GMM-PCA based Speaker-Timbre Conversion on Full-Quality Speech

Fernando Villavicencio, Esteban Maestre

Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain
villavicencio@upf.edu, esteban.maestre@upf.edu,

Abstract
This work addresses a study of the GMM-based approach to achieve full-quality speaker timbre conversion. In general, high-quality voice conversion requires accurate spectral envelope estimates, resulting in high-dimensional feature vectors and relatively high computational. Aiming to achieve low-dimensional processing, accurate envelope estimates of the speakers are mel-frequency scaled and projected onto the space defined by a subset of the principal components. The GMM-based features conversion is then performed in the reduced space. Our experimental findings confirm that this strategy provides benefits, especially observed on the resulting converted speech quality, with a significant computational cost reduction.

Index Terms: Speech synthesis, speech analysis, linear prediction, pattern recognition

1. Introduction
Voice conversion (VC) commonly relies on mapping the timbre features of a source speaker onto that of a target one. Timbre information is considered to be contained in the short term spectral envelope of the speech signal. Early approaches to voice conversion performed envelope estimation by means of classical techniques such as cepstral analysis and linear prediction. These techniques provide estimates of the main features of the envelope in a space of relatively low dimensionality. However, accurate spectral envelope estimates are required to achieve high-quality voice conversion [1]. Timbre features corresponding to autoregressive models are usually parameterized in the form of Line Spectral Frequencies (LSFs) [2]. Depending on the sample-rate of the speech signal and the pitch of the speakers, precise envelope estimation leads to a significant increase of the dimensionality of the resulting feature vectors. Typically, the timbre transformation function in current voice conversion frameworks is based on a Gaussian Mixture Model (GMM) used for statistically mapping the timbre features of a source onto those of a target speaker [3]. Voice conversion frameworks have been mostly limited to low and medium quality speech signals. The resulting converted speech quality is not always found to be satisfactory or artifact-free. No attempt has been made to perform full-quality Voice Conversion [4].

Our recent work [1] presents the benefits of performing accurate spectral envelope modeling to carry out timbre conversion on higher-quality speech signals. Improved envelope models have been reported to contribute to an increase of both the timbre conversion performance and the perceived quality as compared to linear prediction. These approaches require, however, accurate spectral envelope estimation, resulting in high-dimensional vectors, principally due to a dependence of a proper model order selection on the sample-rate of the speech signals and the average fundamental frequency of the speakers. Despite the benefits provided by increasing the detail level of the spectral envelope description, both the complexity and the computational cost of the statistical model increase significantly with the dimensionality of the features.

In this work we evaluate the GMM-based spectral conversion framework for timbre conversion on full-quality speech. The performance of a novel mel-frequency based autoregressive (AR) model for spectral envelope estimation is studied. We also consider the use of principal components analysis (PCA) to consistently reduce the dimensionality of the timbre features in order to increase the clustering capacity of the GMM. In our approach, the source-target mapping is defined in the space given by a subset of the principal components of the features of the given speakers. The benefits of this proposition are addressed by the results of an experimental work carried out to objectively and subjectively evaluate the timbre conversion performance.

The paper is structured as follows. Section 2 address the modeling of the timbre information, based on an autoregressive approximation of the spectral envelope. The experimental framework of our study is described at Section 3. In section 4, the conventional GMM approach is applied on full-quality speech followed by a proposed methodology based on PCA in Section 5. The results of objective and subjective evaluations are presented in Section 6. Finally, conclusions are presented in Section 7.

2. Timbre modeling
2.1. Improved AR modeling
Voice Conversion systems have been generally restricted to low-medium quality speech (8 – 16 kHz). Clearly, this limits their potential of applicability if aiming to achieve high-quality audio processing. In [1] the extension of the GMM-based approach to higher quality speech (24kHz) was proposed. This method is centered on the use of accurate spectral envelope models to improve the timbre conversion performance, partially limited by errors when estimating the envelope information. Considering the spectral envelope as a smooth function passing by the prominent peaks of the spectrum, the typical estimation techniques found in most VC strategies (e.g., Linear Prediction, MFCC) do not provide accurate envelope information. By applying improved envelope modeling an increasing conversion effect and naturalness on the converted speech was observed [1].

Considering the well-known advantages of a representation based on LSF for envelope modification proposals, we followed the idea of [4] to fit an all-pole model on a spline interpolation of the spectral peaks to perform an estimation of the spectral envelope. The resulting autoregressive filter is parameterized in terms of Line Spectral Frequencies to represent the timbre features of the speakers. The spectral analysis is obtained by...
The means of Wide-Band harmonic sinusoidal modeling (WB) as described in [5], leading to a reliable estimation of the spectral peaks. Then, a spline interpolation yields the smoothed envelope for fitting the final all-pole structure. Accordingly, as the order model is increased, the resulting autoregressive model more closely fits the underlying spectral peaks interpolation.

