# NON-PARALLEL SINGING-VOICE CONVERSION BY PHONEME-BASED MAPPING AND COVARIANCE APPROXIMATION

*Fernando Villavicencio and Hideki Kenmochi*

Corporate Research
Development Center
Yamaha Corporation
Hamamatsu, Shizuoka, Japan
{villavicencio,kenmochi}@beat.yamaha.co.jp

## ABSTRACT

In this work we present an approach to perform voice timbre conversion from unpaired data. Voice Conversion strategies are commonly restricted to the use of parallel speech corpora. Our proposition is based on two main concepts: the modeling of the timbre space based on phonetic information and a simple approximation of the cross-covariance of source-target features. The experimental results based on the mentioned strategy in singing-voice data of the VOCALOID synthesizer showed a conversion performance comparable to that obtained by Maximum-Likelihood, thereby allowing us to achieve singer-timbre conversion from real singing performances.

## 1. INTRODUCTION

One of the main limitations of current Voice Conversion technologies are the use of a *parallel* corpora of the source and the target speakers to perform training of a conversion model. This corpus consists of a set of recordings in which both speakers pronounce the same utterances (same phonetic content) without applying any distinctive emotion or vocal quality. The acquisition of such parallel data may represent a number of difficulties, especially if aiming to apply it on target *voices* which are hardly available; for example: past celebrities.

Some proposals have been reported to achieve non-parallel conversion based on the alignment of originally unpaired data by exhaustive similarity search or the adaptation of an original parallel model. An approach following the latter concept [1], is based on the assumption of a linear relation between the timbre features of originally *paired* speakers and *unpaired* ones. A conversion model trained from paired data is adapted accordingly; however, the source-to-target mapping is not defined directly from the unpaired data.

Previously, the authors introduced a strategy to derive the timbre conversion model exclusively from unpaired data considering that the phonetic segmentation is available [2]. The proposition consists of a modification of the original Gaussian Mixture Model(GMM) based approach of [3] and [4] by applying phoneme-constrained modeling of the timbre space and an approximation of the joint-statistics following the same assumption considered in [1]. In terms of spectral conversion error, the conversion performance was found comparable to that obtained by parallel training without perceiving a significant reduction of the conversion effect on the converted signals.

In this work we extend the study of the proposition presented in [2]. In particular, we are interested in clarifying issues as the learning conditions of the phoneme-constrained modeling and the performance of the proposed non-parallel approach when the nature of the source target corpora differs. We remark on our interest in applying this technology to the concatenative singing-voice synthesizer VOCALOID [5] in order to perform singer-timbre conversion on the system databases by exclusively using real performances from target singers. According to the work presented in [6], the experimental study was carried out on full-quality singing-voice data ($Sr = 44.1KHz$). However, the proposal presented in this work may represent a generalized solution for Voice Conversion purposes.

This paper is structured as follows: the phoneme-constrained Multi Gaussian Modeling is presented in section 2, in section 3 we show study of simple strategy to approximate the source-target cross-covariance, the experimental framework of our study is described in section 4, to evaluate the performance of the proposed method and compare it with the one based on ML, the results of objective and subjective evaluations are reported and discussed in section 5, and the paper concludes with observations and proposition for further study in section 6.

## 2. PHONEME-BASED ENVELOPE MAPPING

### 2.1. GMM-ML for features conversion

The conversion of the voice timbre is commonly achieved by modification of the short-term spectral envelope information based on a probabilistic time-continuous transformation function [3]. The conversion function is commonly derived from a Gaussian Mixture Model of joint timbre features trained in a ML basis. The timbre features correspond to all-pole based estimations of the spectral envelope parameterized as Line Spectral Frequencies (LSF) [4]. We remind, for clarity, the main expressions followed on this strategy

$$\hat{y} = \sum_{q=1}^{Q} p(q|x) \; [\mu_q^y + \Sigma_q^{yx} \Sigma_q^{xx\,-1} (x - \mu_q^x)] \tag{1}$$

$$p(q|x) = \frac{\mathcal{N}(x; \mu_q^x; \Sigma_q^{xx})}{\sum_{q=1}^{Q} \mathcal{N}(x; \mu_q^x; \Sigma_q^{xx})} \tag{2}$$

Eq.1 depicts the conversion function, denoting $x$,$y$ and $\hat{y}$ the source, target and converted envelope features respectively. The GMM size (number of Gaussian components) is given by $Q$. Note that an *a priori* weighting of the mixture components is not considered. The term $p(q|x)$ corresponds to the conditional probability or *class membership*, according to Eq.2.

In general, concerning the configuration of the GMM, the number of Gaussian components depends on the amount of training data as well as the form of the covariance matrices (full or diagonal). Normally, an eight-sized GMM with full covariance matrices is in use to achieve learning generalization for voice conversion purposes [3]. Commonly, the resulting Gaussian means, when translated to the spectral domain, depict spectral envelopes with formantic features. This is principally due to the restriction of using only voiced speech and the significant amount of vocalic content on the data. Note, however, that a one-to-one correspondence cannot be straightforwardly stated between those envelope patterns and the vocalic phonetic classes assumed to be contained in the data.

