Using Zero-Frequency Resonator to Extract Multilingual Intonation Structure

Jinfu Ni, Yoshinori Shiga, and Hisashi Kawai

Advanced Speech Technology Laboratory, ASTREC, National Institute of Information and Communications Technology, Japan

\{jinfu.ni, yoshinori.shiga, hisashi.kawai\}@nict.go.jp

Abstract

Human uses expressive intonation to convey linguistic and paralinguistic meaning, especially making focal prominence to give emphasis that highlights the focus of speech. Automatic extraction of dynamic intonation feature from a speech corpus and representing it in a continuous form are desired in multilingual speech synthesis. This paper presents a method to extract dynamic prosodic structure from speech signal using zero-frequency resonator to detect glottal cycle epoch and filter both voice amplitude and fundamental frequency (F0) contours. We choose stable voice F0 segments free from micro-prosodic effect to recover relevant F0 trajectory of an utterance, taking into consideration of inter-correlation of micro-prosody with phonetic segments and syllable structure of the utterance, and further filter out long-term global pitch movements. The method is evaluated by objective tests upon multilingual speech corpora including Chinese, Japanese, Korean, and Myanmar. Our experiment results show that the estimated intonation contour can match F0 contour by conventional approach in very high accuracy and the estimated long-term pitch movements demonstrate regular characteristics of intonation across languages. The proposed method is language-independent and robust to noisy speech.

Index Terms: Fundamental frequency analysis, zero-frequency filtering, glottal cycle epoch, multilingual speech synthesis, and speech prosody

1. Introduction

Human uses expressive intonation to convey linguistic meaning and paralinguistic information [1]. Changes in pitch or fundamental frequency \((F_0)\) along a sentence enable listeners to perceive the sentence’s intonation, and changes in time and intensity aspects also serve as acoustic cues in the perception process. Intonation structure here focuses particularly on the acoustic aspect of \(F_0\) or pitch and terms \(F_0\) and pitch are exchangeably used in this paper. In the context of text-to-speech synthesis [2][3], synthesis of appropriate intonation from input text is important for accurately conveying all of the nuances of the message [4].

In asian languages, local changes in pitch distinguish word meaning. In standard Japanese, for a \(n\)-syllable word, there are \(n + 1\) possible accent types marked by type 0 (non-accent), 1, 2, ..., \(n\). For example, in disyllabic words there exist three-way minimal contrasts, such as \(kaki\) with type 0 (meaning: persimmon), \(ka’ki\) with type 1 (oyster), and \(kaki’\) with type 2 (fence). In Myanmar, there are three tones. When “ma” is pronounced with different tones, it has different meaning: “hard” with Tone 1 (written as \(/\text{ma}/\), “lift” with Tone 2 \((/\text{ma}^2/\) and \(/\text{ma}/\), and “doctor” with Tone 3 (in word \(/\text{ma}/\) and \(/\text{ma}/\) [5]. In standard Chinese, there are five tones, e.g., “ma1” (mother), “ma2” (numb), “ma3” (horse), “ma4” (to curse), and “ma0” (question particle) [6]. A consistent treatment of interactions of tone/accent with intonation in \(F_0\) is desired in multilingual speech synthesis [7].

The principle of superposition is attractive as it is intuitive to model different components or functions of pitch separately [8][9][10]. However, automatic pitch decomposition into its constituent part turns out not to be a trivial task [11][12][13][14][15][16]. The main difficulty may come from three aspects. First, there is no unique solution to decomposition of \(F_0\) contour in general [16], because several components can trade to produce the same \(F_0\) contour. Second, \(F_0\) contours are often not smooth, interrupted by non-sonorant sounds, and perturbed by segmental effects known as micro prosody [10]. Third, phonetic implementation of intonation and accents is rather complex; the boundary between them is not always clearly cut [17].

