Hierarchical modeling of F0 contours for voice conversion

Gerard Sanchez\(^1\), Hanna Silen\(^1\), Jani Nurminen\(^2\), Moncef Gabbouj\(^3\)

\(^1\)Department of Signal Processing, Tampere University of Technology, Finland
\(^2\)Signalit Oy, Lempäälä, Finland
gerard.sanchez.gasulla@gmail.com, hanna.silen@tut.fi,
jan.nurminen@signalit.fi, moncef.gabbouj@tut.fi

Abstract

Voice conversion systems deal with the conversion of a speech signal to sound as if it was uttered by another speaker. The conversion of the spectral features has attracted a lot of research attention but the conversion of pitch, modeling the speaker-dependent prosody, is often achieved by just controlling the F0 level and range. However, the detailed prosody, including different linguistic units at several distinct temporal scales, can carry a significant amount of speaker identity related information. This paper introduces a new method for the conversion of the prosody, using wavelets to decompose the pitch contour into ten temporal scales ranging from microprosody to the utterance level, which allows modeling the different timings of the prosody phenomena. The prosody conversion is carried out in the wavelet domain, using regression techniques originally developed for the spectral conversion of speech. The performance of the proposed prosody conversion method is evaluated within a real voice conversion system. The results for cross-gender conversion indicate a significant improvement in naturalness when compared to the traditional approach of shifting and scaling the F0 to match the target speaker’s mean and variance.

Index Terms: voice conversion, prosody conversion, wavelet transform

1. Introduction

Voice conversion algorithms aim to modify the utterance of a first speaker to sound as if it was uttered by a second speaker. The speaker identity is composed by different factors, including short-term spectral characteristics, prosody and linguistic style, thus in voice conversion it is important to take into account all of them, in order to transform the global identity of the speaker. A significant amount of work has focused on the conversion of spectral parameters, which reflect the timbre of the voice (how the voice itself sounds). However, the modeling and conversion of the prosody is still one of the most challenging areas within the voice conversion framework, and the employed methods are still slightly simplistic.

The prosody of speech, the speaking style of the speaker, is an idiosyncratic feature of the speaker. Nevertheless, depending on what the speaker wants to reflect on the utterance, e.g. irony, sarcasm, question utterances or commands; the emotional state of the speaker, or the target audience of the speech, the prosody of the exactly same utterance may change significantly. Due to these reasons, the goal in prosodic conversion can be declared to be the generation of credible prosody, i.e. prosody that the target speaker could use in a certain situation.

Prosody is created by several factors, such as the phone duration, loudness and pause location, however, the main manifestation and the most expressive one is the pitch. Different speakers have different pitch ranges, which can be represented by calculating the mean pitch and pitch variance for each speaker. Nevertheless, the shape of the pitch contour, not just the mean and variance, contain speaker specific information that needs to be extracted and transformed in a voice conversion framework, since changing pitch throughout an utterance is usually the most powerful way of expressing emotion or emphasis based on the meaning of the message.

The most popular method for the prosody conversion is mean and variance scaling (MS) method. In MS method, only the mean and variance of the source speaker fundamental frequency (F0) are adjusted according to the target speaker F0 mean and variance. Therefore, the MS method retains the shape of the source F0 contour and it is unable to model the details of the target speaker prosody.

The assumption of a hierarchic model of prosody by phonologists and phoneticians is largely accepted [1, 2, 3, 4]. Recently, some approaches working with a hierarchic prosody model [5, 6], instead of a framewise of F0, have put in relevance that there is important information at every linguistic level of the utterance, from microprosody information on phonemes, to the prosody of the whole sentence and utterance.

The method introduced in this paper uses a well-known technique in signal processing, the wavelet transformation, to extract information related to different linguistic levels from the initial pitch. Similar F0 decomposition was recently proposed for the framework of statistical speech synthesis in [7]. The wavelet analysis allows discerning the different prosodic phenomena present in the speech, and modeling them separately according to each prosodic phenomena timing scale. The conversion is performed in the wavelet domain, using dynamic kernel partial least square (DKPLS), a regression technique originally developed for the conversion of the spectral features of speech [8].

The paper is organized as follows. In Section 2, the wavelet transform of F0 is introduced, highlighting the observable prosodic phenomena. Section 3 describes the proposed conversion method, achieved in the wavelet domain. The comparison between the proposed conversion method and the MS method is presented in Section 4, jointly with the results of a listening test. Finally, Section 5 concludes the work.