The vocalic speech is widely considered as provider for the most important perceptual cues for timbre identification. However, they represent only a subset of the phonetic elements of a language. Subsequently, we claim that if aiming to perform full timbre conversion we might map the envelope characteristics regardless, in general, of their vocalic or voiced nature. Accordingly, a clustering of the envelope space by only eight gaussian components may lead to a large averaging of the phonetic content. Note also the highly competitive behavior observed on the ML-based mixture, resulting in a full modeling of speech segments of different phonetic nature by the same Gaussian distribution. These phenomena lead to a significant simplification of the phonetic space on the mapping process and are found at the origin of some "reduction" or modification of the phonetic content perceived in some converted utterances.

### 2.2. Phoneme-constrained Multi-Gaussian Model

Moreover, by means of setting the GMM size close to the assumed number of phonetic events and restricting the covariance matrices to be diagonal, the behavior of the mixture was found to be more cooperative but unstable. We show in Fig.1 the resulting component-to-phoneme correspondence for a GMM-ML in terms of the average membership of each gaussian per phonetic class. The results were obtained by evaluating $p(q|x)$ after training the GMM with labeled data. The vertical axis represents the GMM components whereas the horizontal axis lists the phonemes included in VOCALOID according to the Japanese language (SAMPA standard) ordered by phonetic group (vowels, nasals, voiced plosives, voiced affricates, liquids, semivowels, unvoiced plossives, fricatives, unvoiced affricates).

Clearly, following Fig.1, relationships between the clustering achieved by the GMM-ML and the phonetic class of the features can hardly be established. An unstable activation of the mixture components along with phonetic content may produce irregular evolution of the converted envelopes, representing a potential factor of degradations on the converted signals.

Consequently, we propose to control the fitting of the statistical model by using the phonetic information; therefore, we restrict the computation of each Gaussian distribution to the data corresponding to a same phoneme. A phoneme-based modeling (pho-GMM) was already introduced in [7], showing some benefits in terms of one-to-many mapping reduction compared to conventional GMM-ML.

Following this strategy the resulting component-to-phoneme correspondence is clearly increased [2], as shown in Fig.2. The model was therefore able to extract characteristic information for most of the phonemes, and to increase, consequently, the discrimination between them.

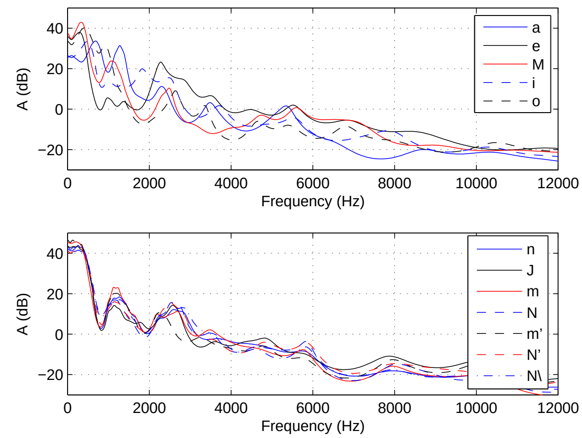


Figure 3: Corresponding spectral envelopes of the MGM means within phonetic groups. Vowels (top), nasals (bottom).

Note however some "shared" regions on the grid within elements of a same phonetic group (e.g. nassals, plosives). Unlike the case of the vowels, where the differences between the formantic structures represent an important discriminaton factor, the average spectral patterns at these groups are relatively close. This can be appreciated in Fig.3, where are shown the resulting envelope patterns of the vowels (top) and nasals (bottom) sets. Although we have not a theoretical basis to explain these similiarities, a further simplifaction of the phonetic space and the role of such a "characteristic envelope" on non-stationary phonemes (e.g. plosives) may be studied.

Finally, keeping consideration that the phonetic information is available, the conditional probability can be replaced by a phonetic flag to directly assign the corresponding component at the conversion stage. However, this "forced" membership should be smoothed at the phonetic boundaries to avoid abrupt changes when transforming the signal. As was already described, by forcing a full-competitive behavior we do not significantly differ from the real role of $p(q|x)$ observed in a GMM-ML. Moreover, following this proposition we aim to refine the envelope mapping in a phonetic basis. Note however that, as comented in [7], without including more meaningful context information some mapping losses can be hardly alleviated if the acoustic characteristics of same-phoneme data significantly differs. This is demonstrated further in our experimentation by using data of increasing heterogeneity.

Accordingly, the original conversion function expressed in Eq. 1 is modified as

$$\hat{y} = \mu^{y}_{q(x)} + \Sigma^{yx}_{q(x)} {\Sigma^{xx}_{q(x)}}^{-1} \left[x - \mu^{x}_{q(x)}\right] \quad (3)$$

Moreover, the sub-index $q(x)$, denotes the phonetic class of the source input and therefore, defines the only gaussian component involved in the mapping. Subsequently, since the resulting model does not keep the "mixture" characteristic anymore, we refer to it as a "Multi-Gaussian Model" (MGM) [2].