This paper presents a method for extracting multilingual intonation structure from speech signal in light of the superpositional principle [8][18]. There is no direct pitch decomposition. Instead, we extract both \(F_0\) trajectory and virtual pitch register (baseline of tonal space) [19] for an utterance, separately. The former gives accurate local pitch changes responsible for lexical tone or accent and the latter global pitch movement for proper intonation. A combination of both yields a means for pitch decomposition, if needed. We apply a zero-frequency filtering (ZFF) method [20] to extract characteristics of glottal activity from speech signal to deal with micro-prosodic effect, taking into account the inter-correlation of it with syllabic segments. The ZFF is further used as a super-smoother; this is based on its integration function. An advantage of the proposed method is language-independent.

The rest of this paper is structured as follows. Section 2 outlines zero-frequency resonator, ZFF, and the proposed method. Experimental results and discussions are presented in Sections 3 and 4, respectively. Section 5 concludes this paper.

2. Approach outline

2.1. Zero-frequency resonator

In previous work [6][19], a tone transformation technique was used to compute global pitch movement (virtual pitch register) from the observed \(F_0\) contour of an utterance, taking into account amplitude-frequency response mechanism:

\[
A(\lambda, \omega) = \frac{1}{\sqrt{1 - (1 - 2\zeta^2)\lambda^2 + 4\zeta^2(1 - 2\zeta^2)\lambda^2}},
\]

where \(\zeta \leq 1\), squared ratio of driving frequency to natural frequency of a vibrating system, \(\zeta^2 \leq 0.5\), system damping ratio. Given \(\cos(\omega/2) = 1 - 2\zeta^2\), Eq. (1) is rewritten as

\[
A(\lambda, \omega) = \frac{1}{\sqrt{1 + \lambda^2 - 2\lambda \cos(\omega/2)}},
\]

where \(\omega\) is angular frequency. Eq. (2) is equivalent to the frequency response of a resonator, an infinite impulse response.
(IIR) filter [21]. Let $\omega = 0$ (zero-frequency), or $\zeta = 0$ (non-damping), an ideal resonator results that is an IIR with a pair of poles located on the unit circle. The ideal zero-frequency resonator (ZFR) can be expressed as

$$y[k] = x[k] + 2y[k - 1] - y[k - 2].$$

(3)

2.2. Zero-frequency filtering (ZFF) method [20]

A ZFR-based filter was proposed to extract epoch from speech [20]. An advantage of ZFF is that the characteristics of time-varying vocal tract do not affect the discontinuities of impulses in the filter output. The ZFF method includes 3 steps.

1. Removing any slowly varying component of signal $s[k]$.

$$x[k] = s[k] - s[k - 1].$$

(4)

2. Passing $x[k]$ through a cascade of two ideal ZFRs, i.e.,

$$g[k] = x[k] + 4y[k - 1] - 6y[k - 2] + 4y[k - 3] - y[k - 4].$$

(5)

The resulting $g[k]$ grows approximately as a polynomial function of time. The trend in $g[k]$ is removed next.

3. Removing the local mean at each sample with a window.

$$\hat{z}[k] = g[k] - \frac{1}{2N + 1} \sum_{n = -N}^{N} g[k + n]$$

(6)

is called zero-frequency-filtered (ZFF) signal. $2N + 1$ indicates the size of the window used for trend removal.

2.3. Extraction of intonation structure

In the work, we employ the ZFF method to filter $F_0$ and amplitude contours besides extraction of epoch from speech signal. For the former, an iterative algorithm is developed below.

**Algorithm 1**: Iterative zero-frequency filtering of signal.

- **Input**: $s[n]$ (input signal), $K$ (number of iterations), $N$ (half size of window in Eq. (6)) required for ZFF.
- Linear interpolation for zero portion of $s[n] \rightarrow s_0[n]$.
- Pass $s_0[n]$ through ZFF $\rightarrow \hat{s}_0[n]$, set $i = 0$.
- While $i < K$, iteratively do the following steps.
  - Pass $s_0[n] - \hat{s}_i[n]$ through ZFF $\rightarrow \Delta \hat{s}_i[n]$.
  - Set $\hat{s}_{i+1}[n] := \hat{s}_i[n] + \Delta \hat{s}_i[n]$ and $i := i + 1$.
- **Output**: $\hat{s}_K[n]$.

