Estimation of GMM in voice conversion including unaligned data

Helenca Duxans and Antonio Bonafonte

Department of Signal Theory and Communications
TALP Research Center
Universitat Politècnica de Catalunya, Barcelona
http://www.talp.upc.es
{hduxans,antonio}@gps.tsc.upc.es

Abstract

Voice conversion consists in transforming a source speaker voice into a target speaker voice. There are many applications of voice conversion systems where the amount of training data from the source speaker and the target speaker is different. Usually, the amount of source data available is large, but it is desired to estimate the transformation with a small amount of target data.

Systems based on joint Gaussian Mixture Models (GMM) are well suited to voice conversion [1], but they can't deal with source data without its corresponding aligned target data.

In this paper, two alternatives are studied to incorporate unaligned source data in the estimation of a GMM for a voice conversion task. It is shown that when a limited amount of aligned parameters are available in the training step, to only include data from the source speaker increases the performance of the voice transformation.

1. Introduction

In many speech applications new synthesized voices are required. Among theses applications voice personalization of speech synthesizers can be found, for example to read an e-mail with its sender's voice, or to conserve speaker's voice characteristics through a translation system. Also, some entertainment products [2] are interested in modifying a reference voice in order to involve more the client, and in languages learning [3] it could be a great help to be able to listen to your own voice speaking a foreign language correctly.

The task of creating a new voice is time-consuming and expensive, this is a reason why voice conversion could be a good alternative. Voice conversion consists in changing the voice characteristics of one speaker (who is called source speaker) in such a way that a listener could think that the transformed voice belongs to a different speaker (target speaker).

One of the main difficulties in voice conversion is to find a minimum set of features that describes the voice individuality for one speaker. It is well known that spectral characteristics (such as formant positions and bandwidths or spectral envelope), as well as prosodic characteristics (pitch, phone duration) and syntactic and lexical properties identify the speaker personality.

One approach to voice conversion is to model speech signals according to the Vocal Tract + Excitation architecture [1], [4], [5]. Then, different transformations can be learned from each kind of information.

This work is focused on vocal tract conversion systems. In the following section, a system based in joint-GMM is revised in more detail. Then, in section 3, some applications in which the amount of target information is smaller than the amount of source information are presented, and two methods to incorporate this unaligned data in the training phase are studied. In section 4, the experimental work is presented and discussed. The paper ends with a brief summary and conclusions.

2. Vocal tract transformation

The vocal tract transformation system that was chosen as a start point in this study is based on joint Gaussian Mixture Models, as the system proposed by A. Kain [1].

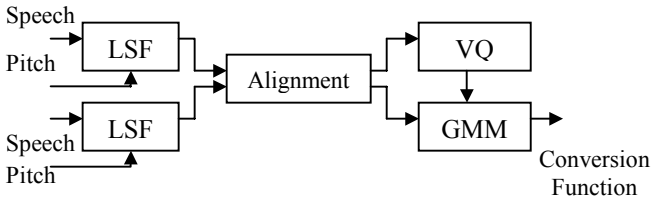

Figure 1: Training step system block diagram

This system is based on modeling the joint acoustic space of vocal tract parameters of the source and target speaker with a GMM.

This class of models is highly suitable to the regression problem in voice conversion because its properties, such as: high capacity to represent arbitrary density functions, acoustic space division by classes and space division with continuous functions.

The GMM is estimated maximizing the likelihood function (ML) of the joint source-target probability, applying an Expectation-Maximization (EM) algorithm. As initial values for the parameters, all the mixture weights are set $\mu_q=1/Q$ (Q: number of mixtures), covariance matrices are set to diagonal matrices and means and variances come from an initial clustering of the training data. To avoid unstable parameters in the estimation, from covariance mixture matrices becoming close to singular, a constant value is added to their diagonals in each EM iteration.

The spectral parameters modeled by GMM are line spectrum frequencies (LSF), extracted pitch synchronously from speech signals and aligned between both speakers (refer to section. 4). Therefore, data consist of pairs of source-target vectors.

