1}^{-1}, Dm_{i2}^{-1}, \ldots, Dm_{iN}^{-1}]$$

$$E_n(m_i) = \mu_{(i,Y)} + \Sigma_{(i,YX)}{\Sigma_{(i,XX)}}^{-1}(x_n - \mu_{(i,X)})$$

$$Dm_i = \Sigma_{(i,YY)} - \Sigma_{(i,YX)}{\Sigma_{(i,XX)}}^{-1}\Sigma_{(i,XY)}$$

n: takes the values from 1 to N, N: is the number of vectors, M: is the total number of GMM components, W: transformation matrix [10]. In the following, this solution will be called **toda**.

2.4. Conversion and Over-fitting risk

This phenomenon was already described by Stylianou in [6]: the problem is principally linked to the choice of a model that is too complex compared to the size of the learning set. *Over-fitting* is characterized by the fact that performances on the learning set increase and performances on a validation corpus decrease when the number of parameters of the model rises. The resulting model loses its generalization capability. In order to limit the *over-smoothing* issue while still obtaining a minimal distortion between the transformed and target vectors, we slackened the equality constraint on covariances introduced in [2] by directly binding these covariances to a diagonal matrix A_m (see equation 1). A diagonal matrix prohibits the cross-correlation between coordinates of the acoustic vectors. A_m is replaced by a global diagonal matrix, noted Γ. The coordinates of Γ, γ^j for $1 \leq j \leq q$, are estimated by a least squares estimation:

$$\gamma^j = \frac{\sum_{n=1}^{N}\left(y_n^j - \sum_{m=1}^{M} P_m(x_n)\mu_{(m,Y)}^j\right)\left(x_n^j - \sum_{m=1}^{M} P_m(x_n)\mu_{(m,X)}^j\right)}{\sum_{n=1}^{N}\left(x_n^j - \sum_{m=1}^{M} P_m(x_n)\mu_{(m,X)}^j\right)^2}$$

$B_m = \mu_{(m,Y)}$. We note this transformation by **gamma-vector**.

3. Data reduction effect

Within a general framework of automatic learning, the techniques of data reduction aim to reduce the computing time necessary to the transformation operations. The reduction is reached by selecting optimal databases, classically, either by techniques like K-means, or vectorial quantification [11]. Other approaches estimate a reduction in order to minimize the complexity of certain optimization problems [12].

In our study, we do not impose a specific acoustic criterion on the reduction mechanism. The adopted heuristics consist in reducing the initial database uniformly. Various progressive reduction are applied to the original training corpus in order to assess the influence on the quality of the transformation. Moreover, we try to establish a link between the size of the learning corpus and the parameters of the voice transformation model (number of training sentences, number of GMM components, dimension of the acoustic vector, etc.).

4. Experimental methodology

4.1. Experimental methodology

The comparative study is carried on an english database, noted *bdl-jmk*. This corpus corresponds to the speakers *bdl* and *jmk* of the ARCTIC speech database [13]. The methodology applied is as follows: 70% of the sentences in the corpus define the learning set. The remaining 30% define the test set. The sentences are chosen randomly. Based on this first learning and

test partition, we defined various reduced learning corpora. The learning corpus corresponding to x% reduction of the full learning corpus will be noted as x%-(*bdl-jmk*). To summarize, the methodological conditions are as follows:

1. $\forall x, y \in \{100, 75, 50, 25, 10, 5\}$ such as $x < y$, we have $x\%$-(*bdl-jmk*) $\subset y\%$-(*bdl-jmk*).
2. The *bdl-jmk* test corpus is the same for all reduction models and contain 90 utterances.

For each learning corpus, the number of speech utterances are as follows: 210 utterances for 100%-(*bdl-jmk*), 157 for 75%-(*bdl-jmk*), 105 for 50%-(*bdl-jmk*),52 for 25%-(*bdl-jmk*), 21 for 10%-(*bdl-jmk*) and 10 utterances for 5%-(*bdl-jmk*).

