SINUSOIDAL+ALL-POLE MODIFICATION BASED SPECTRAL SMOOTHING FOR CONCATENATIVE SPEECH SYNTHESIS

Heng Kang, Wen-Ju Liu

National Laboratory of Pattern Recognition, Chinese Academy of Sciences
{hkang, lwj}@nlpr.ia.ac.cn

ABSTRACT

The conventional LPC spectral smoothing algorithm causes an evident degradation in the speech quality when the smoothing amount is large. To improve speech quality of the smoothed speech, we proposed a new spectral smoothing algorithm. The source LPC spectral envelopes are first interpolated to generate the smoothed target spectra. Then the sinusoidal + all-pole modification is performed on the source speech to get the spectra of the modified speech which will coincide with the target spectra. Experimental results show that this method can get smooth spectral envelope even if the speech boundaries have large spectral distance. Experimental results prove that this algorithm is effective on avoiding degradation in quality of smoothed speech.

1. INTRODUCTION

In modern concatenative speech synthesis system with small or medium scale database, spectral smoothing often need to be applied when the speech segments have different spectral or formant structures. If the segments are concatenated directly, listeners can perceive noise or discontinuity, which makes the synthesized speech unnatural. In order to eliminate this phenomenon and make the synthesized utterances sound natural, effective spectral smoothing should be applied to the segment boundaries.

LPC smoothing[1] technology is frequently used in conventional speech synthesis system. This method first decomposes speech signal of segments into LPC parameters and residual, and then LPC and residual are smoothed separately. Finally the interpolated LPC and residual are re-combined to generate new frames. Pole Shifting[2] or LSF interpolation[3] algorithms can be used to interpolate LPC parameters. WI[4] be used to interpolate residual. Because LPC parameters are closely related with formant location, this method can manipulate formants easily and thus can get desirable smooth spectral envelope. But in another hand, interpolation LPC and residual in different ways causes evident degradation in speech quality when the smoothing amount is large[5,6].

In order to keep good speech quality of the smoothed speech, we propose a spectral smoothing algorithm based on sinusoidal + all-pole modification[7]. First spectra of segment boundaries are interpolated to get the smoothed target spectral envelope. Then sinusoidal + all-pole modification is performed on the boundaries frames to get sinusoidal parameters of the modified speech frames, which will coincide with the target spectra. At last the smoothing frames are synthesized from the sinusoidal parameters. This method can get smooth speech with good quality, because it not only smooth the spectral envelope but also keep spectral detailed information of speech.

This paper first gives a short introduction to sinusoidal + all-pole modification. Then implementational issues of this algorithm are discussed in detail. Finally we give some experimental results to prove the effectiveness of this algorithm.

2. SINUSOIDAL + ALL-POLE MODIFICATION

Sinusoidal + all-pole modification is a high-quality speech modification algorithm based on sinusoidal parameters[8,9]. The key point of this method is to...
modify the sinusoidal parameters \( \{a_k\} \) of speech segment \( x \), making the modified parameters coincide with a target spectral envelope \( S' (\omega) \).

Let us assume that speech \( x \) can be analyzed to sinusoidal parameters \( \{a_k\} \) and all-pole spectral envelope \( S(\omega) \). First a frequency warping function \( f: \omega \rightarrow \omega' \) is used to map points of \( S(\omega) \) to points of \( S'(\omega) \). In [7] Johan proposed a piecewise linear function as the frequency warping function, which is determined by dominant poles of the two spectra. Fig. 1 shows the piecewise linear frequency warping function. Then \( S'(\omega) \) is nonuniformly sampled at frequency points \( \omega_i = f(k\omega_0) \), which means

\[
b_i = S(\omega_i) = S(f(k\omega_0))
\]

Next, spectral detailed information in original sinusoidal parameters \( \{a_k\} \) is mapped to corresponding locations of target spectrum, which means

\[
b'_k = \frac{a_k}{S(k\omega_0)} b_i
\]

Because parameters in equation (1) and (2) are all complex variables, the mapping algorithm can also be applied to sinusoidal phase \( \phi_k \) as well as amplitude \( A_k \).

Finally \( \{b'_k\} \) should be resampled at the target fundamental frequency \( \omega_0 \) to make the new sinusoidal parameters \( \{a'_{k}\} \) coincide with harmonics of the target spectrum. To do the resampling, cubic interpolation on \( \log(b_i) \) and linear interpolation on unwrapped phase \( \angle b_i \) are applied.