Finally, a transformation function can be obtained through the regression of the target given the source parameters:

$$y_{trans} = E(y/x) = \sum_{q=1}^{Q} \alpha_q \left(\mu_y^q + \Sigma_{yx}^q \left(\Sigma_{xx}^q \right)^{-1} (x - \mu_x^q) \right) p(q/x) \quad (1)$$

where the parameters of the GMM distribution are: $\theta=\{\theta_q\}=\{(\alpha,\mu_x,\mu_y,\Sigma_{xx},\Sigma_{xy},\Sigma_{yx},\Sigma_{yy})_q\}$ for $q=1,\ldots,Q$.

The mixture of gaussians splits the acoustic input space and learns a linear regression surface in each partition.

In the next section we will extend this model to consider the use of unaligned data.

3. Unaligned Data

There are many applications of voice conversion that, in the training step, it is available more data from the source speaker than from the target. For example, if we are going to personalize the output of a TTS system we can generate as many sentences as we want. However, we would like to use as few sentences as possible from the target speaker. This difference of data amount between source and target is also present in incremental training voice conversion systems. For instance, in speech recognition, we can apply a transformation function to adapt the input speech to the system. Then, this function can be updated incrementally, using unaligned data, as more spectral parameters are available.

So, in this section some modifications to the estimation algorithm for the GMM are presented, in order to study if a set of source parameters, without the corresponding aligned target ones, can improve the performance of the reference system.

Previous studies [6] have shown that, for specific applications, including unlabeled data in classification problems increases the performance of the classification. Our purpose is to apply this idea in the regression field.

In section 2, the parameters of the mixture model are calculated with the criterion of maximizing the likelihood function of the source-target vector pairs available. Now, this likelihood function is modified to include unaligned data:

$$\begin{aligned} L(\theta/x,y) &= \prod_{i=1}^{N} p(x_i,y_i) \prod_{j=1}^{M} p(x_j) \\ p(x_i,y_i) &= \sum_{q=1}^{Q} \alpha_q N\left((x_i,y_i), \begin{bmatrix} \mu_x^q \\ \mu_y^q \end{bmatrix}, \begin{bmatrix} \Sigma_{xx}^q & \Sigma_{xy}^q \\ \Sigma_{yx}^q & \Sigma_{yy}^q \end{bmatrix} \right) \\ p(x_j) &= \sum_{q=1}^{Q} \alpha_q N\left(x_j, \mu_x^q, \Sigma_x^q \right) \end{aligned} \quad (2)$$

where N is the number of source-target vector pairs and M the number of source vectors without aligned target ones.

In the following sections, some assumptions are made to simplify the estimation of the parameters.

3.1. Fixed covariance matrices

For the sake of simplicity, let's assume that the unaligned source vectors will only significantly affect the means and weights of the mixtures.

So, a first GMM can be estimated from the aligned data as in section 2. Afterwards, an EM algorithm can be derived from expression (2) to recalculate the means and weights, without modifying the covariance matrices. The parameter expressions of this second EM algorithm at each iteration are:

$$\begin{aligned} \alpha_q &= \frac{1}{N+M} \left(\sum_{i=1}^{N} p(q/x_i,y_i,\theta^{n-1}) + \sum_{j=1}^{M} p(q/x_j,\theta^{n-1}) \right) \\ \mu_x^q &= \frac{\sum_{i=1}^{N} x_i\, p(q/x_i,y_i,\theta^{n-1}) + \sum_{j=1}^{M} x_j\, p(q/x_j,\theta^{n-1})}{\sum_{i=1}^{N} p(q/x_i,y_i,\theta^{n-1}) + \sum_{j=1}^{M} p(q/x_j,\theta^{n-1})} \\ \mu_y^q &= \frac{-\Sigma_{yx}^q \left(\Sigma_{xx}^q\right)^{-1} \sum_{i=1}^{N} (x_i - \mu_i^q)\, p(q/x_i,y_i,\theta^{n-1})}{\sum_{i=1}^{N} p(q/x_i,y_i,\theta^{n-1})} + \\ &\quad + \frac{\sum_{j=1}^{M} y_i\, p(q/x_i,y_i,\theta^{n-1})}{\sum_{i=1}^{N} p(q/x_i,y_i,\theta^{n-1})} \end{aligned} \quad (3)$$