Figure 1 outlines the proposed method for extracting intonation structure of an utterance: continuous $F_0$ trajectory free from micro-prosodic effect and virtual pitch register to feature proper intonation. The following describes this method.

1. **Compute ZFF signal from speech signal.** Pass input speech signal through the ZFF method, where $N$ is estimated by the mean of $F_0$ values extracted by tool get_f0 [23]; the output is ZFF signal.

2. **Detect glottal cycle epoch from the ZFF signal.** The epoch is at the instant when ZFF signal changes from negative to positive values as suggested in [20].

3. **Compute amplitudes of the ZFF signal.** At each glottal cycle, compute the maximum of absolute amplitudes of the ZFF signal and sample them using 5 ms window with 5 ms shift. The resulting amplitude sequence, $s_a[k]$, codes information related to source excitation and the status of vocal-cord vibration.

Pass amplitude $s_a[k]$ through **Algorithm 1** to obtain

- ZFF-amplitude (with $N = 100$ and $K = 10$).
- Fitted amplitude (with $N = 100$ and $K = 10$).
- Smoothed amplitude (with $N = 300$ and $K = 5$).

These parameters are jointly used to select stable voice frames to remove micro-prosodic effect on the process of recovering $F_0$ trajectory next.

4. **Detect voice frames.** Assign a frame as voiced if normalized ZFF-amplitude $\hat{s}_a[k] \geq 0.08$, and delete isolated voice frames, if any.

5. **Select stable voice frames for recovering $F_0$ trajectory.** Select stable frames for recovering $F_0$ trajectory taking into account dynamic features of both amplitude and $F_0$.

- Compute mean $\mu_a$ and variance $\sigma_a$ of $\Delta \hat{s}_a[k]$.
- Compute the intersection between the fitted and smoothed amplitudes and initially mark such frames that are located at the intersecting points or the peaks of the fitted amplitude as stable frames.
- Assume any frame, say, $i$, next to an existing stable frame, $j$, as stable one if $|\hat{s}_a[i] - \hat{s}_a[j]| \leq \mu_a + \sigma_a$.
- Remove critical stable frames, e.g., difference of its $F_0$ with adjacent frame’s $F_0 \geq 0.8$ semitones.

6. **Recover $F_0$ trajectory from the stable frames.**

- Compute $F_0$ for the selected stable frames from the detected epoches but take 0 for the others.
- Pass the resulting $F_0$ sequence through **Algorithm 1** with $N = 100$ and $K = 15$.

7. **Estimate virtual pitch register from the $F_0$ trajectory.**

- Pass the continuous $F_0$ trajectory through **Algorithm 1** with $N = 150$ and $K = 1$. 

Figure 1: Outline of extracting intonation structure from speech.
• Shift the output contour downward with a base $f_0[k] = 2.5$ semitones. The resulting contour is assumed as virtual pitch register for the register.

Note that the values of control parameters ($N$ and $K$) in Algorithm 1 could be changed to certain extent.

### 3. Experimental setup and results

Evaluation on automatic extraction of intonation structure from utterances is conducted on multilingual speech corpora amounting to 28.4 hours. Table 1 lists the dataset in more details. Both laryngograph signals and professional K-ToBI transcription [22] are available in the Korean speech corpus. Three tests are carried out. First, the detection of voice frames is evaluated by using the laryngograph signals. Second, the recovered $F_0$ trajectory is compared with the observed $F_0$ contours by ESPS get$f_0$ function in the Snack Sound Toolkit [23]. Third, the extracted virtual pitch register is evaluated by predicting K-ToBI break 3 (a strong phrasal disjuncture such as intonation phrase (IP) [22]). An algorithm of predicting IP break is as follows.

1. Predict IP break 3 at time $t_p$, when the $F_0$ trajectory digs into the virtual pitch register (PR), or at PR’s valley ($t_p$) with PR rising and/or falling magnitude $\geq 2$ semitones.