3. SPECTRAL SMOOTHING METHOD

3.1. Process of algorithm

In order to smooth the segment boundaries, smoothing frames need to be added between the concatenation boundaries. Modern speech synthesis system requires not only the spectral envelope evolve smoothly in the smoothing frames but also speech quality of the interpolated frames be high.

Figure 2 shows the flow chart of our spectral smoothing method. First all-pole spectra of segment boundaries are interpolated to get the target spectral envelope of the smoothing frames. Then sinusoidal + all-pole modification is performed on the boundaries frames to get sinusoidal parameters of the smoothing frames. This modification can ensure that the sinusoidal parameters coincide with the target spectra. At last the smoothing frames are synthesized from the sinusoidal parameters. In another word, the all-pole spectral envelope is only target spectrum. The smoothing frames are generated from sinusoidal parameters.

3.2. Computing target spectra of smoothing frames

In our algorithm, DFW interpolation[10] is used to compute target spectral envelope of smoothing frames. DFW can interpolate LPC parameters very quickly, and furthermore, spectral matching pairs in interpolation process can be used to get the frequency warping function, which is very import to sinusoidal + all-pole modification.

DFW first try to match the frequency points on LPC spectra of the two frames to be smoothed. In the matching process, DFW minimizes the sum of local distances between all matching pairs. After we get the matching pairs, linear interpolation is performed to get the target spectra of smoothing frames.

We assume that \( S'(\omega) \) and \( S''(\omega) \) are LPC spectral envelope of the two frames to be smoothed, respectively.

\[
S'^1 = \{s'^1_1, s'^1_2, s'^1_3, ..., s'^1_N\}
\]

And

\[
S'^2 = \{s'^2_1, s'^2_2, s'^2_3, ..., s'^2_N\}
\]
in which, $s_n$ are series of points represented spectral amplitudes and distributed on frequency axis $\omega$. That is

$$s_n = S(\omega_n)$$

$\omega_n = n\pi/N$

And we assume that the matching pairs series is

$$P = \{p(1), p(2), \ldots, p(K)\}$$

in which $K$ is number of pairs and $p(k) = (i(k), j(k))$ represents the $k$th pair is $s_{i(k)}$ and $s_{j(k)}$.

We define local distance between a matching pair $s_{i(k)}$ and $s_{j(k)}$ as the difference of derivative magnitude between them:

$$d(i, j) = |\Delta s_{i} - \Delta s_{j}|$$

Searching routine of DFW is very similar to DTW, which is commonly used in speech recognition. After searching we can get optimized matching pairs series $P_{opt}$, which minimizes sum of distances between all matching pairs. And then linearly interpolate the spectra.

$$\omega_i'(a) = \omega_{i(k)} \cdot (1 - a) + \omega_{j(k)} \cdot a$$

$$S_k'(a) = S_{i(k)} \cdot (1 - a) + S_{j(k)} \cdot a$$

where $1 \leq k \leq K$, $0 < a < 1$, $S_k'(a)$ and $\omega_i'(a)$ are target spectra amplitude and the corresponding frequency, respectively. Fig. 3 shows the spectral envelope matching and interpolation.

Because sinusoidal + all-pole modification only needs the target LPC spectrum, we won’t convert the target spectrum into target LPC parameters. This can reduce much calculation in comparison to original DFW smoothing algorithm.

### 3.3. Frequency warping function

Frequency warping function plays a very important role in sinusoidal + all-pole modification process, because it is the bridge between the frames to be smoothed and the target spectra. Only through this warping function, can we combine them together and perform the modification. Original sinusoidal + all-pole modification uses a piecewise linear function, whose turning points are dominant poles or formant frequencies. The piecewise linear function looks easy to be understood, but it has some problems in practice:

1. Turning points are determined by formants frequencies. This requires computing speech formants. As we all know, formants cannot be determined very precisely in state-of-art technology.
2. Formants of two segments are required to be matched one by one, and then we can get the piecewise function. This is a difficult problem because sometimes formants cannot be matched one by one.
3. Piecewise linear function is formally too simple and behaves not well in our experiments.

Considering all the facts above, we try to give a simple and general method to compute the frequency warping function.

In fact, in last sector of this paper, we have got the relationship between the spectra to be smoothed and the target spectra. This relationship is the frequency warping function.