It is expected that this re-estimation of the GMM will be fast and converge in few iterations, as the highest dimensional parameters are not re-calculated and the first GMM should be near a maximum of the aligned-unaligned likelihood function.

3.2. Alignment through transformation

Another alternative to reduce the computational complexity of the estimation of the aligned-unaligned likelihood function given by expression (2) is to complete the missing data, which means to make a guess about the alignment of source vectors. One possible way of completing data is to include transformed vectors as alignments. Then, the new likelihood function will be:

$$L(\theta/x,y) = \prod_{i=1}^{N} p(x_i,y_i) \prod_{j=1}^{M} p(x_j,y_j^{trans}) \quad (4)$$

It should be accepted that a starting transformation function, estimated as in section 2, is good enough to estimate a first set of transformed vectors. When source-transformed vector pairs are built, an EM algorithm as in section 2 can be applied over the joint parameter set of source-target and source-transformed vectors.

This idea can be iterated, transforming the non-aligned source vectors with the new regression function and re-calculating another GMM over the new joint set.

4. Experimental results

In this section, different experiment results are presented. The objective task is to transform the vocal tract of a male speaker into the vocal tract of a female speaker, testing the alternatives described in previous sections in order to evaluate the performance variation, taking into account the computational cost.

The speech corpus used in this work was produced to generate speakers of our unit selection TTS system. Speech and laringograph signals were recorded in an acoustically isolated room. A sample frequency of 32KHz and 16 bits per sample were used. For this work, signals were decimated to 8KHz. The corpus was phonetically segmented. To align source and target LSF vectors we applied dynamic time warping constrained by the phonetic segmentation. The total

corpus size is more than one hour for each speaker, but we will only use a few sentences from each one.

The evaluation index (EI) that will be used in this paper measures the reduction of distance between source and target speakers when the transformation is applied:

$$D_{LSF}(A,B) = \frac{1}{M}\sum_{m=1}^{M}\sqrt{\frac{1}{p}\sum_{i=1}^{p}\left(LSF_A^m(i) - LSF_B^m(i)\right)^2} \qquad (5)$$

$$EI = 1 - \frac{D_{LSF}(\text{target}, transformed)}{D_{LSF}(\text{target}, source)}$$

where LSF_A and LSF_B are the source and target LFS vectors, p its dimension and M the number of frames used in the test.

All the experiments have been done twice with different sets of frames and each regression function have been also tested twice, in order to assure that the results aren't adapted to training data.

4.1. Reference system

The reference system has been tested using 20 pairs of aligned sentences (training S20; 11K vectors pairs) and also with a reduced set of 5 pairs of aligned sentences (training S5; 2.4K vectors pairs). For each training set, different numbers of mixtures have being considered. The evaluation index has being computed with independent test data. The results are shown in table 1.

#sent. \ Q	2	4	8	16	32	64
S20	-	-	0.259	0.255	0.239	0.209
S5	0.202	0.210	0.179	0.144	-	-

Table 1: Evaluation index for training sets S20 and S5 as a function of the number of mixtures.

According to *table 1*, a regression function trained with 20 sentences performs the best for 8 gaussian mixtures, and if it is trained with 5 sentences the best is for 4 mixtures. When more mixtures are trained an overfiting problem appears. The experiments were repeated with different training and test sets, with the same conclusions.

4.2. Unaligned Data

We have evaluated the effect of adding unaligned data using both methods explained in section 4, for the two training sets (S20 and S5) and for each number of mixtures (Q). The results of reference systems trained with enlarged sets of aligned data will be also presented as a comparative value. We will name this configuration "enlarged-GMM". Note that in this configuration we use more data from the target speaker.