2. Search the nearest break index label to $t_p$, say, at $t_b$.

3. Compute accuracy, error rate, and mean span $|t_b - t_p|$.

### Table 1: List of multilingual speech corpora used in the work.

<table>
<thead>
<tr>
<th>Korean</th>
<th>Japanese</th>
<th>Chinese</th>
<th>Myanmar</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 F (female)</td>
<td>1 F</td>
<td>1 M</td>
<td>1 M</td>
</tr>
<tr>
<td>8.8 hours</td>
<td>4.5 h</td>
<td>0.6 h</td>
<td>3.5 h</td>
</tr>
</tbody>
</table>

The metrics in the second test are RMSE and Pearson’s correlation coefficients (hereafter, corre(lation)) of the recovered $F_0$ trajectory $(x_i)$ with respect to the observed $F_0$’s $(y_i)$, $i = 1, \ldots, n$ (the total number of voice frames upon voice detection).

\[
\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - y_i)^2}
\]

\[
\text{Corre(lation)} = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \sqrt{n \sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2}}
\]

Generally, RMSE indicates the average mismatch of two input contours, while correlation indicates the mismatch between the shape and alignment of the two contours.

### Figures 2, 3, 4, and 5 show examples of recovering $F_0$ trajectory in Japanese, Chinese, Myanmar, and Korean, respectively, and the corresponding virtual pitch register estimated by the method. In short, the recovered $F_0$ trajectory faithfully track the observed $F_0$ contours by conventional method. Also, it is not difficult to see from them that there exists clear intercorrelation of micro-prosody with phonetic segments.

Table 2 shows the result for voice frame detection. Compared to the conventional method (get$f_0$), the proposed method not only significantly suppresses unvoiced-to-voice error but also reduces voice-to-unvoiced error.

Table 3 gives objective test of recovering $F_0$ trajectory in comparison with the observed $F_0$ contours. Considering varying $F_0$ range of individual speakers, $F_0$ is converted to semitone using $12 \log_2(\frac{f_0}{176.0})$. Generally, the RMSE is small and the correlation is very high for each speaker. After removing the micro-prosodic effect, the mean of $F_0$ trajectory slightly increases compared to that of the observed $F_0$ contours.

Table 4 shows the performance of using the virtual pitch register to predict K-ToBI IP break 3. The mean span of prediction is 58 ms. Among the 25,253 break-3 samples, 78.8% are successfully predicted by using the extracted intonation structure (particularly the virtual pitch register), but 21.2% missed, such as the two “(3)” breaks in Fig. 5. Among the 25,532 predicted breaks, 22.0% are linked to break 2 (minimal phrasal disjuncture such as accent phrase [22]) (18.32%), break 1 (3.23%), or break 0 (0.51%). Basically, the results demonstrate that the extracted virtual pitch register agrees with the intonation structure labeled by linguistic experts.
Six observations are made from the work as follows.

- A ZFR-based filter has integration function as expressed in Eq. (3). A cascade of two ZFRs in Eq. (5), for example, is equivalent to successive integration four times [20]. The integration function of ZFR with the advantage of keeping impulse discontinuities [20] are useful for a task of recovering underlying trajectory from relevant sparse targets. In this sense, the ZFF method can be used as a super-smoother besides such applications as described in [24].

- It is worth noting that the proposed method is robust to some kinds of noise like white noise. This is because the method is built upon ZFF signals in which the high-frequency components are effectively filtered out as demonstrated in Fig. 1 (top panel).

- According to the results shown in Table 2, output from get_f0 is sometimes quite noisy, although get_f0 is more or less the industry standard and very well proven. In this test, there exist 2.78% voice-to-unvoiced (V2U) error and 9.23% unvoiced-to-voice error. In HMM-based speech synthesis, for example, V2U error has negative impact on the quality of synthetic speech. So it is very positive to see a significant improvement in V2U error; benefiting from the epoch-based voice detection.

- In [19], some “new” forms of expressive intonation in Japanese were discussed, namely, initial rise plus gradual decaying (form A), high plateau (form B), and gradual rise plus sharp fall (form C). An informal inspection of the extracted virtual pitch register shows that these forms are also quite common across languages. For examples, form A occurs in Figs. 3 (tone language) and 5 (pitch accent language), form B in Fig. 2 (pitch accent language) and Figs. 3 and 4 (tone language), form C in Figs. 2 (pitch accent language) and 4 (tone language). Further work is needed at this aspect.

