Japanese Pitch Conversion for Voice Morphing Based on Differential Modeling

Ryuki Tachibana¹, Zhiwei Shuang², Masafumi Nishimura¹

¹ Tokyo Research Lab., IBM Research, Kanagawa-ken, Japan
² China Research Lab., IBM Research, Beijing, China
{ryuki,nishimura}@jp.ibm.com, shuangzw@cn.ibm.com

Abstract
In this paper, we convert the pitch contours predicted by a TTS system that models a source speaker to resemble the pitch contours of a target speaker. When the speaking styles of the speakers are very different, complex conversions such as adding or deleting pitch peaks may be required. Our method does the conversions by modeling the direct pitch features and differential pitch features at the same time based on linguistic features. The differential pitch features are calculated from matched pairs of source and target pitch values. We show experimental results in which the target speaker’s characteristics are successfully modeled based on a very limited training corpus. The proposed pitch conversion method stretches the possibilities of TTS customization for various speaking styles.

Index Terms: pitch conversion, voice conversion, voice morphing, speech synthesis, differential modeling.

1. Introduction
Voice conversion changes the characteristics of the voice of an SPS (SPeaker Source) to that of an SPT (SPeaker Target) for various applications. One important application is to build customized text-to-speech (TTS) systems for different companies, so a TTS system with each company’s favorite voice can be created quickly and inexpensively by modifying the speech corpus of some original speaker.

Spectra and prosody are the two major characteristics of voice. For spectral conversion, recent work such as [1, 2] has achieved significant improvements in the naturalness and similarity of the voices converted using only a limited amount of training data. However, not much research has been done into prosody conversion. Most spectral conversion research uses simple linear transformations for the prosody. It is true that the detailed prosody difference is sometimes difficult for listeners to distinguish [3], especially when the speakers are monotonously reading for TTS corpus recording. However, to generate TTS voices with a lively colloquial speaking style, reproduction of the detailed prosody characteristics is important.

Our objective is to reproduce the SPT’s speaking style of pitch contours based on limited training data. We assume 100 sentences as training data is a reasonable amount to require for the SPT’s speech corpus. A speech corpus with that size can easily be recorded in a thirty-minute recording session. We do not assume the existence of a parallel corpus, which can be a difficult condition to satisfy. We focus on the pitch changes around the syllable level, because the important pitch changes in Japanese are mainly at the syllable level.

Figure 1 illustrates examples of pitch contour pairs that the proposed pitch conversion method can handle. They are (a) asymmetrical slope changes, (b) adding or deleting peaks, and (c) adding or deleting phrase-final rises.

2. Proposed Method

2.1. Main Ideas
The problem is to predict the SPT’s pitch contour \(y_t = \{y_{i|1} | i = 1 \ldots N_f\}\) for an input sentence by modeling the speaking style, where \(N_f\) is the number of frames in the voice. Our main ideas are:

(1) Use both static and dynamic features of the pitch contour

A speaker’s speaking style must be characterized in its patterns both in static pitch values \(y_i\) and dynamic pitch changes \(\Delta y_i\) and \(\Delta^2 y_i\). We calculate the optimum pitch contour by taking into consideration the distributions of static and dynamic features in a way similar to how an HMM-based TTS [4] works. The target feature vector is defined by \(z_{yt} \equiv (y_t, \Delta y_t, \Delta^2 y_t)\).

(2) Differential modeling

Since we assume the amount of training data is small, we cannot directly model the SPT’s pitch features. Hence, we model the differences of the SPT’s pitch contour from the SPS’s pitch contours. For this purpose, we predict the pitch contours \(y_s = \{y_{j|1} | i = 1 \ldots N_f\}\) of the SPS by using the SPS’s pitch model of our TTS system. This procedure makes parallel corpora of the SPT human voices and the SPS synthetic voices. Next, we calculate the differences of the pitch contours \(\delta y = y_t - y_s\), and model the differences with a stochastic model (Fig. 2). In the runtime, we first predict the SPS’s pitch contour and then...
predict the SPT’s pitch contour based on the first prediction. By doing this, the SPS’s pitch model serves as a safety net in a sense that it is automatically used for the context where the training data is insufficient. The differential feature vector is defined by \( d_o \equiv (\delta_y, \Delta \delta_y, \Delta^2 \delta_y) \). 

(3) Combination of differential and direct modeling
Since modeling the pitch differences between an SPT and an SPS can sometimes be more complicated than directly modeling the SPT pitch, we include the target features \( z_{sp} \) in the model as well as the differential features \( d_o \). The runtime procedure calculates the optimum path with respect to the likelihood of both features. Since the variances of the features serve reliability indicators, more reliable features automatically contribute more to the optimization.