4.2.1. Fixed covariance matrices

Figure 1 and *2* show the results for S20 and S5 training sets. The first one presents, from left to right, groups of experiments where 10, 20, 40 and 80 (6.309, 12.267, 24.680 and 51.306 frames) source sentences have being used. In the second one 5, 10 and 15 sentences (2.985, 5.794 and 8.856 frames) were added. For each group, the evaluation index is shown, as a function of the number of mixtures and for three methods: reference system, enlarged training (more aligned data) and fixed covariance (more unaligned data).

As it was expected, adding more aligned data resolves the overfiting and increases the performance of the conversion system. So enlarged-GMMs perform better than other methods.

It can be seen from figure 1 (S20), that a slightly increasing in the performance is achieved using unaligned data, but it isn't significant, independently from the number of unaligned sentences used.

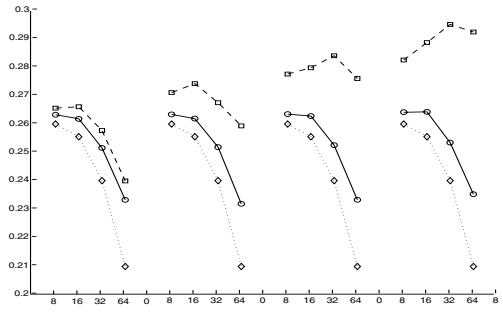

Figure 1: Training set S20: ✧) Reference GMM, □) Enlarged-GMM, o) fixed covariance method.

On the other hand, the results presented in *figure 2* for the S5 training set show significant improvements in the performance. Actually, adding 10 unaligned sentences and estimating a GMM with 4 mixtures is computationally fast, and the performance increases 18% (from 0.190 to 0.224).

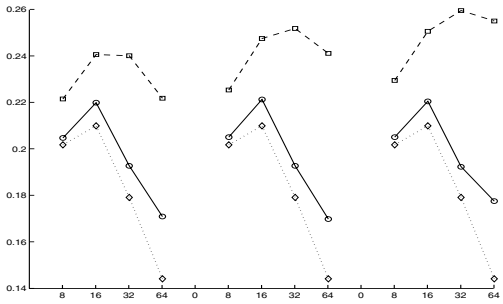

Figure 2:Training set S5: ✧) Reference GMM, □) Enlarged-GMM, o) Fixed covariance method

We can conclude that when we are dealing with models estimated over few data, modifications only in the mixture weights and means increase the performance. But, when the GMM has been estimated with enough data to be a good model, the restriction over the covariance matrices is a hard constraint and the results do not improve significantly.

4.2.2. Alignment through transformation

The next figure shows the performance indexes for experiments about alignment through transformation. To build the source-transformed pairs, the reference GMM with 8 gaussian mixtures was used in S20, and the reference GMM with 4 mixtures in S5. The same sets of unaligned data as in the previous section have been added.

It can be observe that a better performance is obtained than with fixed covariance matrices for both reference GMMs.

In fact, for the 20 initial sentences model, we can achieve an equivalent performance adding 10 sentences aligned (50% more of the initial set), than adding 20 only from the source speaker. The inclusion of more unaligned sentences doesn't increase the performance; this means that it exists a relationship between the two different sets of data.

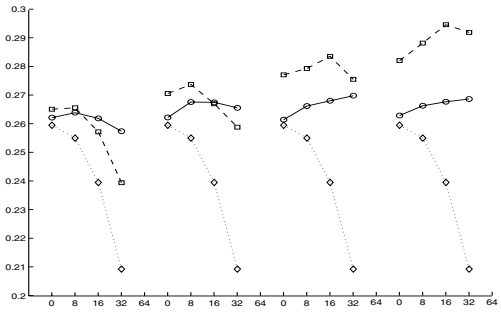

Figure 3: Training set S20. ✧) Reference GMM, □) Enlarged-GMM, o) Alignment through transformation.