- This method provides a means of decomposing $F_0$ contours into micro, accent/tone, and register components as done in [19][4]. Those methods in [19][4] were based on some assumptions made from Japanese, while the proposed method is language-independent.

- Observed $F_0$ contours are quasi-continuous and affected by micro-prosody. Due to HMM-based $F_0$ modeling at the state level, traditional MSD-HMM [25] has a limitation to tracking global $F_0$ behaviors and suffers from over micro-prosodic effects [26]. The proposed method makes it possible to apply the superpositional intonation synthesis [4] to, hopefully, overcome the limitation in multilingual speech synthesis.

### Table 2: Result for voice frame detection.

<table>
<thead>
<tr>
<th>Language</th>
<th>Method</th>
<th>#Voice frame</th>
<th>#Unvoiced fra.</th>
<th>#Unvoice fra.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Korean</td>
<td>get_f0</td>
<td>4.24293e+06</td>
<td>3.20146e+07</td>
<td>9.23%</td>
</tr>
<tr>
<td>Japanese</td>
<td>Proposed</td>
<td>97.72%</td>
<td>1.12%</td>
<td>1.16%</td>
</tr>
<tr>
<td>Japanese</td>
<td>Proposed</td>
<td>87.99%</td>
<td>2.78%</td>
<td>9.23%</td>
</tr>
</tbody>
</table>

- V2U: Voice-to-unvoiced; U2V: Unvoiced-to-voice

### Table 3: Result for recovering $F_0$ trajectory.

<table>
<thead>
<tr>
<th>Language</th>
<th>Method</th>
<th>$F_0$ mean (semitone)</th>
<th>RMSE</th>
<th>CoRe</th>
</tr>
</thead>
<tbody>
<tr>
<td>Korean</td>
<td>get_f0</td>
<td>42.000</td>
<td>0.595</td>
<td>0.985</td>
</tr>
<tr>
<td>Japanese</td>
<td>Proposed</td>
<td>42.073</td>
<td>0.595</td>
<td>0.985</td>
</tr>
<tr>
<td>Japanese</td>
<td>Proposed</td>
<td>47.771</td>
<td>0.602</td>
<td>0.997</td>
</tr>
<tr>
<td>Chinese</td>
<td>M</td>
<td>24.755</td>
<td>0.702</td>
<td>0.989</td>
</tr>
<tr>
<td>Chinese</td>
<td>M</td>
<td>45.720</td>
<td>0.580</td>
<td>0.993</td>
</tr>
<tr>
<td>Myanmar</td>
<td>M</td>
<td>34.320</td>
<td>0.716</td>
<td>0.994</td>
</tr>
<tr>
<td>Myanmar</td>
<td>M</td>
<td>35.638</td>
<td>0.623</td>
<td>0.994</td>
</tr>
</tbody>
</table>

b Recovered $F_0$ trajectory by the proposed method.

### Table 4: Result for predicting intonation phrase break 3.

<table>
<thead>
<tr>
<th>#IP break</th>
<th>#Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>25, 253</td>
<td>25, 532</td>
</tr>
</tbody>
</table>

- In [19], some “new” forms of expressive intonation in Japanese were discussed, namely, initial rise plus gradual decaying (form A), high plateau (form B), and gradual rise plus sharp fall (form C).

### 4. Discussions

A novel method is proposed for extraction of intonation structure from speech signals and yields two outputs for an utterance. One is continuous $F_0$ trajectory free from micro-prosodic effect. The other is virtual pitch register, global pitch movement capable of capturing proper intonation. The experimental results in four languages indicate that the continuous $F_0$ trajectory can faithfully track $F_0$ contours extracted by conventional method. In comparison of professional K-ToBI transcription for a large-scale speech corpus, 78.8% of intonation phrase (IP) breaks are successfully predicted by the extracted intonation structures. The method is language-independent. In practice, this method provides a means for pitch decomposition to train superpositional HMM-based intonation model from a speech corpus. Future work shall include investigation of relations of virtual pitch register with linguistic structures and further improvement for multilingual speech synthesis.
6. References