### 2.2. Training Procedure
The model is trained using the following procedure (Fig. 3).

(1) SPT Voice Collection
We collect a speech corpus of the SPT human voices that are to be modeled.

(2) Alignment
We use forced alignment in an automatic speech recognition (ASR) system to find the phonemic alignment points in the SPT voices.

(3) Synthesis
The TTS system first uses the SPS’s model to generate the synthetic voice for the transcribed sentences. For the prosodic structures of the synthetic voices, such as the positions of pauses and phrase boundaries, as close to those of the SPT as possible, we manually correct the text processing results before the synthesis.

(4) Pitch Detection
For both the SPS synthetic voices and the SPT human voices, a wavelet-based tool detects pitch values for constant-spaced frames. We interpolate pitch values for the unvoiced frames and calculate a smoothed pitch value for each frame by convoluting a Gaussian smoothing function. In addition, we normalize the mean and the standard deviation of the pitch values for each intonational phrase. This process extracts syllable-level pitch changes by removing the micro-prosody and discourse-level prosody. For each frame, we sample a point \((x, y)\) on the pitch contour. The pitch \( y \) is logarithmic and has a unit of logarithmic Hertz.

(5) Matching
Each point on the SPS pitch contour is matched to a point on the SPT contour. A Dynamic Time Warping (DTW) algorithm calculates the matching points based on the Euclidean distances of spectral feature vectors. We use the phonemic boundaries as anchor points for the DTW procedure.

(6) Output Feature Extraction
Let us assume the \( i \)-th SPS frame is mapped to the \( j \)-th SPT frame. The target pitch features for the SPT frame are \( z_{sp}[j] \equiv \left( y_i[j], \Delta y_i[j], \Delta^2 y_i[j] \right) \). The differential pitch features are \( d_o[i] \equiv \left( \delta_y[i], \Delta \delta_y[i], \Delta^2 \delta_y[i] \right) \), where \( \delta_y[i] \equiv y_i[j] - y_s[i] \). These features define an output feature vector for the decision tree (DT) and the Gaussian model (GM).

\[
(z_{sp}[j]^T, d_o[i]^T)^T = \left( (y_i[j], \Delta y_i[j], \Delta^2 y_i[j])^T \right).
\]

(7) Input feature extraction
The input feature vector for the DT training has 50 dimensions, which includes (a) 25 features to track the positions of the current and neighboring frames, segments, syllables, and phrases, (b) 12 linguistic features including part-of-speech, phrase type, syllable stress, and accentual type, (c) 9 features for identifying the current and neighboring phonemes, (d) 5 features for the quantized static and dynamic pitch values of the current SPS frame, and (e) 1 feature for the voice of the current SPS frame.

(8) Decision Tree (DT) and Gaussian Model (GM) training
We then train the DT to predict a 6-dimension output feature vector based on the 50-dimensional input feature vector. The training feature vectors are classified to the leaves of the DT. A multi-variate Gaussian model with a 6-dimension feature vector is constructed for each leaf.

### 2.3. Runtime Procedure
The runtime procedure starts by generating a synthetic voice for the input sentence. The voice is analyzed to determine the pitch values and input feature vectors for each frame as in the training steps (4) and (7). We also calculate the SPS pitch feature \( z_{sp}[i] \).

Next, by traversing the DT with the 50-dimension input feature vectors, we calculate a DT leaf for each frame, which also determines a Gaussian distribution tagged for the leaf.

Now we know the SPS pitch feature \( z_{sp}[i] \), which we can use as a reference, and the Gaussian distribution, for which we want to maximize the likelihood, of the 6-dimension feature vector for each frame.

The next problem is to determine the optimum sequence of \( y_s[i] \) based on this information. The observation vector whose likelihood we want to maximize is \( o = [o[i]|i = 1..N_f] \), where \( o[i] = (z_{sp}[i]^T, d_o[i]^T)^T \), which is a combination of direct features and differential features. Because the \( \Delta \) and \( \Delta^2 \) sequence can be calculated by matrix operations on the static feature sequence \( y_i \) or \( \delta_y[4] \), and the differential feature vector \( \delta_y \) has been defined as \( \delta_y \equiv y_i - y_s \), \( o \) can be written as:

\[
o = \begin{pmatrix} z_{sp} \\ d_o \end{pmatrix} = \begin{pmatrix} W_{y_i} \\ W_{\delta_y} \end{pmatrix} y_s + \begin{pmatrix} W_{y_i} \\ W_{\delta_y} \end{pmatrix} (y_i - y_s) = U y_i - V y_s \quad (4)
\]
The values of $\frac{\partial L}{\partial \mu}$ can be calculated as