2.2. melAR modeling

The benefits of using a mel-scale resolution on the frequency axis to perceptually emphasize the spectral information are well-known. Moreover, when increasing the order of AR-based envelope estimations of high-quality speech signals, a significant number of LSFs falling in the high frequency region show noise-like behavior. Typically, this will result in significant oversmoothed features after the conversion based on statistical modeling. The information represented at these frequency regions is not assumed to be strongly correlated with the voice timbre. Accordingly, as presented in [6] for True-Envelope based modeling [7], we apply mel-scaling on the frequency axis of the peaks interpolation before fit the AR model.

As a result, for the same model order, the number of noisy LSFs is significantly reduced. This is shown in Fig. 1 where the LSF temporal trajectories estimated on a sentence are plotted using both linear and mel frequency scales. It can be appreciated that for the linear case, most of the LSFs exhibit noisy behavior, whereas this effect is reduced for the mel-scaled case. We will refer to the linear and mel-scaled models as AR and melAR, respectively. Note that, unlike our melAR model, mel-scale based Warped Linear Prediction [8] includes the harmonic structure on the resulting envelope. This characteristic is not suitable for envelope modification.

Denoting $\tilde{X}(\omega_p)$ as the resulting WB-based harmonic analysis (size $P$), $E$ as the operator yielding the spline interpolation (size $K$), and $\omega_k = F_{\text{mel}}(\omega)$ as the mel-scaled frequencies, we derive the melAR model as

$$ E: \tilde{X}(\omega_p) \rightarrow E(\omega_k) $$

(1)

$$ R_{E}(i) = \frac{1}{K} \sum_{k=1}^{K} E(\omega_k)^2 e^{j\omega_k i} $$

(2)

The value of $K$ should be set large relative to $P$, resulting in increased smoothing of the feature interpolation for larger $K$. In our case we used $K = 2^{(N+1)}$ such that $2^N$ is the smallest value greater than the number of peaks. Note that, as long as the resulting interpolation matches the given spectral points (Fig. 2, top), a significant conversion performance difference following this criterion compared to the use of estimates obtained by optimal order selection on the True-Envelope estimator [6] was not observed. Moreover, in a similar way for such an estimator, $K$ can be fixed considering a theoretical maximal bound for $P$ (considered as the expected maximal number of partials) given the smallest pitch value observed on the speaker

$$ P_{\text{max}} = \frac{\delta_f/2}{f_{\text{min}}} $$

(3)

Finally, the autocorrelation function $R_{E}(i)$ of the mel-frequency scaled envelope $E(\omega_k)$ is obtained by IDFT as shown in eq. 2 and used in the Yule-Walker equations system to compute the final autoregressive model. In Fig 2 an example of the resulting melAR estimation is shown.

3. Experimental framework

The conversion performance was evaluated through 10-fold cross-validation of a database of 200 utterances recorded from two male speakers in French language, sampled at 44100Hz. We only considered a single gender case to avoid the influence of the pitch information in order to exclusively evaluate the timbre conversion performance. Also, note that the fidelity of the approximation of the underlying transfer function of the signal in terms of the envelope estimation is limited by the number of support points (spectral peaks). Accordingly, the use of high-pitched signals may not allow us to clarify whether or not the conversion performance is affected by the missing timbre information. Accordingly, the use of male speech (low pitched) appears to be a reasonable choice for our study.

The speech modification stage consisted of a pitch synchronous application of a filter corresponding to the difference between the converted and the source envelopes. As for the envelope estimation stage, the frame processing was based
4. High-dimensional GMM conversion

4.1. Envelope modeling performance

The necessity to perform accurate envelope estimations to perform high-quality timbre conversion and the increased precision of the AR modeling for increased order has been addressed. Considering the extension proposed in this work to full-quality spectra, we were firstly interested in studying the envelope precision that can be captured on the statistical mapping by using both AR models and its effect on the learning conditions of the GMM. Nevertheless, following our experimentation, converted speech based on melAR modeling was perceived as more natural, exhibiting clearly better time-domain evolution. This fact confirmed the aforementioned benefits of applying increased precision on the low-frequency region. We therefore kept melAR modeling as our unique estimation technique.

