Applying Voice Conversion To Concatenative Singing-Voice Synthesis

Fernando Villavicencio¹, Jordi Bonada²

¹Corporate Research and Development Center, YAMAHA Corporation, Japan
²Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain

villavicencio@beat.yamaha.co.jp, jordi.bonada@upf.edu

Abstract

This work address the application of Voice Conversion to singing-voice. The GMM-based approach was applied to VOCALOID, a concatenative singing synthesizer, to perform singer timbre conversion. The conversion framework was applied to full-quality singing databases, achieving a satisfactory conversion effect on the synthesized utterances. We report in this paper the results of our experimentation focused to study the spectral conversion performance when applied to specific pitch-range data.

Index Terms: speech synthesis, speech analysis, singing synthesis, linear prediction, envelope extraction

1. Introduction

VOCALOID is a commercial concatenative singing synthesizer widely popular in Japan. Users of this technology are found among audio and music professionals as well as general public thanks to its easy conception and friendly interface. Lyrics and musical score are the only information required to obtain a sung utterance based on a given singer database. Unique in its nature, VOCALOID is widely considered as representative of the state of the art of singing synthesizers. Mental rendering is based on the modification of the speech of a source speaker in order to render it perceptually similar to that of a specific target one. The conversion process is based typically on a statistical mapping of the timbre space of the speakers (represented by spectral envelope estimates) and a matching of the average pitch of the target speaker.

The concept of imitating singer characteristics by a model-based procedure was already introduced in [3] in the perspective of future applications of singing-voice synthesis. Voice Conversion technology may represent an effective way to provide a multi-singer capability to the VOCALOID system by reproducing several DBs from a single one by simply applying a conversion model to a single singer database. Moreover, compared to typical Voice Conversion frameworks, the use of stable singing signals might represent some advantages in terms of the accuracy of the timbre patterns that are captured on the statistical model on which the source-target mapping is based.

Based on our previous works on high-quality Voice Conversion ([4],[5]), we present in this paper our application to the singing voice in order to achieve singer-timbre conversion on VOCALOID’s databases. An experimental study is presented to exhaustively evaluate the training and performance of the conversion framework when applied to singing voice. Following our previous works we were especially interested in studying the effect of using uniform pitch-range data instead of continuous speech on the envelope modeling performance.

The paper is structured as follows. We start at Section 2, giving a functional description of the VOCALOID system. Section 3 address the application of a GMM-based framework for the conversion of the timbre on high-quality singing-voice databases, followed at Section 4 by the description of an experimental study carried out to evaluate our implementation. The work finishes at Section 5 with our final conclusions.

2. Singing system overview

2.1. VOCALOID singing synthesizer

The system generates singing synthesis based on the concatenation of singing samples (diphones) selected from a singer database given the corresponding lyrics and musical score. Accordingly, as shown in Fig. 1, the main elements of the system are: the synthesis engine, which includes the functionality for the selection, modification and concatenation of the sequence of diphones; the score editor, consisting of an interface in which the user inputs the lyrics and desired musical score; and the singer database which contains the diphones samples and related analysis at three pitch ranges (low, medium, high).

The processing stream of the system can be briefly described as follows: firstly, given the lyrics and musical score provided by the user, the system transforms the text stream in a sequence of diphones and selects the corresponding samples from the databases based on the phonetic case and target pitch range. Then, the selected samples are processed and concatenated. Time stretching and F0 transposition are applied to match the desired musical score. More details of the system processing can be found in [1].

2.2. Singers databases

The sample units consist mostly of diphones as well as sustained vowels. Accordingly, the singer databases should contain all possible combinations of phonemes of a language. The diphone samples are extracted from full-quality recordings consisting of pre-defined phonetic sequences sung following similar tempo and melodic patterns. The pitch remains almost constant within each phoneme, being restricted to a small variation related to a predefined musical height along the melodic pattern. The process is repeated at three representative pitch ranges tagged as low, medium and high according to the nature of the singer (gender, age).
3. High-quality singer timbre conversion

To obtain a conversion of the singer identity we consider a strategy widely used in Voice Conversion approaches to perform speaker timbre conversion. The concept consists of a statistical modeling of the timbre space of the speakers by GMM in order to define a time-continuous mapping of features of the source speaker to those of the target [2]. The timbre information is represented by short-term estimations of the spectral envelope of voiced speech. In particular, we use the approach of [6] based on fitting the probabilistic model in joint source-target data and the parameterization of the envelope estimates on Line Spectral Frequencies (autoregressive modeling).