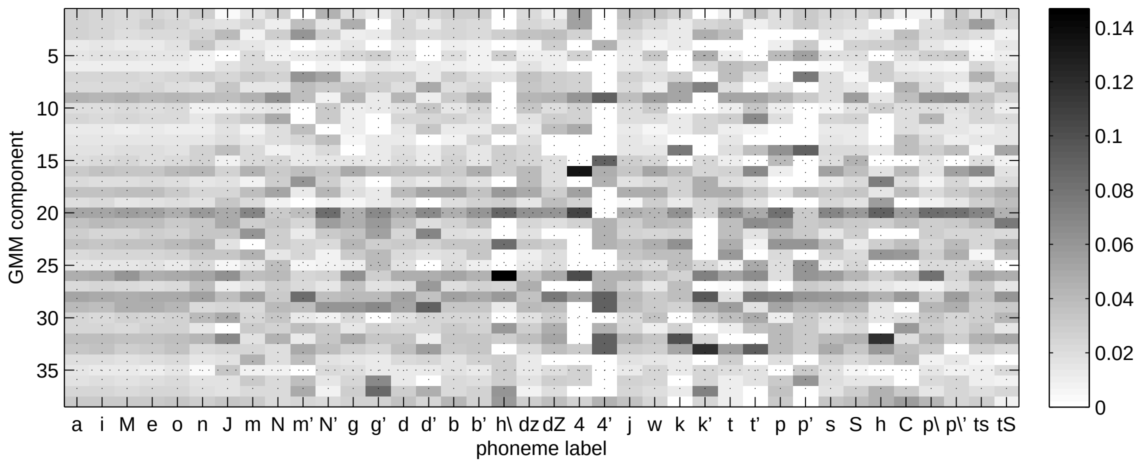


Figure 1: Average conditional probability at each GMM component per phonetic class. ML-based fitting.

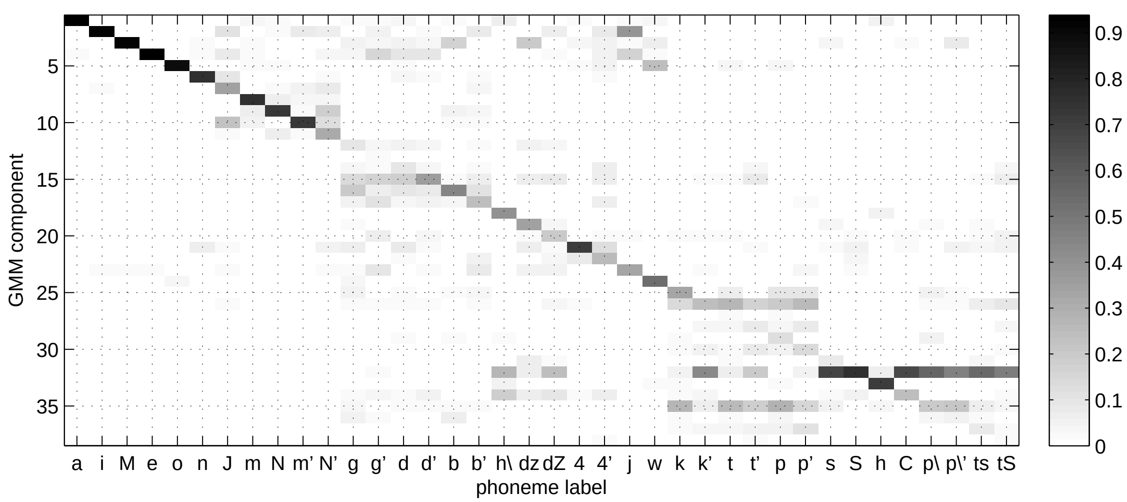


Figure 2: Average conditional probability at each GMM component per phonetic class. Phoneme-based fitting.

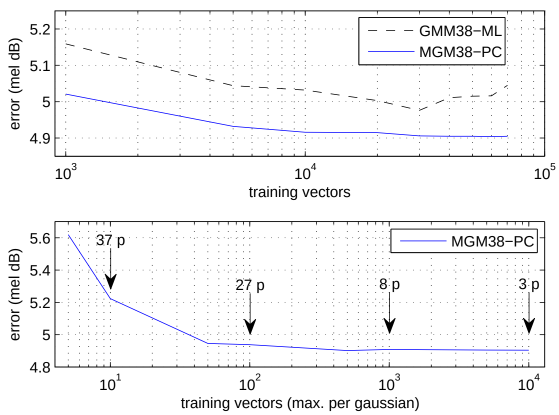


Figure 4: GMM-ML and pho-MGM conversion error by overall training size (top). pho-MGM Error by maximum training size per component (bottom). The error measure (mel dB) coresponds to the spectral distortion averaged over mel-scaled spectra.

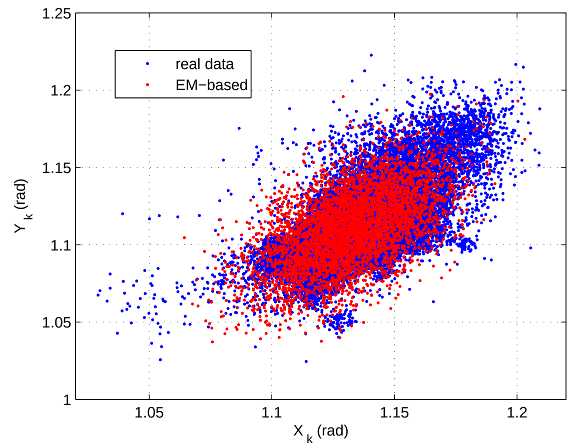


Figure 5: Example of LSF data within a phoneme-class (one-dimension). Real data (blue) and generated from the resulting ML-based Gaussian distribution (red).

