Piecewise Linear Stylization of Pitch Via Wavelet Analysis

Dagen Wang, Shrikanth Narayanan

Viterbi School of Engineering
University of Southern California
dagenwan, shri@usc.edu

Abstract

In this paper, we propose a wavelet analysis based piecewise linear stylization of the pitch trajectory. We also address the often-faced difficulty in handling the tradeoff between mean squared error and the number of lines used for fitting, where a heuristic approach is typically used to make the stylization choice. We pose the piecewise linear stylization task as a minimization problem by defining a penalty function that is a linear combination of the stylization mean squared error and line number to seek an optimal tradeoff. The weights for the penalty function are selected in a semi-supervised way using a development set. We also provide an objective statistical measure based on such penalty function for evaluating the performance of the stylization problem. Results show that our algorithm provides 16.7% penalty reduction than the baseline system based on heuristics. Also, we found that the wavelet decomposition combination selection approach outperforms the low pass level selection approach by 15.6%.

1. Introduction

Prosody, especially that conveyed by pitch dynamics, plays an important role in human speech communication [1]. It has been widely investigated in many spoken language processing areas. Text to speech synthesis systems strive to better assign and realize pitch contours to produce more natural utterances. Natural language understanding system uses prosody to disambiguate structural or scope confusions. Dialog management systems look for prosodic cues to predict speech acts and hierarchical discourse structure. Some studies also use prosody in automatic speech/speaker recognition and emotion recognition.

1.1. Pitch abstraction

Generally, the estimate of pitch from the speech signal has fluctuations on the calculated pitch contours due to both estimation errors and inherent unstable vibration of the vocal folds [2].

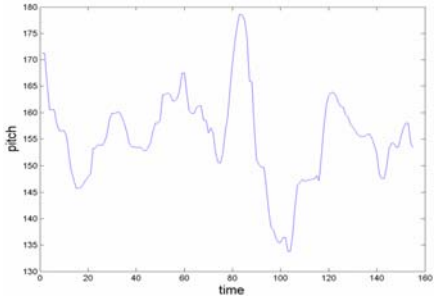

Figure 1: Sample pitch contour for the utterance "You know Jeans whatever um" (from switchboard sw2008B)

The pitch contour as illustrated in figure 1 has 2 types of variations: Type I includes large-scale variations related to intonation and carries various linguistic information. Type II refers to the small variations attached to type I resulting from estimation error, unstable vibration and microprosody. Even though type II variations carry some information to make speech natural in some speech synthesis applications [3], such information is very difficult to model and most of the times deemed redundant in speech analysis systems.

In order to get the type I information, and also for the purposes of computational linguistic studies, pitch contours are normally transcribed or parameterized. It could be either low level descriptions such as the Fujisaki model [4], the Hirst model [5], RFC model [6], and the tilt model [11] or high level descriptions such as the IPO model [9], or even phonological systems such as Pierrehumbert's [7], Ladd's [8] and the TOBI system [10].

Taylor [11] listed the 3 major requirements for such (type I) transcription: constrained, wide coverage and linguistically meaningful. However, there is never a consensus which system is suitable for achieving these requirements, even for the widely referenced TOBI system [12].

1.2. Why piecewise linear stylization?

Though various labeling schemes have been proposed and used in many speech systems, prosody plays relatively a limited role in such applications in spite of the useful information it carries. The major challenges come not only from the transcription scheme, but also the transcribing process itself. Such transcriptions demand a lot of resources and it is often very difficult to get large amounts of data to do statistically meaningful studies. Moreover, much of the transcription process is very subjective, and limited annotation agreement restricts the usage of prosody [12].

There is increasing interest among researchers to detect the prosodic events and labels automatically [11] [13] [14]. Statistical learning approaches like CART and HMM are applied on training data. However, bootstrapping labeling efforts, data type restrictiveness, and limited accuracy still pose major problem s for such approaches.

Interestingly, many successful speech applications use piecewise linear stylization of pitch as a prosodic feature. This method has been successfully used in a variety of contexts such as the study of sentence boundary [18], disfluency [17], dialog act [19], and speaker verification [16]. One common feature of these studies is that they all require large amount of data to do statistical learning and hence unsupervised (pitch) transcription becomes necessary. Piecewise linear stylization has been found suitable for this task.