2. Wavelet transform of pitch

2.1. Preprocessing

Using the multiresolution properties of the wavelet analysis, the F0 contours are transformed to the wavelet domain in order to
achieve a better capability of understanding the prosodic phenomena present on it. However, the wavelet analysis is sensitive to the gaps of the signal, due to the unvoiced frames, besides mean and variance, therefore a couple of preprocessing steps, proposed originally by [7], are required to the precise conversion of the signal to the wavelet space.

The first step applied to the F0 contour is a transformation to the logarithmic scale, since the relevant information in the pitch signal is closely related to the logarithmic perceptual scale. To fill the gaps, produced by the unvoiced frames, a simple linear interpolation is applied to a smooth version of the F0 contour, created using a 3-point mean filter. The interpolated gaps are added to the original logarithmic F0, and a 3-point median filter is applied to the final interpolated signal to reduce continuities.

In order to reduce the effect of the edges, constant F0 is added prior and after the utterance. The pre-utterance F0 is set to the mean of the first half F0, while the post-utterance is set to the minimum of the second half F0. Finally, the interpolated F0 contour is normalized to zero mean and unit variance, required by the wavelet analysis, leading to the final F0 preprocessed contour depicted in Fig. 1.

2.2. Pitch contour decomposition

The continuous wavelet transform is the basis of the wavelet analysis, defined by

\[ T_{\text{wav}}(s, \tau) = |s|^{-1/2} \int f(t) \psi \left( \frac{t - \tau}{s} \right) dt \]  \hspace{1cm} (1)

where \( \psi \) is the mother wavelet, and \( s, \tau \) are the scaling and translation parameters. In this paper, we use the Mexican hat (second derivative of the Gaussian) as mother wavelet, and the decomposition of F0 is discretized to ten scales, one octave apart

\[ T_{\text{wav}}(i, \tau) = T_{\text{wav}} \left( 2^{i+1} s_0, \tau \right) \]  \hspace{1cm} (2)

with \( i = 1, \ldots, 10 \) and \( s_0 = 5 \text{ ms} \). These timing scales were originally proposed by [7], due to its proven relation with the prosodic formants of the hierarchic prosody model [10]. However, these timings were set for the Finnish language and in this work the approach is applied for English data, thus some modifications are required. Since Finnish is a predominantly polysyllabic language, word level and syllable level are quite different, thus the corresponding frequencies are clearly differentiated. On the other hand, English is closer to a monosyllabic language, hence syllable frequencies are closer to word frequencies.

Consequently, it is proposed increasing the scales by one octave, thus \( i = 2, \ldots, 11 \). With this new timing, the wavelet levels have been displaced fitting the prosodic phenomena observed at each level, such as the stress of the syllables and words, or the intonation pattern, to its corresponding frequency, as shown in Fig. 2.

3. Conversion of pitch

3.1. Wavelet domain conversion

The transformation of F0 contours to the wavelet domain allows separating the different prosodic phenomena to their corresponding temporal scale, for instance, microprosody events are represented in the first levels, the stress of the syllables is represented at levels 3 and 4, or the global intonation pattern is depicted at the sentence levels. However, the stress of the syllables, the stress of the words, and the intonation patterns, which are represented in the syllable, word and sentence levels (3, 4, 5, 6, 7 and 8), are important phenomena among speakers. Thus, these levels are energy-normalized and selected for the conversion.

The conversion of the wavelet levels is achieved using dynamic kernel partial least squares regression, a statistical mapping technique, originally developed for the conversion of the spectral parameters [8], that allows non-linear conversion and improves temporal continuity, since it allows handling the dynamics of speech. All the levels 3-8 are converted simultaneously, similarly to the spectral feature conversion of [8].

Once the conversion is achieved, transforming the syllable, word and sentence levels and copying the phoneme and utterance levels from the source, a final filtering stage is applied to
4. Experimental evaluation

4.1. Experimental setup

The performance of the proposed pitch conversion scheme was evaluated in the framework of DKPLS-based voice conversion [8]. Speech was parameterized by STRAIGHT [9] into spectral and excitation parameters, that were further encoded as mel-cepstral coefficients, F0, and mean band aperiodicity of five frequency bands.

For the hierarchical modeling, F0 contour was decomposed into ten levels using wavelet transformation. The scales were shifted by one octave from [7] to better match with the structure and timing of the English data. To model the suprasegmental features of prosody, wavelet levels 3-8 were converted using a model learned separately for each speaker pair by DKPLS regression. The two lowest and the two highest levels were copied from the source speaker as such. The traditional F0 scaling with the MS served as a reference approach.

The speech data from CMU Arctic dataset [11] was used for the evaluations. Two male voices (BDL, RMS) and two female voices (SLT, CLB) were considered in the conversion. For each possible speaker pair, 40 parallel sentences aligned by dynamic time warping were used for training. For spectral parameters, standard DKPLS regression approach was used [8]. Energy and band aperiodicity features were copied from the source speaker.