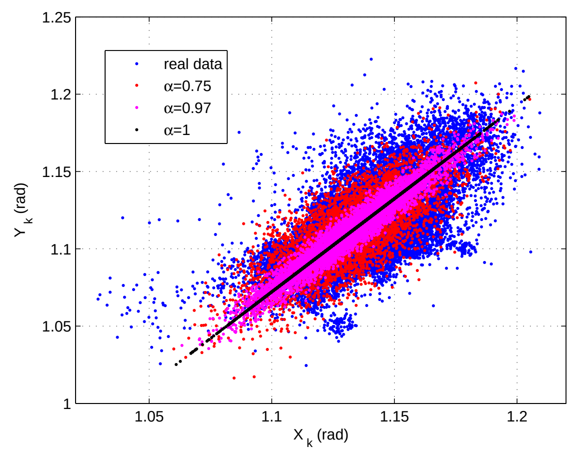


Figure 6: Example of LSF data within a phoneme-class (one-dimension). Real data (blue) and generated from the approximated statistics for variable $\alpha$ (black, magenta, red).

### 2.3. MGM performance and training

We intended to study the minimal amount of data per phoneme required to generalize the timbre mapping. However, the amount of frames per phoneme can barely be equilibrated since the vocalic content is predominant in the Japanese language. Thus, we limit our data size control to an upper bound, or maximal training size, for the number of frames of a same phoneme used to fit a Gaussian component. A regularization of the covariance matrices was required for phonemes from which only a small amount of data was available.

The results are shown in Fig. 4. The timbre features correspond to LSF parameters of accurate spectral envelope estimates obtained by a mel-based autoregressive model [6] with envelope order set to 50. The cost function corresponds to the spectral conversion error between the converted envelopes and the target spectra. We compared GMM-ML and MGM models with similar complexity (diagonal matrices, 38 components). In general, the resulting conversion error levels are similar (top graph), showing the MGM with slightly increased performance. An over-fitting effect was not found to be affecting, though a small training set was used (1000 vectors). We remark that the conversion performance was always evaluated on unknown data (test set) in order to observe the stabilization of the learning; that explains the decreasing behavior of the error curve.

The maximal amount of training data per MGM component was also evaluated (bottom graph). The arrows denote the number of phonetic classes reaching the corresponding maximal number of vectors at each case. The results show that it is not necessary to have a large amount of frames (around 100) to approach the high performance region.

## 3. CROSS-COVARIANCE APPROXIMATION

### 3.1. Motivation

From eq. 3 we remark that the only term for which paired data is required is the source-target cross-covariance ($\Sigma^{yx}$). By simplifying the proposition of [1] via assuming directly a linear transformation between the source and target features their joint statistics can be approximated. Moreover, the phoneme-constrained modeling presented in the past section limits this term, for each Gaussian distribution, to depend exclusively on the data of the corresponding phonetic class.

According to eq. 3, the term $\Sigma^{yx}$, commonly called *transformation matrix* after normalization by the source variance, acts actually as a weight of the variance of the converted features. The values observed on this term on the GMM-ML based models are rather small, resulting in poor dynamics of the converted features. This well-known and characteristic over-smoothing, already addressed in works [8], is commonly perceived as a *muffling* quality, affecting the naturalness of the converted signals.

Notably, an augmentation of the variance of the oversmoothed converted parameters has been found to reduce significantly this *muffling* effect. Therefore, we assert that this term when estimated by ML represents a limitation on the resulting conversion quality. Furthermore, having control of this value might represent an effective way to increase the naturalness of the converted signals.

### 3.2. Covariance approximation by linear transformation

Following the phoneme-constrained modeling, the probabilistic linear transformation between the timbre features of two speakers proposed in [1] can be simplified as $y = A_{q(x)}x + b_{q(x)}$, where $A_q$ is, in general, a square matrix according to the dimensionality of $x$, and $b_q$ is a bias vector. Therefore, considering the mentioned relation in the computation of $\Sigma^{yx}$ for each phonetic-component of the MGM we obtain

$$\check{\Sigma}^{yx} = E[(\check{y} - \mu^{\check{y}})(x - \mu^{x})] \tag{4}$$

$$= E\{[(Ax + b) - (A\mu^{x} + b)]\ (x - \mu^{x})\} \tag{5}$$

$$= E[(A(x - \mu^{x})^2] = A\Sigma^{xx} \tag{6}$$

Where $A$ can be approximated similarly by evaluating $\Sigma^{yy}$

$$\check{\Sigma}^{yy} = E\{[(Ax + b) - (A\mu^{x} + b)]^2\} \tag{7}$$

$$= E[(A^2(x - \mu^{x})^2] = A^2\Sigma^{xx} \tag{8}$$

$$A = \sqrt{\Sigma^{yy}\Sigma^{xx^{-1}}} \tag{9}$$

Although the relation $y = Ax + b$ is assumed between features corresponding to the same phoneme imposes a strong assumption and, by using diagonal covariance matrices, the resulting one-dimensional distributions restricts to narrow regions. As the norm of $A$ decreases, the "width" of the covariance region increases until it reaches a circular form at the full-uncorrelated case ($A = 0$). Thus, since the orientation of the modeled distribution is given exclusively by $\Sigma^{xx}$ and $\Sigma^{yy}$ the proposed $\check{\Sigma}^{yx}$ may be rather seen as a lower bound of the real distribution width. Accordingly, we apply a weighting factor ($0 < \alpha < 1$) to $\check{\Sigma}^{yx}$ on the conversion function in order to impose a more realistic form on the approximated distribution.