In the case of a reference GMM trained with 5 sentences, the performance increment adding unaligned data is more remarkable, since in the reference system a limited amount of information about the source and the target is available.

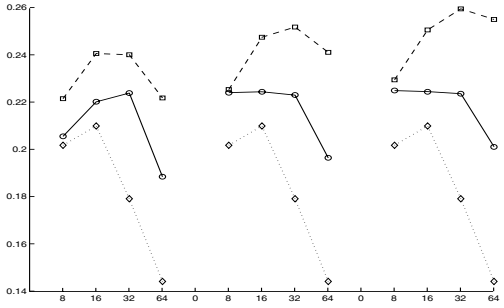

Figure 4: Training set S5: ✧) Reference GMM, □) Enlarged-GMM, o) Alignment through transformation.

According to the results, this technique is appropriated to increase the conversion performance when a limited set of aligned training vectors are available. Also, it should be noted that, in the case that voice conversion was integrated in another system and it was forced to used a fix number of mixtures higher than the optimal one, we can used alignment through transformation to estimate a GMM with more mixtures with the same performance than the reference one.

The iterative application of this method has been tested. So, starting from new models after adding unaligned sentences, two more iterations of this method have been done, each time using as a transformation function the last estimated regression. The performance slightly increased, but do not justify the increase on time computing.

As final results, next figure shows how the performance index evolves in an incremental training system, using a reference system or using alignment from transformation. When more training data is available, the increment in performance using unaligned data is smaller.

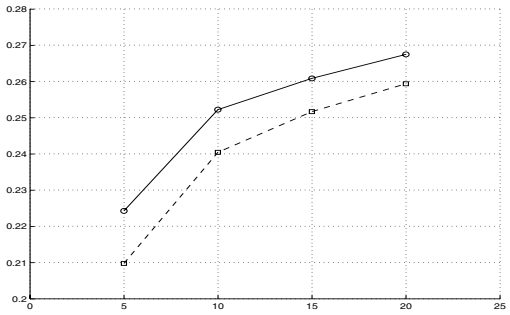

Figure 5: Performance of an incremental training system □) Reference GMM, o) New method

5. Conclusions

This work is focused on increasing the performance of a vocal tract conversion system based on joint source-target GMM. We are concerned on applications in which the amount of source and target speaker data, in the training step, isn't the same. It has been shown that a combined learning with aligned source-target data and source-transformed data increases the conversion performance, mainly when few training data is available. In this latest situation, to re-estimate only means and weights mixtures also increases the performance, with very reduced computational time.

6. Acknowledgements

This work has been sponsored by the Spanish Government under grant TIC2002-04447-C02.

7. References

[1] Kain, A.; Macon, M.W.; "Spectral voice conversion for text-to-speech synthesis", in *Proc. ICASSP 1998*, vol. 1, pp 285-288.

[2] Gustafson, J.; Sjölander, K.; "Voice transformations for improving children's speech recognition in a publicity available dialogue system", *ICSLP 2002*.

[3] Mashimo, M.; Toda, T.; Kawanami, H.; Kashioka, H.; Shikano, K.; Campbell, N.; "Evaluation of Cross-Language Voice Conversion Using Bilingual and Non-Bilingual Databases", *ICSLP 2002*.

[4] Masanobu, A.; Satoshi, N.; Shikano, K.; Kuwabara, H.; "Voice conversión through vector quantization", *Proc. ICASSP 1988*, vol. 1, pp 655-658

[5] Stylianou, Y.; Cappé, O.; Moulines, E.; "Continous probabilistic transform for voice conversion", *IEEE Trans. Speech and Audio Proc.*, march 1998, pp 131-142

[6] Miller, D. J.; Uyar, H. S.; "A mixture of experts classifier with learning based on both Labelled and Unlabelled Data" in *Advances in Neural Information Processing Systems* 1997, vol. 9, pp. 571--577.