For each corpus, we respect the following methodology:

1. 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) except for the **toda** transformation, where a vector is of dimension 26 (MFCC with their deltas).
2. Dynamic time warping between the *source* and *target* sequences using an euclidian norm on the MFCC vectors.
3. Parameter estimation of the GMM models (means, covariances and weights). We estimate joint source/target models. The source or target models are obtained by marginalizing a joint model. The learning process is carried out with a relative convergence threshold on the likelihood set to $1e^{-5}$. GMM models with 8, 32 and 64 components have been calculated. According to the studied conversion techniques, covariance matrices are full or diagonal.
4. Conversion of the source MFCC vectors by applying one of the conversion techniques described previously.

In this paper, we use a distortion score to measure the performance of the studied conversion functions with respect to various reduction ratios.

This distortion is defined as the mean distance between target and converted speech and normalized by the distance between source and target (Normalized Cepstral Distance). For this purpose, we used the following normalized cepstral distance for our objective tests:

$$e(\hat{c}^s, c^t) = \frac{\sum_{i=1}^{N}\sum_{j=1}^{P}(\hat{c}_{ij}^s - c_{ij}^t)^2}{\sum_{i=1}^{N}\sum_{j=1}^{P}(c_{ij}^s - c_{ij}^t)^2}$$

such as:$\hat{c}^s$ is the transformed source vector, c^t is the target vector and c^s the source vector.

In order to consider reliable confidence intervals on these average scores, experiments are conducted 16 times (the complete process from the definition of a training and test sets). The scores are estimated both on test and learning corpora so as to appreciate the *over-fitting* effect. From the set of utterances issued in the ARCTIC corpus for two speakers *bdl* and *jmk*, 16 initial learning and test corpus were drawn randomly according to a 70/30 proportion. From each one of these 16 training corpora, 6 reduced corpora are drawn: from 100%-(*bdl-jmk*) to 5%-(*bdl-jmk*). Consequently, conversion techniques have to be tested on 96 corpora. For each one of these experiments, 3 models of acoustic space representation are calculated: GMM with 8, 32 and 64 components. Finally 6 transformation systems are tested, 2 of them required the training of parameters in addition to those of the GMM.

4.2. Results and discussion

The tables 1 and 2 present the Normalized Cepstral Distance between source and target for respectively the learning and test corpus. By column, the various reduction ratios applied to the learning corpora: 75%, 50%, 25%, 10% and 5%. By raw, various models of transformation for all GMM : 8, 32 and 64 components. An average score with a 95% confidence interval is estimated on 16 different experiments.

On reading these two tables, we can first note that each studied conversion system reacts in accordance with the following relations:

1. Learning[x%-(*bdl-jmk*)] $\leq$ Test[x%-(*bdl-jmk*)], $\forall x$.
2. Test[x%-(*bdl-jmk*)]$\leq$ Test[y%-(*bdl-jmk*)], $\forall x \leq y$.
3. Learning[x%-(*bdl-jmk*)] $\leq$ Learning[y%-(*bdl-jmk*)], $\forall$ x $\leq$ y.

Based on our discussion in section 3, we try to establish a relationship between the reduction ratio of the training corpus and parameters of transformation systems:

1. When the number of training sentences decreases, degradation increases. It will be seen that this degradation is not proportional to reduction ratio.
2. In certain extreme situations, it is not possible to compute a GMM model for lack of a sufficient number of data. For instance, the 64 components GMM cannot be computed on the 5%-(*bdl-jmk*) corpus. For the 10%-(*bdl-jmk*) corpus, the **stylianou-diag** with 64 components GMM cannot be carried out because the transformation matrices estimated by least square methods become singular.
3. The **toda** transformation system, which uses a source/target joint GMM of 52 dimensional acoustical vectors (MFCC coefficients source and target plus theirs derived), gives an error higher than any other method whatever the corpus reduction and the number of GMM component.