$$L = \frac{1}{2} (\mu - \mu_0)^T \Sigma^{-1} (\mu - \mu_0) \quad (5)$$

$$= \frac{1}{2} \{U y_t - V y_s - \mu_0\}^T \Sigma^{-1}_s \{U y_t - V y_s - \mu_0\}$$

$$= \frac{1}{2} \{U y_t - \mu_{o'}\}^T \Sigma^{-1}_o \{U y_t - \mu_{o'}\}, \quad (6)$$

where $\mu_{o'} \equiv V y_s + \mu_o$. The optimum sequence $\hat{y}_i$ satisfying $\partial L / \partial y_i = 0$ can be calculated as

$$\hat{y}_i = (U^T \Sigma_i^{-1} U)^{-1} U^T \Sigma_i^{-1} y_s$$

$$= \Sigma^{-1} r. \quad (7)$$

The values of $R$ and $r$ are defined and calculated as $R \equiv U^T \Sigma_i^{-1} U$ and $r \equiv U^T \Sigma_i^{-1} y_s$. The covariance matrix $\Sigma_o$ can be written as a combination of covariance matrices as

$$\Sigma_o = \begin{pmatrix} \Sigma_{x_i y_i} & \Sigma_{x_i y_d} \\ \Sigma_{y_d x_i} & \Sigma_{y_d y_d} \end{pmatrix} \quad (9)$$

Since we regard the covariance matrices, $\Sigma_{x_i y_i}$, $\Sigma_{x_i y_d}$, and $\Sigma_{y_d y_d}$ as diagonal, we can easily calculate the inverse matrix $\Sigma_{x_i y_d}^{-1}$. For example, its $(i, i)$-th component is calculated as $c[i] / (a[i] c[i] - b[i] b[i])$ where $a[i]$, $b[i]$ and $c[i]$ are the $i$-th component of $\Sigma_{x_i y_i}$, $\Sigma_{x_i y_d}$, and $\Sigma_{y_d y_d}$, respectively.

By using the special structure of $R$, we can calculate Eq. (8) analytically by using Cholesky decomposition [4].

Now we finally have the optimum sequence $\hat{y}_i$ as the SPT pitch contour, we can generate the pitch-converted synthetic voice as the final output, by running the TTS system again using the obtained pitch values both as the target pitch for unit selection and the final pitch for signal processing.

### 3. Experiments

We conducted subjective listening tests to evaluate the individuality of the speaker and the naturalness in the pitch contours based on Japanese corpora and a unit selection TTS system [5].

For each test sentence, we first produced the SPS pitch contour by using the TTS system, and then converted the contour by using either the proposed method or the conventional Gaussian Normalization method. By using the converted pitch both as the target pitch for unit selection and the final pitch for signal processing, a synthetic voice was generated by the TTS system. Then we used the spectral conversion algorithm [2] to convert the spectra to match those of the SPT. The spectral conversion algorithm is based on frequency warping and the warping function was trained with a single pair of vowel utterances of the SPT and SPS. The voice conversion algorithm also linearly converted the duration of the synthetic voices to agree with the average speech rate of the SPT.

In the conventional Gaussian Normalization method, the pitch $y_s[i]$ for each frame was calculated as

$$y_s[i] = (y_s[i] - \mu_s) \frac{\sigma_r}{\sigma_s} + \mu_r, \quad (10)$$

where $\mu_r, \sigma_r$ ($\mu_s$ and $\sigma_s$) were the global mean and standard deviation of the SPS (SPT), respectively.

For the proposed method, we used 7 frames for calculating the $\Delta$ and $\Delta^2$ pitch features. The interval between adjacent frames was 5 msec. The number of the leaves in the trained DT was 655. Figure 4 shows an example of the SPS pitch, the pitch converted by the method, and the SPT pitch, two of which are shifted upward for better visualization. We can see the proposed method successfully reproduced a huge pitch rise near the end of the phrase.

Table 1 shows the characteristics of the speakers used for the tests. Since the objective of our research is to reproduce the pitch details, we chose female speakers with similarly high pitch means. While SPS was a recording by a professional speaker reading for TTS corpora, SPT was a podcast recording by a young woman, who is not a professional speaker, speaking in a colloquial style.