In Fig. 3 we show the resulting conversion error as a function of the order. The performance measured on the test and training sets are shown in solid and dotted lines, respectively. The error corresponds to the average MSE between the converted spectrum and the target peaks interpolation measured at each frame over the mel-frequency axis. By using this measure we aim to avoid a bias on the conversion error produced by the estimation error between target estimates and the real target spectra as well as to perceptually emphasize the spectral information. As expected, the conversion error decreases with the increasing envelope precision until convergence. The error was also measured in non-overlapping mel-scaled frequency bands to clarify the effect of the increasing precision over the frequency axis. In Fig. 4 we compare the conversion error when using low and high order melAR modeling. Note that the error increases along the frequency axis, confirming the increasing difficulty to model correlations at high-frequency regions.

Following these results and the resulting computational cost, we fixed the envelope order to 70 aiming to keep a reasonable error-cost trade-off.

Figure 3: Conversion error for increasing melAR model order for the test (solid) and training (dotted) sets at several GMM sizes. Full (left) and diagonal (right) covariance matrices.

Figure 4: Conversion error at perceptual bands using low and high envelope estimation precision (melAR model order).

4.2. Learning conditions

In Fig. 3 the results when using full covariance matrices (left plot) for several GMM sizes are shown. The use of a LSF parameterization suggest us to consider full-covariance matrices on the GMM in order to capture the correlation between successive LSF parameters. This was verified by observing a corridor extended along the diagonal on the covariance matrices of the mixture. However, such model complexity (the number of parameters) restricts us to use few components in the mixture to avoid overfitting. Note that the classification behavior of the GMM on this task is commonly found to be highly competitive; the membership of each frame is mainly represented by one single component. Accordingly, a reduction of the size of the mixture will result in the extension of a common membership over a larger phonetic space. As a result in our experiments we perceived an increased smoothing of the phonetic content on the converted speech when using the typical number of components (8) compared to the case of lower quality signals. On the other hand, as result of the overfitting, by increasing the number of components, the resulting quality was progressively degraded.

We therefore used diagonal matrices to reduce the overfitting and increase the size of the mixture. The results are shown in Fig. 3 (right plot). We note that the performance of the full-matrices case was not significantly improved. However, the observed overfitting (the gap between training and test sets error) was found to be smaller despite a substantial augmentation of the number of components compared to the full case (256 instead of 32). Also, by listening to the converted speech, the phonetic smoothing and degradations previously addressed were not detected. Moreover, an increased conversion effect was perceived. These improvements are attributed to the increased sampling of the timbre space by the source-target mapping while keeping a good learning generalization.

By using diagonal matrices, the benefits provided by a progressive augmentation of the number of components (for the values shown) in terms of the perceived quality were obvious. However, the computational cost became prohibitive for the given envelope order (70) using such sized mixtures. In the next section we present a strategy based on principal components analysis to efficiently reduce the features dimensionality and to derive the source-target mapping in a more meaningful
5. Efficient full-quality timbre conversion

5.1. Features dimensionality reduction

As already outlined, a considerable amount of information on the accurate envelope estimates (mostly, values of LSFs in the high frequency region presenting low variance) fails to show a clear correlation with the timbre characteristics of a speaker, despite the benefits naturally brought by the mel-warped of the frequency axis. By carrying out an appropriate analysis of the feature vectors for a given speaker, such noiselessness can be filtered out by conveniently reducing the dimensionality of the space where feature vectors live while keeping a high percentage of information. This is achieved in this work by applying PCA to the feature vectors of a given speaker. Besides a significant computational cost reduction, the projection of the features to a reduced space based on the correlation of the data may provide some benefits in terms of the timbre space clustering.

5.2. GMM-PCA based timbre conversion

To achieve timbre features conversion, the GMM-based conversion framework is applied after reduction of the features of each speaker. The length of the features was fixed by limiting the number of principal components according to a set of information rates within the range \([8, 99]\), resulting in the dimensionality range \([3, 45]\) based in the PCA of the source speaker. Finally, the predicted envelope parameters were obtained by bringing back the conversion results onto the original space (LSF parametrization). For all cases, the original mel-AR model order was 70.

6. Experimental results

6.1. Objective evaluation

We carried out a second evaluation of the conversion performance using the GMM-PCA based strategy just described. The results are displayed in Fig. 5. The reduced error found at a lower dimensionality than the original order suggests to us a better adaptation of the envelopes clustering when the reduction of information is applied, showing a slight improvement compared to the non-PCA case (Fig. 3). Following the plots, by considering a number of components around one third of the original dimensionality (24 instead of 70) the error already reaches a stable behavior. The corresponding reduced features allow us to keep a sufficient precision on the main features of the original envelope (formants). This is shown in Fig. 6. By using only very few components (5) the original formants cannot be reproduced, however, this is possible by using 24 components. Also, we remark that converted speech based on these representations showed an increased conversion effect over the case where the envelope order is just set to a similar value.