Voice Conversion has been mainly applied to low and medium quality speech ($s_r = [8 \, \text{kHz}]$). However, previous works of the authors address the successful application to high-quality speech ([4],[5]). In these works the use of precise envelope modeling is reported as an important factor of the converted speech quality. Accordingly, we aimed to extend and to evaluate our findings in spoken speech to the case of singing voice.

3.1. Singing vs spoken voice conversion

In the GMM approach, the components of the mixture are expected to model representative phonetic patterns of the speaker. Briefly, this is based on the attribution of a particular spectral shape to each phoneme, considered as representative of the spectrum of the signal within a stationary criterion. This assumption is justified as much as similar phonemes are uttered in an uniform way (voice quality, pitch range) along the data of the speaker. However, this can hardly be assumed for the case of real singing performances since the phonation conditions are continuously varying depending on the vocal register, pitch and tempo. In particular, note the existing dependencies between the configuration of the vocal tract and the pitch range for the case of singing voice [7]. Accordingly, data corresponding to similar phonetic content may observe significant differences on their spectral characteristics, resulting in ambiguous envelope patterns on the statistical model.

However, by using singing data recorded following uniform performance conditions (register, pitch, tempo) and selected phonetic content, the task appears feasible. Moreover, data following such controlled characteristics might observe, in principle, more stable envelope information than those coming from naturally spoken utterances. This is expected to have a favorable impact on the objective spectral conversion performance regarding an increasing precision of the source-target envelope mapping. Also, note that, following our interest, the conversion will be only applied to data of the same nature since we are exclusively focused to convert similar recordings. Therefore, by restricting the application of the conversion to such stable characteristics we expect to avoid some degradations observed when modifying speech segments where the spectral characteristics of the signal are strongly influenced by the speakers’ pronunciation.

3.2. Experimental framework

We were interested in clarifying several issues in our experimental study. Firstly, we wanted to study the effect of the pitch regarding the envelope precision and the resulting conversion performance. Note that by estimating the spectral envelope we aim to obtain an approximation of the underlying transfer function of the signal in a source-filter basis. However, for harmonic signals, this information is limited in the spectrum by the number of F0 partials. Clearly, the pitch range may have an impact on the resulting estimation and furthermore conversion of the envelope information. Note that, to perform singing synthesis on VOCALOID, the final pitch is defined by the target musical score and applied by the synthesis engine of the system. Accordingly, in our general use-case, the conversion process will be restricted to the modification of the timbre of the databases recordings of three different pitch ranges regardless of the nature of the singers (gender, age).

The experimentation done in this work was based on a female to male singer conversion. The average musical heights were C3 (131Hz), C3 (196Hz) and C4 (392Hz) for the female singer databases and G2 (98Hz), C3 (131Hz) and E3 (165Hz) for the male ones. The conversion framework was firstly applied between single-pitch databases in order to evaluate the conversion towards similar, higher and lower pitch ranges. Then, we considered the case of pitch-independent conversion by building a mixed database containing data of the three different pitches without keeping a similar pitch correspondence between the paired source-target utterances. By doing this we aimed at comparing the conversion performance when the source and target data keep the same phonetic content but not similar pitch values, as commonly done in Voice Conversion.

The size of the singer databases provides a significant amount of data. Nevertheless, due to the computational cost we restricted the maximal training size to 50,000 feature vectors. This amount appears to be higher than the common values reported in the bibliography on Voice Conversion, allowing us to closely verify the learning conditions of the GMM for the different data cases, as it was just described.