Therefore the pitch conversion from SPS to SPT was a challenging task in stretching the possibilities of TTS customization. The SPT training corpus contains 100-sentences, which were not included in the training corpus when building the TTS system. Since colloquial speech does not always have clear sentence boundaries, we split sentences also at semantic phrase boundaries. The total duration of the SPT training corpus was 338 seconds excluding silences. We took 24 test sentences from another podcast recording of the SPT. To reflect the SPT’s speaking style in the synthetic voices, we manually corrected the text processing results for the test sentences. The same manually corrected results were used for both of the methods.

Each of the 15 subjects listened to 10 randomly chosen sets of synthetic voices. One set consists of the SPT human voice and synthetic voices generated with the proposed method and the conventional method for the same sentence. The order of the synthetic voices in each set was random and we did not tell the subjects which method generated which voice. We asked the subjects to choose which of the synthetic voices sounded more similar to the human voice and to rate the naturalness of the prosody in the synthetic voices on a scale of 1 (very unnatural) to 5 (very natural). The subjects were exposed to 20 sentences of the SPT human voices to learn the SPT’s characteristics in advance of the tests.

Figure 5(a) shows the results of the preference tests for speaker individuality. The proposed method was preferred in 59.3 % of the tests. According to a binomial test, there was a significant difference between the preferences of the methods ($p = 0.0136$). The listeners’ perception of the speaker identity was dependent on the sentences.

However, the MOS for the naturalness of the proposed method was slightly below that of the conventional method (Fig. 5(b)). We can also see the scores for both of the methods were in a relatively low level. We consider a major reason...
for these is the mismatch between the colloquial speaking style in the test sentences and the converted pitch, and the reading style in the other aspects of the synthetic voices such as energy and duration.

4. Related Work

Many approaches to spectral conversion as part of voice conversion (VC) have been studied. Major approaches include GMM-based methods [6, 7], frequency warping (FW) [8], a combination of GMM and FW [9], a method incorporating FW and unit selection [2], and a method using maximum likelihood estimation similar to HMM-based TTS [1]. In contrast, relatively small consideration has been given to pitch conversion. A common approach for pitch conversion is Gaussian normalization, which changes the mean and the standard deviation of the pitch to match the SPT, basically assuming the pitch is distributed as a Gaussian distribution. However, in recent years increasing research work is focusing on this topic, as speech research is getting more targeted on spontaneous speaking styles, in which the detailed pitch variations can also carry a significant amount of the speaker’s identity. As improvements for Gaussian normalization, Gaussian normalization around a deviation line for each phrase [10] and a piecewise linear mapping method [3] have been proposed. Though these methods have advantages with small amounts of training data and don’t require parallel training corpora, a disadvantage is that they cannot perform complex conversions such as deleting or adding peaks.

A GMM-based approach maximizing the joint probability of the source and target feature vectors was deeply studied [6, 7]. One of the advantages is good sound quality because the GMM model can handle the coupling between pitch values and spectral coefficients (especially for the first formant). The disadvantage is that it requires a large amount of training data to model the combined feature vectors. In addition, some linguistic features with large effects on the pitch can not yet be handled by these methods.

Another method in [11, 12] also calculates and optimizes the joint feature vector of the pitch and spectral coefficients. The main difference of the method in [11, 12] from the GMM-based approach is that it uses a DT based on linguistic features to model the pitch-related feature vectors including the differential features. We believe the pitch features are better modeled by the linguistic features than by acoustical features when the training data amount is limited. We chose not to do a joint conversion, though the coupling of pitch and spectra is known to be important. Our primary reason was to reduce the amount of training data. By decoupling the spectral feature vectors, we can use more data (from our small amount of training data) for each linguistic context.

The optimization operation proposed in our method is a variation of the optimization operations proposed by Tokuda’s group [4, 11, 12]. The major difference is that our method uses a DT based on linguistic features to model the pitch-related feature vectors including the differential features. We believe the pitch features are better modeled by the linguistic features than by acoustical features when the training data amount is limited. We chose not to do a joint conversion, though the coupling of pitch and spectra is known to be important. Our primary reason was to reduce the amount of training data. By decoupling the spectral feature vectors, we can use more data (from our small amount of training data) for each linguistic context.

5. Conclusions

In this paper, we proposed a pitch conversion algorithm that uses the decision tree based on the linguistic features to model the joint feature vectors of the direct pitch features and the differential pitch features. The subjective listening tests in Japanese showed the method successfully reproduced the colloquial speaking style of the target speaker based on a limited training corpus whose substantial size was below 7 minutes.

Though we chose not to use joint conversion, it is true that the overall sound quality was degraded by this configuration. Our future works include use of global variances and asymmetric components of covariance matrices as well as Gaussian Mixture Models.

6. References