These observations lead to the following comments. The GMM parameters used by the studied transformations do not model efficiently the test data if the reduction of learning corpus is to important. Among those parameters, we note the probability distribution of a vector x_n for the m^{th} component of a GMM, noted as $P_m(x_n)$. This parameter influences largely the quality of the transformation. Moreover, the dimension of the acoustic vector should be counted as a parameter that influences the conversion quality. Indeed, in a similar framework, [14] shows that the classification error decreases as the dimension of observed vector increases (for a constant number of acoustic samples). What ever the learning method of the transformation is, a reduction of the learning set entails an increase of the distortion on the test set. One can search for a reduction threshold that keeps the distortion in an acceptable range. Yet, the studied conversion methods do not use the same number of parameters nor the same type to describe their transformations. Thus, they do not have the same behavior with regard to the reduction of the learning set. Consequently, a common reduction threshold, for all the conversion technique, is quite unimaginable. We rather propose to establish a range of reduction threshold that would establish a compromise between a light learning set and a high conversion precision. A parallel can be established between this last remark and [11] where "safety regions" are established in machine learning.

Transformation		Reduced training corpora					
		100%-(*bdl-jmk*)	75%-(*bdl-jmk*)	50%-(*bdl-jmk*)	25%-(*bdl-jmk*)	10%-(*bdl-jmk*)	5%-(*bdl-jmk*)
kain Learning	GMM 8	0.391 ± 0.001	0.390 ± 0.001	0.389 ± 0.002	0.383 ± 0.003	0.371 ± 0.007	0.351 ± 0.009
	GMM 32	0.375 ± 0.001	0.373 ± 0.001	0.370 ± 0.002	0.360 ± 0.003	0.328 ± 0.007	0.311 ± 0.010
	GMM 64	0.365 ± 0.001	0.363 ± 0.001	0.358 ± 0.002	0.342 ± 0.002	0.318 ± 0.007	–
gamma-vector Learning	GMM 8	0.423 ± 0.003	0.423 ± 0.003	0.423 ± 0.003	0.419 ± 0.003	0.414 ± 0.007	0.408 ± 0.007
	GMM 32	0.398 ± 0.001	0.397 ± 0.001	0.396 ± 0.002	0.391 ± 0.003	0.376 ± 0.007	0.370 ± 0.008
	GMM 64	0.385 ± 0.001	0.384 ± 0.001	0.382 ± 0.002	0.374 ± 0.002	0.363 ± 0.007	–
stylianou-diag Learning	GMM 8	0.422 ± 0.003	0.422 ± 0.003	0.422 ± 0.003	0.419 ± 0.003	0.415 ± 0.007	0.413 ± 0.006
	GMM 32	0.396 ± 0.001	0.395 ± 0.001	0.395 ± 0.002	0.392 ± 0.003	0.384 ± 0.007	–
	GMM 64	0.386 ± 0.001	0.385 ± 0.001	0.385 ± 0.002	0.380 ± 0.003	–	–
toda Learning	GMM 8	0.497 ± 0.001	0.497 ± 0.002	0.497 ± 0.002	0.496 ± 0.004	0.490 ± 0.008	0.482 ± 0.006
	GMM 32	0.456 ± 0.001	0.455 ± 0.001	0.454 ± 0.002	0.449 ± 0.004	0.437 ± 0.009	0.418 ± 0.010
	GMM 64	0.445 ± 0.001	0.444 ± 0.002	0.443 ± 0.002	0.436 ± 0.003	0.414 ± 0.009	–

Table 1: This table presents the Normalized Cepstral Distance between source and target for learning corpus. By column, the various reduction ratios applied to the training corpora are 75%, 50%, 25%, 10% and 5%. By raw, various models of transformation for all GMM : 8, 32 and 64 components. An average score is estimated on 16 different experiments associate with a 95% confidence interval.