As an envelope model we used the melAR method presented in [5], consisting of an autoregressive model computed from a mel-scaled spline interpolation of the harmonics. The estimation of the harmonic information was done by means of the Wide-Band Harmonic Sinusoidal technique (WB) [8]. As usual, the resulting envelope estimates were represented by means of Line Spectral Frequencies (LSF). The conversion of the signals was achieved by applying at each frame the filter corresponding
melAR estimates in order to perceptually measure the impact of the spline interpolation of the harmonic analysis used to fit the scaled frequency axis (mel-dB). As target information we used envelope and original target-spectra evaluated over the mel-performance measure, the resulting MSE from the converted speech. The following results were obtained by using as a conversion and naturalness of the converted speech. (glottal excitation) should lead to an increased conversion effect on the behavior of both configurations on the GMM for increasing data. The envelope order was fixed to 70 since, as will be described later, it was the maximal allowed value showing increasing conversion performance. Following the results, significant differences on the learning conditions compared to the case of speech were not found. A generalization of the learning was achieved by using a number of vectors in the range [20000 – 40000] for the single-pitch data whereas some benefits were still observed for the mixed case at higher values. Note, however, the improved statistics of the error (reduced mean and variance) and learning stability of the single-pitch case. The results presented in the next sections were computed using [40000] training vectors.

4. Evaluation

The following results were obtained by using as a conversion performance measure, the resulting MSE from the converted envelope and original target-spectra evaluated over the mel-scaled frequency axis (mel-dB). As target information we used the spline interpolation of the harmonic analysis used to fit the melAR estimates in order to perceptually measure the impact of the error, not on envelope estimates, but on real target spectra.

4.1. Training size

We started by studying the effect of the training size on the learning performance. We evaluated the performance when using single and mixed pitch content in order to compare the learning capacity of the models and to verify the training size values commonly reported in the Voice Conversion bibliography.

The results are shown in Fig. 2 and Fig. 3. We tested both full and diagonal covariance matrices in order to evaluate the behavior of both configurations on the GMM for increasing data. The envelope order was fixed to 70 since, as will be described later, it was the maximal allowed value showing increasing conversion performance. Following the results, significant differences on the learning conditions compared to the case of speech were not found. A generalization of the learning was achieved by using a number of vectors in the range [20000 – 40000] for the single-pitch data whereas some benefits were still observed for the mixed case at higher values. Note, however, the improved statistics of the error (reduced mean and variance) and learning stability of the single-pitch case. The results presented in the next sections were computed using [40000] training vectors.

4.2. Pitch effect on the envelope modeling

Once a convenient training size was fixed, we proceeded to evaluate the effect of the pitch on the envelope modeling. Firstly, In Fig. 4 we show the resulting conversion error by using several GMM sizes between databases with similar pitch data of increasing envelope order. As expected, the resulting error was lower than the speech case [5], showing the benefits of using stable vocal phonations instead of continuous speech. There were not observed significant differences regarding the number of components on the GMM, suggesting a similar behavior of the clustering of the timbre space. Considering the resulting performance, the overfitting effect, and the required computational cost, we restricted our experiments to the consideration of a full and a diagonal GMM with 8 and 64 components, respectively.

We show in Fig. 5 the results of converting data considering several pitch cases. The 4 cases shown correspond to the conversion towards similar, lower, higher, and between mixed musical height. The latter case should be seen as the pairing of singing performances following similar lyrics and tempo but observing different melodic evolution. As expected, the performance corresponding to single pitch-range data observed lower conversion error than the mixed case. However, note the reduction on the performance for the case involving the conversion towards a lower pitch (C3-to-G2). The augmentation of the error should be partially attributed to the increased resolution of the target information: as the \( F_0 \) decreases, an increasing number of support points are available to fit the interpolated spectrum.

As a last figure of merit we show in Fig. 6 the resulting error measured on non-overlapping frequency bands perceptually scaled. The stable envelope information of the single-pitch data provides benefits on the conversion performance along the entire frequency axis. The results obtained for singing voice are comparable to the speech case, confirming the capacity of the conversion framework to specially map the envelope information within the vocal band ([0 – 4]KHz).
4.3. Subjective evaluation

An ABX and MOS tests were carried out to subjectively evaluate the timbre similarity and synthesis quality respectively. The results are shown in Fig. 7 and Table 1. As result, it was observed an increased conversion effect and naturalness compared to the case of speech [5].

<table>
<thead>
<tr>
<th>source</th>
<th>target</th>
<th>conversion</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOS</td>
<td>4.3</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 1: Synthesis quality: MOS evaluation.

5. Conclusions

We presented in this work a successful application of a Voice Conversion to perform timbre conversion on the singing-voice synthesizer, VOCALOID. A description of the implementation framework and the results of an experimental study were presented, confirming the benefits of performing spectral conversion by using accurate spectral envelope estimates of restricted pitch-range data to achieve singer timbre conversion.

6. References