It is obvious that piecewise linear stylization is a highly approximating process. Surprisingly, Hart [20] found that "..a piecewise linear approximation of an F0 curve in speech is, perceptually, not inferior to an approximation by means of fragments of parabolas, which gives—visually, at least--a better fit to the original F0 curve than does the rectilinear

approximation." Piecewise linear approach is also intuitively good at conveying type I information in the pitch contour if properly fitted.

1.3. How to do piecewise linear stylization?

Piecewise linear stylization could be formulated as the following mathematical function [16]:

$$g(x) = \sum_{k=1}^{K} (a_k x + b_k) I[x_{k-1} < x \le x_k] \quad (1)$$

Here a_k and b_k are the slope and intercept of each line. x is defined on the voice region. K is the total number of lines.

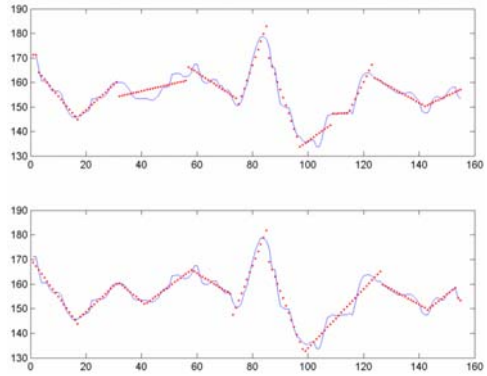

Figure 2*: Sample piecewise linear stylization*

Figure 2 illustrates two possible choices for stylization. The key problem is that we need an optimal criterion to choose one stylization from among all possible choices. In this context, people have defined various criteria.

In [16], K is chosen proportional to the duration of pitch. Then the stylization is determined by minimization of the mean squared error. Apparently, the choice of K in this approach is heuristic and not derived from the nature of the pitch contour.

In [2], piecewise linearization is taken within the syllable range. K=2 is fixed in the stylization process. This setting greatly simplified the process. However, prosody is a suprasegmental feature and such an approach will not be natural for most speech analysis applications.

In this paper, we propose an approach to select K optimally with respect to the pitch contour behavior by wavelet analysis. Meanwhile, a penalty function is setup to statistically evaluate the piecewise linear stylization process.

2. Algorithm Description

By analyzing pitch contours, such as the one shown in Figure 1, we concluded that if there is a high degree of type I variations, K should be chosen larger and vice versa. Also, we note that wavelet decomposition is particularly suited for doing such multi-scale analysis of the signal variations. So we introduce wavelet analysis for pitch stylization.

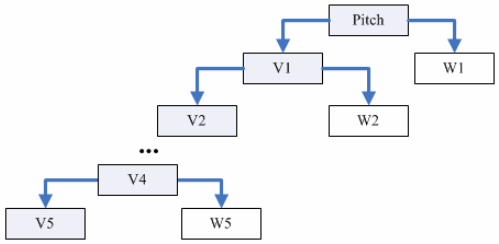

Figure 3: 5 level wavelet decomposition

2.1. Wavelet analysis of pitch contour

We do a five-level signal decomposition by wavelet analysis as illustrated in Figure 3. Figure 4 illustrates this process.

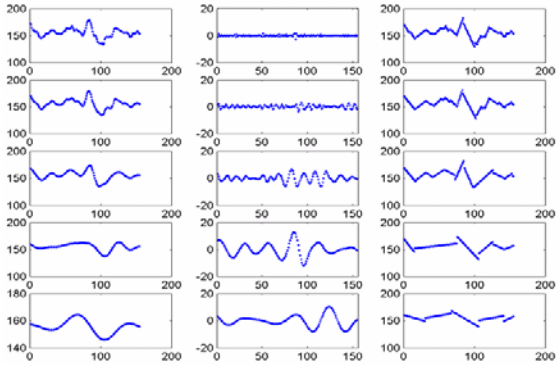

Figure 4: Db10 wavelet analysis of pitch contour

The left column in Fig. 3 is the V domain (low pass). The middle column is the W domain (high pass). The right column is the piecewise linear stylization. Row 1 to Row 5 are related to the level 1 to level 5 wavelet decompositions.

We make the following observations: as the decomposition level increases, the V domain analysis gives a smoother abstraction of the pitch contour and gets fewer lines for the stylization. Meanwhile, of course, the mean square error increases.