Transformation		Reduced training corpora					
		100%-(*bdl-jmk*)	75%-(*bdl-jmk*)	50%-(*bdl-jmk*)	25%-(*bdl-jmk*)	10%-(*bdl-jmk*)	5%-(*bdl-jmk*)
kain Test	GMM 8	0.395 ± 0.002	0.395 ± 0.002	0.397 ± 0.002	0.402 ± 0.002	0.420 ± 0.004	0.446 ± 0.009
	GMM 32	0.387 ± 0.002	0.390 ± 0.002	0.396 ± 0.002	0.416 ± 0.002	0.472 ± 0.006	0.548 ± 0.016
	GMM 64	0.389 ± 0.002	0.395 ± 0.002	0.407 ± 0.002	0.440 ± 0.003	0.534 ± 0.009	–
gamma-vector Test	GMM 8	0.423 ± 0.003	0.423 ± 0.003	0.424 ± 0.004	0.425 ± 0.003	0.432 ± 0.003	0.435 ± 0.005
	GMM 32	0.402 ± 0.002	0.404 ± 0.002	0.406 ± 0.002	0.414 ± 0.002	0.434 ± 0.003	0.460 ± 0.007
	GMM 64	0.395 ± 0.002	0.398 ± 0.002	0.404 ± 0.002	0.419 ± 0.002	0.459 ± 0.006	–
stylianou-diag Test	GMM 8	0.422 ± 0.003	0.422 ± 0.003	0.424 ± 0.004	0.424 ± 0.003	0.430 ± 0.003	0.437 ± 0.004
	GMM 32	0.398 ± 0.002	0.399 ± 0.002	0.401 ± 0.002	0.406 ± 0.002	0.420 ± 0.003	–
	GMM 64	0.391 ± 0.002	0.393 ± 0.002	0.396 ± 0.002	0.405 ± 0.002	–	–
toda Test	GMM 8	0.496 ± 0.002	0.496 ± 0.003	0.497 ± 0.003	0.501 ± 0.004	0.505 ± 0.005	0.508 ± 0.004
	GMM 32	0.458 ± 0.002	0.459 ± 0.002	0.461 ± 0.002	0.463 ± 0.003	0.467 ± 0.004	0.498 ± 0.008
	GMM 64	0.451 ± 0.002	0.452 ± 0.002	0.456 ± 0.002	0.465 ± 0.003	0.488 ± 0.005	–

Table 2: This table presents the Normalized Cepstral Distance between source and target on the test corpus. By column, the various reduction ratios applied to the training corpora are 75%, 50%, 25%, 10% and 5%. By raw, various models of transformation for all GMM : 8, 32 and 64 components. An average score is estimated on 16 different experiments associate with a 95% confidence interval.

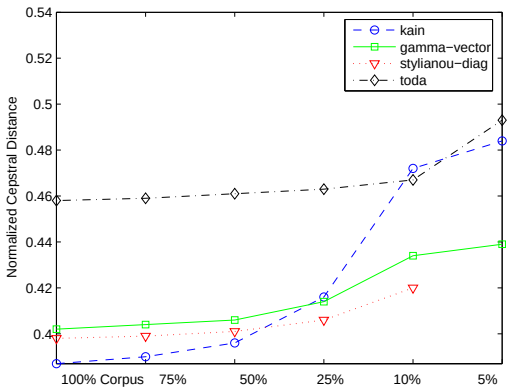

Figure 1: Evolution of normalized cepstral distance scores between transformed and target voices according to the data reduction with 75%, 50%, 25%, 10% and 5%, for the **kain**, **stylianou-diag**, **toda** et **gamma-vector** approaches. These results are obtained on the test corpus with 32 GMM components.

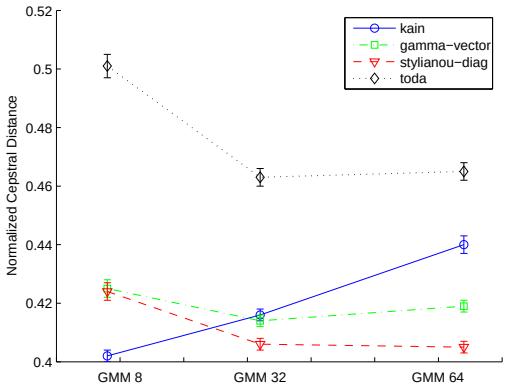

Figure 2: Evolution of normalized cepstral distance scores between transformed and target voices according to the number of GMM components for all approaches **kain**, **stylianou-diag**, **toda** and **gamma-vector**. We have fixed the threshold ratio at 25%. These results are obtained on the test corpus.