In Fig. 6 we show a comparison of real and approximated source-target distributions for several $\alpha$ values of one LSF dimension within a phonetic class. Clearly, the distribution strictly following the relation y = Ax+b ($\alpha = 1$) does not suffice the data. However, by setting $\alpha$ around 0.75, the covariance region approaches the covariance based on ML. This can be seen in Fig. 5, illustrating the case when the Gaussian is fitted in a ML basis.

Then, based on eq. 3, the final expression for the conversion features will be as follows

$$\hat{y} = \mu^{y}_{q(x)} + \alpha\ \sqrt{\Sigma^{yy}_{q(x)}\Sigma^{xx}_{q(x)}{}^{-1}}\ \ [x - \mu^{x}_{q(x)}] \tag{10}$$

Regarding the effect of the parameter $\alpha$ on the conversion performance, values within the range [0.5-0.7] provide the best perceptual results. Further, we observe that, for the dimensions with a low correlation, the imposition of a covariance value higher than the real one was found to be beneficial. The naturalness of the converted signals is improved by increasing the dynamics of the predicted LSFs. Nevertheless, for clarity an objective and subjective evaluation of $\alpha$ is presented in Section 5.

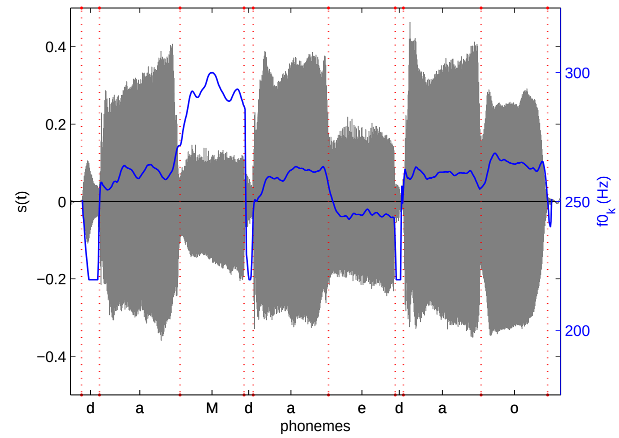


Figure 7: Example of a singing sample of a VOCALOID DB including the phonetic segmentation and F0 estimation.

## 4. EXPERIMENTAL FRAMEWORK

### 4.1. VOCALOID singer databases

The VOCALOID singing-voice synthesizer system consists of three main elements, the user interface (allowing to input lyrics and melody information), a singer database (containing a collection of singing-voice samples from a singer), and the synthesis engine (performing, briefly, the selection, F0-transposition and concatenation of the samples).

In particular, the singer-DB consists of a pre-defined set of phonetic sequences sung at different pitch ranges. The phonetic scripts are assumed to cover principally the consonant-vowel combinations of the Japanese language. All the singing samples are recorded at a same tempo following the same melodic pattern, which is restricted to a one tone variation related to the representative musical height of each pitch set. An example of a singing sample is shown in Fig. 7. Each single-pitch set consists of more than 100 recordings, representing more than 70,000 feature vectors. Typically, a complete VOCALOID-DB includes low, medium, and a high pitch sets.

Singing-voice data from 2 VOCALOID singer-DBs were considered for our experimental study. A C4 ( $261Hz$) pitch set of a female singer was set as source voice whereas G3, C4 and E4 pitch sets $(193, 261, 330Hz)$ as well as 4 real singing performances from a male singer were used as target voice. The configuration of the target data was modified according to the interest of comparing both the mapping strategy and the effect of the pitch on the conversion performance, and is described in the next section.

### 4.2. Effect of the corpora heterogeneity

There is advantage, in terms of envelope mapping performance, to using data which is not only paired but restricted to a small pitch variation [6]. Accordingly, the use of non-parallel corpora may have an impact on the conversion performance if the nature of the source and target corpora differs. We remark that one of our main interests in applying non-parallel timbre conversion on the singing-voice is to use real singing performances to compute the

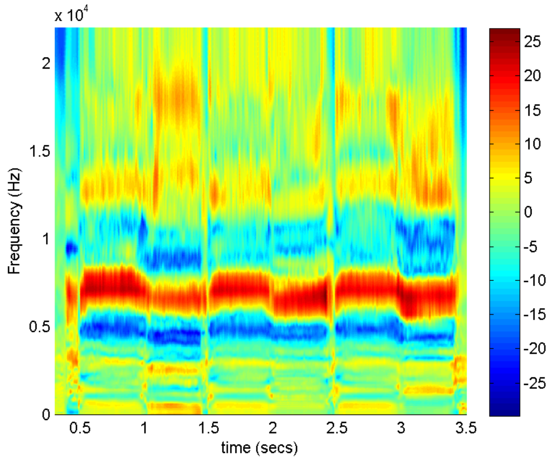


Figure 8: Spectrogram of the conversion filter for the sample of Fig. 7 (energy in dBs).