2.2. Mean squared error and line number

Preferably, we wish to choose a decomposition level to minimize the mean squared error of stylization to get a better fit to the pitch contour. At the same time, we also wish to minimize the line numbers to get better abstraction of the pitch contour in the sense of conveying more accurately the type I information.

In the previous section, we noted the relationship between expected mean squared error and line number. Unfortunately, we can not minimize these 2 factors simultaneously.

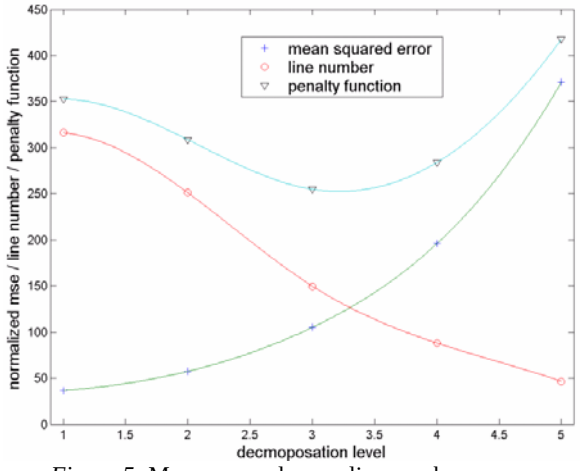

Figure 5: Mean squared error, line number versus decomposition level and their relation to the choice of ***Wt***

Figure 5 illustrates this relationship through a statistical analysis employing a large number of voiced segments. Each level is chosen independently and the mean squared error and line number are recorded respectively and finally statistically evaluated. The results of the figure confirm the observation that mean squared error and line number could not be minimized simultaneously.

2.3. Penalty function selection

Hence, we have to tradeoff between mean square error and line number. We define the following penalty function:

$$Penalty = (mse + Wt * line\#) / duration \quad (2)$$

By choosing a decomposition level to minimize the penalty function, we could handle the tradeoff with a single scalar ***Wt.*** Again, we have to answer how to set ***Wt*** optimally.

One objective method to aid and validate this process is to bring linguistic knowledge constraints wherein piecewise linear stylization of pitch is used to model prosody. However, how pitch dynamics determine various linguistic events is still a largely open problem for prosody researchers. Such relations are normally modeled as hidden channel parameters between pitch contour and various linguistic events. These parameters again need a suitable model of pitch to do the learning. So we have a "chicken-and-egg" loop and such problems tend to be difficult to solve.

Furthermore, pitch contour modeling requirements vary depending on the application. For example, for segmentation and end of utterance problem, the later part of the voiced region in an utterance is the focus [18] while for speaker identification, global variations are taken into consideration [16]. These two systems would hence require different evaluation focus on pitch stylization.

Hence, we decided to use human intervention in specifying this penalty measure. It is noted that manual piecewise linear stylization of pitch is not expected to be good. For instance, the two cases in Figure 2 could be two possible results from human transcribers. Overall, we could expect a very low agreement for direct manual piecewise linear transcription of pitch.

However, wavelet analysis as illustrated in Figure 4 makes consistent, objective human transcription possible. We found that it was relatively easy for a human to systematically choose a level from the 5 levels which best conveys the type I information. The method was simple since the human did not need to handle every local portion of the pitch trajectory heuristically but chose a match from one of the 5 levels. By collecting the statistics of the labeling on a development set, we are able to set ***Wt*** optimally to reflect the observations. (See Section 3 for details).

2.4. Evaluation

With the penalty function defined by ***Wt***, we could evaluate pitch stylization in large amounts of (voiced) data statistically. Different approaches could be compared with respect to the statistical mean of their penalty function value to determine the performance. Also, such a statistically meaningful measure would help to set the parameters of the piecewise linear stylization algorithm.

3. Data and Results

3.1. Data description

We chose 690 voiced regions from the switchboard corpus [21]. Each region is at least 1.5 second long (150 frames). The pitch contour was calculated using the ESPS get_f0 utility and median-filtered with a length-3-window to remove very sudden spikes.

3.2. Determine *Wt* in penalty function

The very first step is to determine ***Wt*** in the penalty function. This process was described in the previous section and illustrated in Figure 6. After the 5 level wavelet analysis, change points (local maxima or minima) are located in each level of the V domain. The piecewise linear stylization is performed in each region bounded by such change points. Then the normalized mean squared error and line number are recorded.