There are conversion samples available online at http://www.cs.tut.fi/~sgn/arg/sanchez/is2014.

4.2. Objective measurements

The performance of the proposed F0 conversion was evaluated in terms of root mean squared error (RMSE) and correlation between converted and target F0 contours. It was computed over 25 test sentences for every possible conversion scenarios. The results are summarized in Table 1.

In general, the RMSE for the proposed system is lower than for the MS method, except for the female-to-male conversion. Furthermore, for the female-to-female and male-to-female conversion, the proposed approach results in higher correlation compared to MS, hence suggesting more accurate modeling for the shape of the target F0 contour. For male-to-male conversion, the correlation is increased as well, but it should be noted, that the the shape of the source and target F0 contours typically was rather similar even without any conversion. For the female-to-male conversion, the correlation of the signals has decreased.

<table>
<thead>
<tr>
<th>Conversion</th>
<th>RMSE</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-to-F</td>
<td>19.92 Hz</td>
<td>61.93 %</td>
</tr>
<tr>
<td>F-to-M</td>
<td>12.07 Hz</td>
<td>59.12 %</td>
</tr>
<tr>
<td>M-to-F</td>
<td>9.22 Hz</td>
<td>74.86 %</td>
</tr>
<tr>
<td>M-to-M</td>
<td>19.09 Hz</td>
<td>59.02 %</td>
</tr>
</tbody>
</table>

Table 1: RMSE and correlation comparison between prosody conversion with wavelets and mean and variance scaling (MS) method, for the four possible scenarios of conversion.

4.3. Perceptual evaluation

The listening test was conducted under the assumption of creating a credible sentence, which might be uttered by the target speaker in some situations. The test was designed in such a way that also non-native listeners with good English skills...
can judge the prosody and relevant issues in the naturalness of the test samples. The test consisted of 100 randomly selected pairs of sentences generated by the proposed approach with wavelets and the traditional MS approach. For each pair, listeners were asked to choose the sample representing higher naturalness. The subjects could also choose equal. For each sentence pair, the target sample was made available as well. Altogether 16 listeners participated in the test.

The results of the listening experiment are shown in Fig. 5. For the intra-gender conversion, no clear preference was found. However, for the inter-gender conversion, the proposed approach clearly outperforms the traditional scaling approach.

![Figure 5: Preference results, with 95% confidence interval for the mean and variance scaling (MS) and the proposed wavelet method (CW).](image)

In the evaluation, the listeners were asked to only rate the naturalness, not how well the prosody matches the expected prosody of the target speaker. The results suggest, that in the case of intra-gender conversion, much less modification is needed compared to the case of inter-gender conversion and, depending on the speaker pair, the simple reference method may already perform quite well in terms of naturalness. And even more importantly, since the listeners did not know the speakers, the detailed prosody does not play that important role, as long as the F0 level and scale are adequate and output quality high enough.

However, in the case of inter-gender conversion, the situation is different. The reference method performs adequately rather rarely because the differences in prosodies are more significant. The simple shifting and scaling might reduce the naturalness rather much in inter-gender conversion. Thus, by more detailed F0 modeling, more significant improvements can be achieved even in the case when the listeners do not know the actual speakers.

5. Conclusions

This paper introduces a new prosody conversion method within a voice conversion framework. The proposed approach uses the wavelet transform to model the different prosodic phenomena, working with an extended set of data for every temporal frame, in contrast with most of the conversion systems present in the literature. The wavelet separation in different temporal streams has shown a considerable potential as an analysis system of the prosodic phenomena of the speech, since the diverse prosodic events are temporally classified and thus, can be easily studied, modeled, and finally converted in the conversion stage.

The selection of the timings of the wavelet is a critical point since the language of the speakers influences the prosody timing, and thus, the wavelet levels of prosody. Nevertheless, comparing the number of peaks in the word levels with the word boundaries, e.g. with a peak prominence study similarly to [10], the timings are set properly to their corresponding prosody level, allowing the correct characterization of the prosodic phenomena.

The method was tested with objective measures, RMSE and correlation, suggesting that the proposed conversion system is better capable to model the target speaker F0 contour than the MS method, in three of the four tested scenarios. Moreover, the proposed method was also tested in a real voice conversion system. Listening tests showed a clear preference for the proposed method in the inter-gender conversion, where pitch differences are noteworthy, improving the naturalness of the speech samples when compared against the baseline MS conversion. Thus, it is suggested that the proposed method allows creating more natural profiles of pitch in the cases where the differences of speaking style are remarkable among the speakers.

6. References