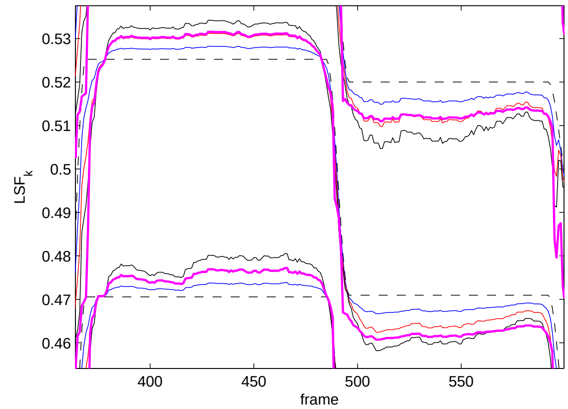


Figure 9: Converted LSF parameters given by $\alpha = 0$ (dotted), $\alpha = 0.25$ (blue), $\alpha = 0.5$ (red), $\alpha = 0.75$ (black) and ML-based conversion (magenta).

conversion model, which may observe a large pitch range (melody), different tempo, and a rich pronunciation variety among the content.

We were therefore interested in studying the performance using data with different characteristics. Therefore, we fixed three different sets as target data: a)VOCALOID's single-pitch data, b) VOCALOID's mixed-pitch data and c) real singing performances (4 songs). Further, we considered the following evaluation cases: a)GMM-ML single-pitch (labeled as GMM-ML P SP), b)MGM single-pitch (MGM-PC P SP), c)non-parallel single-pitch (MGM-PC NP SP), d)non-parallel mixed-pitch (MGM-PC NP MP), and e)non-parallel real songs performances (MGM-PC NP RP). Since objective evaluation on unpaired data is not straightforward, all the approaches were evaluated on the single-pitch paired data i.e., different sets for training but same ones for evaluation.

An evaluation of this nature represents the most exigent case since the single-pitch set observes the most precise and "homogeneous" features, resulting in an increased challenge for the models trained on data corresponding to wider pitch-range and "heterogeneous" phonation characteristics (multi-pitch and real singing performances sets).

### 4.3. Signal modification

As was already described, the timbre conversion process is based in a short-term mapping of the spectral envelope information. The transformation of the timbre is therefore achieved by replacing the original envelope by the one given by the converted features. This is commonly done by analysis-synthesis filtering following the autoregressive modeling of the envelope. However, we perform the modification of the envelope by defining a *conversion filter*, corresponding to the difference at each frame between the corresponding transfer function of the converted LSFs and an interpolation of a harmonic analysis of the source signal. The frame processing is done in a pitch-synchronous basis. We show in Fig. 8, in the form of a spectrogram, an example of resulting conversion filter for the utterance of Fig. 7.

We use the harmonic information instead of the envelope on the source signal aiming to match closely the real source information and discard the risk of estimation errors that occurred during the computation of the autoregressive model. The harmonic analysis and the processing framework itself follow the wide-band technique described in [9].

Besides the capability of the processing method to perform efficient envelope modification, the conversion itself may result in some unnatural or distorted quality on the transformed signals since the characteristics of the converted spectra may not match naturally the original signal (harmonicity, energy, f0). Also, consider that some abrupt changes on the evolution of the source signal cannot be properly reflected by the mapping process.

Moreover, the stable and controlled characteristics of the singing samples might impact positively the conversion quality if compared to the case of spontaneous speech. However, the particular evolution of the source signal and the use of wide-band based information may result in an important variation of the envelope information at successive frames. Accordingly, we consider two parameters to control independently the smoothness of the conversion filter for both time and frequency axes. Although it is not generally required, this strategy was found effective to avoid degradations in some converted utterances and to smooth undesired frame-to-frame variations on the conversion filter.

## 5. EVALUATION

### 5.1. Objective evaluation

We were interested on studying three aspects in our experimental evaluation: first, the impact of the covariance approximation on the converted features, second, to compare the conversion performance of the parallel and non-parallel strategies, and finally to evaluate the effect of the heterogeneity of the target data.

We therefore started analyzing the converted LSFs for different $\alpha$ values. Note the benefits of using this parameterization, seen as temporal trajectories denoting spectral pole locations, to observe differences in terms of variance. This can be seen in Fig. 9. The plot shows a comparison of three converted LSFs at a segment

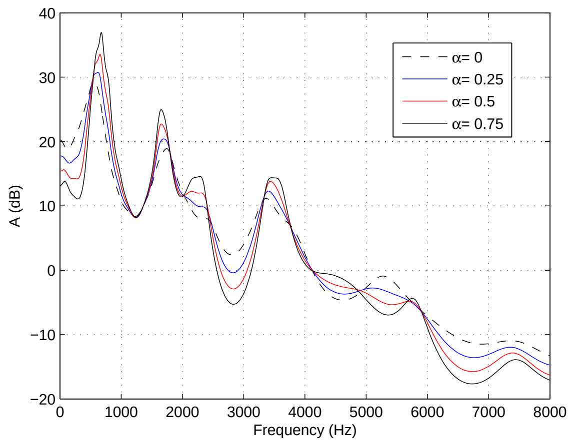


Figure 10: Resulting spectral envelopes from converted LSF parameters for different $\alpha$ values.