Figure 1 represents mean distortion scores between the transformed speaker and the target speaker, for a 32 component GMM. The scores are given for all the tested transformations for reduced learning sets. It can be observed that the performance of **kain** decreases for more than 25% reduction. Note that this transformation remains the more precise if it uses a less complex GMM (for instance a 8 component GMM: see table 2). **toda** is stable up to 10% reduction and crosses **kain** at this point. **stylianou-diag** is beter than **gamma-vector** up to 10%. **gamma-vector** is quite stable up to 5% since this technique uses far less parameters than the others.

We can observe, on this same figure, that the upper bound of the interval containing the optimal reduction thresholds for all the transformations is 25% (that is, a 52 sentences per corpus). Observe that the distortion of all the transformations lies in a stability zone that can go to above 25% reduction. For **gamma-vector**, the lower bound of this stability zone reaches 5% reduction (that is a 10 sentences per corpus). In that case, the conversion distortion is still acceptable (a 2.8% increase relative to the original learning set). Unfortunately, this technique suffers from over-fitting (for a fixed learning set, distortion rises as the number of components rises). For **stylianou-diag**, the threshold is before 10% reduction (21 sentences). It does not suffer from over-fitting. For **toda**, the threshold is before 5% reduction. The precision of this method is lower than any other transformation and is submitted to over-fitting. **kain** always give the best precision, its threshold is about 10%. To conclude, the general observation is all methods using less parameters are more stable.

Figure 2 shows the opposite situation of figure 1. Here, the distortion scores are presented for each transformation and for 25% reduction, since we consider this ratio as an optimal threshold. For each transformation, we used successively 8, 32 and 64 components GMM. We can notice that **kain** suffers from over-fitting even if its precision still overcomes **stylianou-diag** and **gamma-vector** for 8 components GMM. For 32 and 64 components GMM, **stylianou-diag** presents a better score than **gamma-vector**.

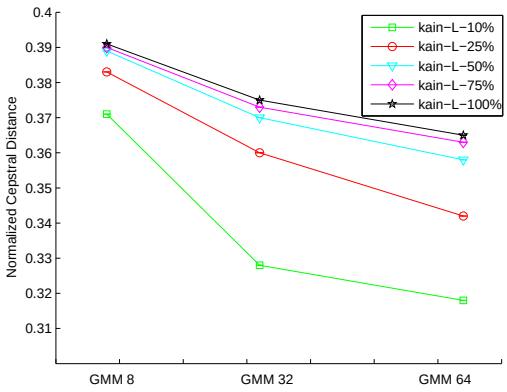

Figure 3: Evolution of normalized cepstral distance scores between transformed and target voices for **kain** approach, with the corresponding reductions 75%, 50%, 25% and 10%, according to the number of GMM components. These results are obtained on the learning corpus.

Figures 3 and 4 show the variations of the distortion scores

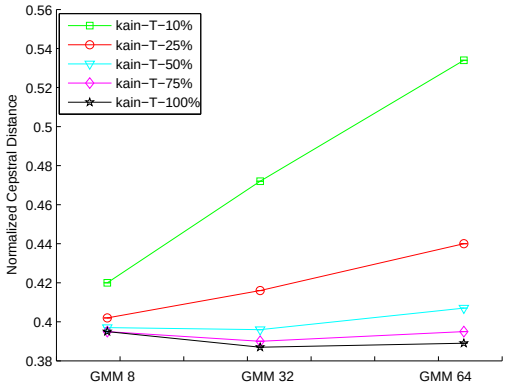

Figure 4: Evolution of normalized cepstral distance scores between transformed and target voices for **kain** approach, with the corresponding reductions 75%, 50%, 25% and 10%, according to the number of GMM components. These results are obtained on the test corpus.