![Comparison between two frequency warping functions](image4.png)
We assume that S1 and S2 are the spectral envelope of frames to be smoothed, respectively. And the optimized matching pairs series is $P_{opt}$. After interpolation, $S_1'(a)$ and $\omega^I(a)$ are one target spectra amplitude and the corresponding frequency, respectively. When we modify spectra $S_1'$ to $S_1^I(a)$, we use the frequency warping function:

$$f_1^{I-x} = \{(x, y) | x = \omega_{a(k)}, y = \omega_1^I(a), \bar{\gamma}k < K\}$$

Similarly, when we modify spectra $S_2$ to $S^I_2$, we use the frequency warping function:

$$f_2^{I-x} = \{(x, y) | x = \omega_{a(j)}, y = \omega_2^I(a), \bar{\gamma}k < K\}$$

As we can see from fig. 4, the DFW determined warping function can precisely depict the matching relationship between two spectra contour. This will improve the fineness of modification and thereby improve the speech quality of smoothed segments.

**4. EXPERIMENTAL RESULT**

To validate the effectiveness of this spectral smoothing method, we did some experiments to test both the smoothing effect and speech quality of the smoothed segments. When we test the smoothing effect, ANBM distance measure is used for our criterion (the smaller the ANBM score, the better the smoothing effect). We used a male speaker speech database as our experimental material. First we segmented all utterances to speech segments according to the boundaries labels. Then we randomly selected about 60,000 segments to form 30,000 segments pairs. Statistical ANBM scores of all this 30,000 pairs can be seen in Fig. 5.

First, we performed our smoothing algorithm on boundaries of all the segments pairs, and computed the ANBM scores between boundaries of the segments pairs, before and after the smoothing. We give a statistical result of the change in scores. The result shows that the average ANBM scores dropped 24.54% after smoothing. Fig 6 shows the percentage of ANBM scores reduction of all segments pairs after smoothing.

From this result, we can draw a conclusion that the proposed method can well improve the smoothness of the segment boundaries, especially when the ANBM score is large. But as we can see from Fig. 6, this method behaves not well when the ANBM score is too small (below 200). The reason is that the distortion introduced by the modification counteracts the smoothing effect.

In addition, to test the speech quality of the smoothing segments, we also made a small scale of subjective listening test. We used LSF interpolation to smooth all the segments pairs, and compared the speech quality to the segments smoothed by our method. Because the workload of listening test is too large, here we only give part of the comparison results of some frequently appeared pairs.

Table 1 shows that speech quality of our method is better than that of LSF interpolation in most cases. Sometimes our method behaves worse than LSF, or has the same preference. This is because that this smoothing method still has some limitation or defect. Firstly, DFW based interpolation sometimes cannot get the proper spectral envelope, whose formants are expected to evolve smoothly. The improper interpolation results in bad target spectrum. When we modify frames to fit this target, the result sounds bad in quality. Secondly, we use LPC spectrum as our target. As LPC is an all-pole model it cannot model adequately nasal consonants or nasalized vowels. This leads to bad results when one of the segments to be smoothed is nasal.

<table>
<thead>
<tr>
<th>Former segment</th>
<th>Latter segment</th>
<th>Number of pairs</th>
<th>Preference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>LSF interpolation</td>
</tr>
<tr>
<td>b</td>
<td>a</td>
<td>10</td>
<td>30%</td>
</tr>
<tr>
<td>a</td>
<td>y</td>
<td>10</td>
<td>30%</td>
</tr>
<tr>
<td>y</td>
<td>ou</td>
<td>10</td>
<td>80%</td>
</tr>
<tr>
<td>w</td>
<td>ei</td>
<td>10</td>
<td>40%</td>
</tr>
<tr>
<td>ong</td>
<td>g</td>
<td>10</td>
<td>30%</td>
</tr>
<tr>
<td>m</td>
<td>ang</td>
<td>10</td>
<td>50%</td>
</tr>
<tr>
<td>n</td>
<td>i</td>
<td>10</td>
<td>40%</td>
</tr>
</tbody>
</table>
5. DISCUSSION

In general, we proposed a spectral smoothing method based on sinusoidal + all-pole modification. This method can remove unsmoothness of concatenation points and result in good speech quality in most cases. Especially when the spectral distance is large between the segments boundaries, this method get expected results.

But in some cases, this method cannot behave better than a pure LPC smoothing. This is because of the limitation of this model. Therefore we advise that some other smoothing methods should be used considering the context of the segments boundaries. When the spectral distance is large, our method can be applied to reduce the spectral difference effectively. When the spectral distance is not big enough, methods in spectral domain should be avoided. In this condition methods like optimal coupling[12] can get expected result.

In [7], the authors also suggest that we should study carefully the type of segments (nasal, vowel, fricative, etc.) and smoothing amount. And then we can determine what method to use.

6. REFERENCES