These results do not let us claim important benefits regarding a reduction of the overfitting. However, the advantages were clearly identified when listening to the converted speech. An increased naturalness and phonetic content preservation (no reduction) were perceived when using the PCA-based conversion for similar GMM settings (size, covariance matrix). The use of a larger number of components was found favorable within a reasonable range (below 128 components). The origin of some degradations beyond this value might be explained by the increasing overfitting.

Fig. 7 and Fig. 8 illustrate an extract of an original source-speaker signal (top), the conventional GMM conversion (middle), and the GMM-PCA one (bottom). For the GMM conversion, an irregular evolution of the energy is observed between the narrow segments limited by the dotted lines. As shown by the spectrogram plot, this is due to the unstable evolution of the converted envelope observed at the first harmonic partials. This behavior, commonly found in the conventional GMM conversion, was leading to perceived degradations of the quality. By observing several cases and considering the competitive behavior of the mixture a better connection at the low-frequency region across the selected GMM components on continuous speech for the GMM-PCA case was verified. We claim that the information reduction resulting from the application of principal components analysis positively affects the envelope clustering at this low-frequency region in terms of continuous speech synthesis.
French language. However, significant correlations on the resulting scores according to the familiarity of the listeners on the language were not found.

As converted material, ten utterances issued of both GMM and GMM-PCA methods were evaluated. As original data five sentences of the source and target speakers were considered. Following the results of the objective evaluation, the melAR envelope order was set to 70 and the number of principal components for the GMM-PCA case was fixed to 24. A slight transition factor (1.1) was applied to all the converted utterances in order to match the average target FO and to reduce accordingly the perceived effect of the pitch level when comparing the timbre conversion.

6.2.1. Conversion effect

For the conversion effect test, the listeners were asked to qualify if the timbre of a presented utterance was perceived as clearly corresponding or close to one of two references (source and target, randomly selected). Original utterances were also evaluated in order to verify the level of discrimination of the listeners between the source and target speakers. The utterances selected as references did not correspond to the same phrases of the evaluated ones in order to reduce the effect of the prosody and to concentrate the attention of the listener on the perceived timbre.

The results are shown in Fig. 9. Although a full conversion effect was not perceived overall, both methods were perceived closer to the target, with the GMM-PCA one exhibiting slightly better performance. These results follow those of the objective evaluation. However, as already presented, the projection of the features onto the PCA space appears to provide some benefits in terms of the mapping of some perceptually important information. This should be exhaustively verified among several conversion cases (speakers).

6.2.2. Quality comparison

The second test consisted of a direct comparison of the quality between GMM and GMM-PCA conversions. The listeners were asked to qualify which method resulted in less perceived artifacts and increased naturalness, or if there were not perceived differences between them. The results, illustrated in Fig. 10 show that in more than half of the trials, some differences were identified, resulting in the GMM-PCA approach being the best choice for around 80% of these cases. We attribute the high score of the "no differences" option to the low-
quality listening conditions of some participants. Following our informal observations, an evaluation performed in reliable listening conditions might increase considerably the capability to discriminate some differences found on the signals.

6.2.3. Overall quality - Mean Opinion Score

Finally, the third test consisted of an overall quality evaluation based on the well-known Mean Opinion Score (MOS). We wanted to measure the perceived quality of the converted speech following this scale and to compare it with the one of the original signals. As shown in Fig. 11, the GMM-PCA conversion resulted in a better MOS than the standard GMM one with, as already described, a reduced computational cost. Note however the remaining gap between the resulting score of the converted speech and the one of the original utterances, demonstrating that we cannot state the achievement of full natural speech conversion. Nevertheless, the results represent a step forward regarding the application to full-quality audio with a significantly reduced cost.

7. Conclusions

We presented in this work a study on the application of Voice Conversion to full-quality audio. A conversion framework based on GMM and precise spectral envelope estimates was applied to full-quality speech signals. The performance of a novel envelope model was studied in terms of the spectral conversion performance, confirming the benefits of using accurate spectral information for timbre conversion. The use of PCA to represent the features space was proposed in order to reduce the computational cost of the statistical model and to evaluate the conversion performance when the timbre mapping is derived in the reduced space. The proposed methodology exhibited benefits regarding the resulting converted speech quality. We plan in future work to validate the findings reported in this article among an increased number of conversion cases.

8. Acknowledgments

The authors would like to thank Jordi Bonada and Oscar Mayor from the Music Technology Group and Alex Fink from Arizona State University for their assistance provided to accomplish this work.

9. References