of **kain** on the learning and test sets, for all reductions. It can be observed that, as the number of gaussian components rise, the score on the learning set improves while the score on the test set worsen. On the test set, the results can be divided in two classes. One of them contains the 100%, 75%, 50% and 25% reductions. For 10% reduction, scores can no longer be compared. This observation drove us to choose 25% as the reduction threshold. The obtained learning set regroups 52 sentences and is the best compromise between the corpus size and the conversion precision.

5. Conclusion

This work presents an experimental evaluation of various voice transformation techniques based on GMM models relative to the learning set's size. We observed that, in order to keep a good conversion score when this size is reduced, the number of parameter describing the transformations (number of gaussian component, the covariance type) must be reduced as well. For instance, for the transformation proposed by Kain with 32 components GMM, the normalized cepstral distance when using 5% of the original learning set, has a 41.6% variation relative to the score obtained with 100% of the learning set. This variation is only of 15.24% for a 8 components GMM. Furthermore, for the same transformation with 8 components GMM when using 25% of the original learning set, the variation of the distortion is only of 1.77% relative to the score obtained with 100% of the learning set. For a smaller learning set, the distortion rises unlinearly. We have observed that, on the Artic database, studied systems give fair conversion scores even if only 52 training sentences are available. Future work will evaluate the presented techniques when using various acoustic parameterizations. By varying the nature and dimension of the acoustic parameters, we seek to study the influence of reducing the learning set on conversion's precisions on an other level.

6. References

[1] T. Toda, A. W. Black, and K. Tokuda, "Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter," in *International Conference on Acoustics, Speech, and Signal Processing*, 2005, pp. I9–I12.

[2] Y. Chen, M. Chu, E. Chang, and J. Liu, "Voice conversion with smoothed gmm and map adaptation," in *EUROSPEECH*, Geneva, Switzerland, September 2003, pp. 2413 – 2416.

[3] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, New York, April 1988, pp. 655–658.

[4] M. Narendranath, H. A. Murthy, S. Rajendran, and B. Yegnanarayana, "Transformation of formants for voice conversion using artificial neural networks," Elsevier Science B. V., pp. 207–216, 1995.

[5] L. M. Arslan, "Speaker transformation algorithm using segmental codebooks (stasc)," *Speech Communication*, vol. 28, pp. 211 – 226, 1999.

[6] Y. Stylianou, O. Capp, and E. Moulines, "Continuous probabilistic transform for voice conversion," in *IEEE Transactions on Speech and Audio Processing*, vol. 6, 1998, pp. 131–142.

[7] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 1998, pp. 285–288.

[8] D. Helenca, B. Antonio, A. Kain, and J. Van Santen, "Including dynamic and phonetic information in voice conversion systems," in *Proceedings of the International Conference on Spoken Language Processing*, 2004, pp. 1193–1196.

[9] Y. Stylianou, O. Capp, and E. Moulines, "Statistical methods for voice quality transformation," in *EUROSPEECH*, Madrid, Espagne, 1995, pp. 447–450.

[10] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for hmm-based speech synthesis," in *ICASSP - International Conference on Acoustics, Speech, and Signal Processing*.

[11] K. Ravindra and H. Saman, "Reducing the number of training samples for fast support vector machine classification," *Neural Information Processing-Letters nd Reviews*, vol. 2, March 2004.

[12] M. Sebban, R. Nock, and S. Lallich, "Stopping criterion for boosting-based data reduction techniques: from binary to multiclass problems," *Journal of machine learning research*, vol. 3, pp. 863–885, 2002.

[13] J. Kominek and A. Black, "The cmu arctic speech databases for speech synthesis research," *Tech. Rep. CMU-LTI-03-177*, 2003.

[14] G. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEETrans. Information Theory*, vol. 14(1), pp. 55–63, 1968.