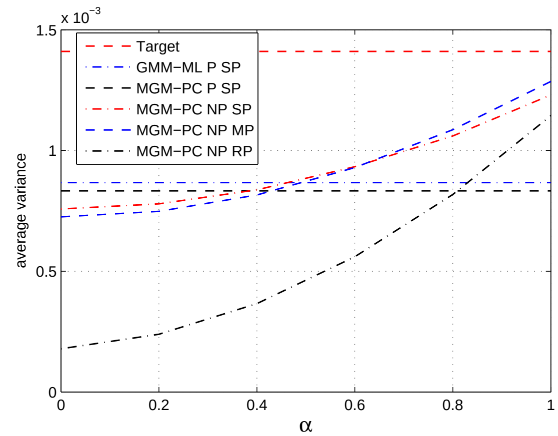


Figure 11: Average variance of target and converted LSFs for the different method and data confirgurations.

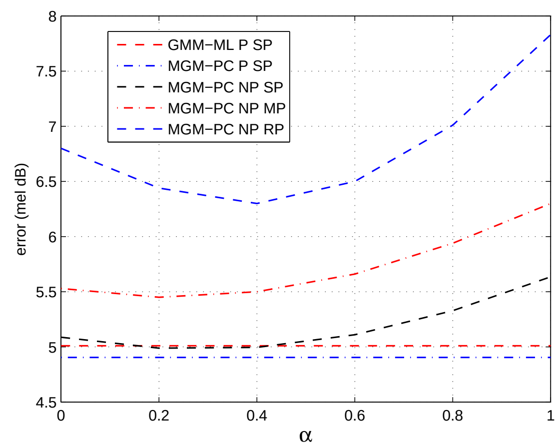


Figure 12: Spectral conversion error for the different method and data confirgurations.

of the utterance example. The different cases correspond to conversions obtained by using increasing $\alpha$ values as well as a ML result issued from a model with similar complexity. As expected, an augmentation of this value was found to increase the temporal variance related to the means position ($\alpha = 0$). The corresponding effect in the spectrum is found as an emphasis of the maxima and minima of energy, as shown in Fig. 10, producing a positive effect, within a reasonable limit, in the perceived naturalness.

Fig. 11 shows the average variance measured on about 5000 evaluation vectors (test set) of target and predicted LSF for the different evaluation cases as described in section 4.2. Note that the resulting variances of the parallel cases are just slightly higher than those given by the gaussian means ($\alpha = 0$), denoting the poor impact of the transformation matrix when it is exclusively derived from the data (whether or not the variance is obtained by ML). On the other hand, note that by setting $\alpha$ we can force a variance on the converted features close to the real one.

Finally, Fig. 12 depicts a conversion performance comparison. The cases involving models trained on single-pitch data (labels ending with "SP") are considered as the references since the training data corresponds to similar corpora with stable characteristics. The proposed non-parallel conversion using single-pitch data performs close of ML and MGM-parallel cases. As expected, the performance decreases as the target data is more heterogeneous. Moreover, the conversion performance shows a maximum related to $\alpha$; however, slightly higher values ([0.5-0.7]) have been found as providing increased naturalness (muffled quality reduction).

### 5.2. Subjective evaluation

A subjective evaluation was designed aiming to compare the conversion effect and the quality of the converted signals. Looking for a strict perceptual evaluation, ten sound technology professionals participated as listeners. Five VOCALOID samples, representative of the different phonetic groups, were selected as the evaluation set.

First, a timbre similarity test was defined considering three conversion strategies: GMM ML SP, MGM NP SP and MGM NP RP. We intended to compare the conversion effect when using both parallel and non-parallel methods and the effect of using a homogenous or heterogenous target corpus on the proposed non-parallel method. The procedure was as follows: the source, target, and the three converted utterances (randomly selected) were presented to the listeners. Then, the timbre similarity between the converted and the reference samples was measured according to the continuous range [0 1] (0 = source, 1 = target). This process was repeated immediately to allow confirmation or modification of the first judgement.

Note that at each sample case, both reference and converted utterances observe stable and similar acoustic characteristics (pitch, energy, and dynamics). A comparison based on this data appears to be an efficient way to exclusively focus on the timbre and vocal quality differences. However, the conversion of vocal quality features is out of the scope of this work. We claim that although an increased perceptual discrimination capacity may result in lower

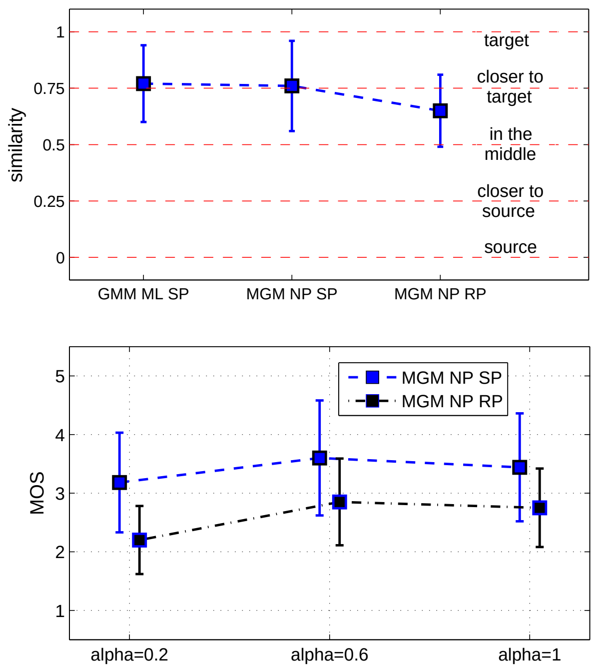


Figure 13: Subjective evaluation. Timbre similarity results (top) according to given reference levels. Signal quality evaluation (MOS) of the non-parallel conversion for different $\alpha$ values.

conversion scores it might lead us to a more robust evaluation of the effect achieved by the spectral envelope conversion process.

The results are shown in Fig. 13 (top). In the figure the five levels in red tagged with a subjective description correspond to the scale references given to the listeners. The scores achieved by both parallel and non-parallel methods when using the same corpora were found similar and denote, in general, a reasonable conversion effect. However, the performance suffers some reduction when real singing corpora is used as target training data.

Second, a MOS-like test was focused on exclusively evaluating the signal quality of the non-parallel conversion for different $\alpha$ values. The test was applied separately for both corpora cases in order to observe exclusively the effect of $\alpha$. The subjective description of the MOS levels was re-defined looking for an exigent evaluation and an association of the measurement levels with some quality phenomena (5=perfect, 4=slight degradations, 3=clean enough, 2=artifact(s), 1=annoying).

The test followed a similar procedure as the similarity test. For each sample case three converted samples, corresponding to three representative $\alpha$ values (small, $\alpha$=0.2; proposed, $\alpha$=0.6; large, $\alpha$=1), were randomly selected and evaluated in the same two-step basis. The results are shown in Fig. 13 (bottom). As expected, best results were found for $\alpha$=0.6. Note however that a large value achieved a comparable performance. This might be explained by a reduced risk of producing undesired amplitude modulations on the spectrum when aplying a high variance to the LSF trajectories on stable signals.

Although the overall results does not allow us to claim full natural-quality conversion the scores achieved when using similar and stable corpora show a general perception of an adequate naturaleness. As for the similarity test, the drop in the performance level is attribuited to the increased heterogeneity of the target corpora, resulting in over-smoothed envelope patterns on the conversion model. The estimation of precise envelope information from singing performances might be studied further.

## 6. CONCLUSIONS AND FUTURE WORK

In this work we presented an approach to perform voice timbre-conversion from non-parallel data. The proposed strategy is based on phoneme-constrained modeling of the statistical space of the timbre features and an approximation of the cross-covariance information and is described and compared with the conventional approach based on parallel data and ML. The results, obtained from an experimental study on singing-voice let us claim the achievement of comparable conversion performance although some dependency was observed according to the heterogeneity of the corpora.

The experimentation done in this work suggest to extend the study in some issues: the estimation of the $\alpha$ parameter individually for each feature dimension; an efficient selection of envelope features from real singing performances; an efficient mapping of non-stationary phonemes, among others. However, the proposition presented in this work was proved to be a step-forward the interests of voice timbre conversion.

## 7. REFERENCES


[1] A. Mouchtaris, J. Van der Spiegel, and P . Mueller, "Non-parallel training for voice conversion based on a parameter adaptation approach," *IEEE-TASLP*, vol. 14, no. 2, pp. 952–963, 2006.

[2] F. Villavicencio and H. Kenmochi, "Ressurecting past singers: Non-parallel singing-voice conversion," in *1th International Workshop on Singing-Voice InterSinging*, 2010.

[3] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE-TASAP*, vol. 6, no. 2, pp. 131–142, 1998.

[4] A. Kain, *High-Resolution Voice Transformation*, Phd. thesis, Oregon Institute of Science and Technology, October 2001.

[5] H. Kenmochi and H. Oshita, "Vocaloid commercial singing synthesizer based on sample concatenation," in *Proc. of INTERSPEECH'07*, 2007.

[6] F. Villavicencio and J. Bonada, "Applying voice conversion to concatenative singing-voice synthesis," in *Proc. of INTERSPEECH'10*, 2010, vol. 1.

[7] E. Godoy, O. Rosec, and X. Chonavel, "Alleviating the one-to-many mapping problem in voice conversion with context-dependent modeling," in *proc. of INTERSPEECH'09*, Brighton, UK., 2009.

[8] T. Toda, A.W. Black, and Tokuda, "Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter," in *Proceedings of IEEE-ICASSP '05*, 2005, pp. I–9–I–12.

[9] J. Bonada, "Wide-band harmonic sinusoidal modeling," in *International Conference on Digital Audio Effects, DAFx'08*, Helsinki, Finland, 2008.