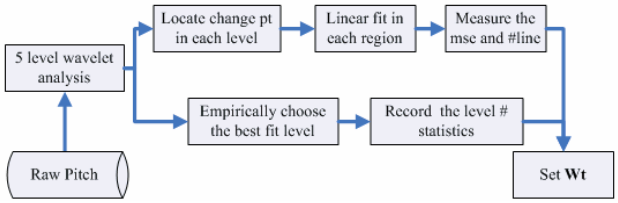

Figure 6: The process to estimate ***Wt***

On the lower branch in Figure 6, the same 5-level wavelet analysis as in Figure 4 is provided to the human transcriber. The transcriber then chooses the best level to reflect the type I information. For the development set, we performed manual transcription on 100 randomly chosen voiced regions and the distribution statistics are given in Table 1.

Level 1	0
Level 2	11
Level 3	60
Level 4	27
Level 5	2

Table 1: Optimal decomposition by human choice (on the development set)

It could be concluded that level 3 dominates most of the decomposition level and with level 4 coming second to it.

Now compare this to Figure 5 which illustrates that the penalty function value for each level forms a parabolic shape with its minimal point providing the optimal choice of the decomposition level. By setting ***Wt=1800***, we get the minimal point close to level 3 and a little tilted to level 4. This indeed is in agreement with the observation we made on Table 1 that was based on human selection.

3.3. Evaluation results

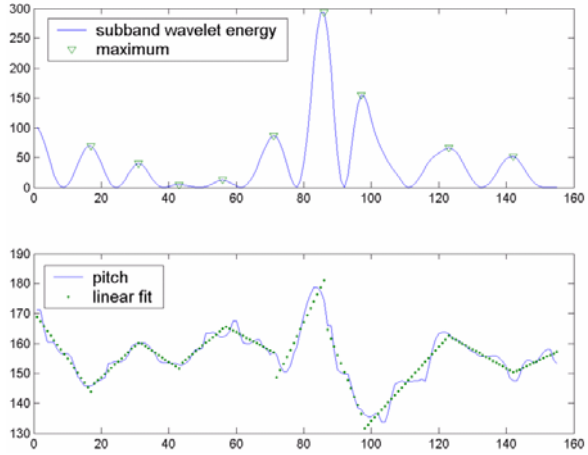

Figure 7: W domain piecewise linear stylization

With the value of ***Wt*** fixed, we specified the penalty function as defined by Equation (2). We evaluate our algorithm both in the V domain and W domain of wavelet analysis. For every voiced region, we first performed a 5 level piecewise linear

stylization in V (low pass) domain as illustrated in Figure 3 and Figure 4. Then 5 pairs of mean squared error and line number (and hence the penalty values) are obtained. Then we choose the level with minimum penalty and keep it in record.
We next try to achieve the same goal in W (high pass) domain. In this case, we are not locating the change points as in V domain (see Figure 6), but the maximal point in the wavelet energy curve (See Figure 7). Such curve is the energy from a subset combination of the wavelet bands. Such combinations are {W5}, {W5, W4}, {W5, W4, W3}, {W5, W4, W3, W2}, {W5, W4, W3, W2, W1}. (See Figure 3) Again, we choose one of the combinations that minimizes the penalty function.
By performing the above process both in V and W domain on all 690 utterance samples (voiced regions only), we compute the mean of the penalty value and list them in Table 2:

	V domain	W domain
Baseline Penalty	255	221
Penalty by optimal selection	218	184

Table 2: Final evaluation result

The baseline result here is based on presetting one fixed optimal decomposition level for all the voiced regions. For example in the V domain analysis baseline, level 3 was chosen as the universal decomposition (based on information in Table I) instead of choosing an optimal level for each data segment. We see the penalty reduction as much as 16.7%, as seen for the W domain. Also, the W domain analysis demonstrates more ability for this piecewise linear stylization task than V domain (by as much as 15.6%).

4. Conclusions

Evaluation results in Table 2 show that the proposed algorithm could automatically select the optimal decomposition level to effectively tradeoff between mean squared error and line number in piecewise linear stylization. This algorithm is also a fully automatic once the penalty function is defined using a small development set. As described in Section 3, the penalty function settings need some human intervention. However, the subjectivity of human annotation is minimized by allowing the transcriber to select the optimal decomposition level match instead of doing a full piecewise linear stylization on the raw pitch contour.
We also find that W domain analysis outperforms V domain analysis by almost 15.6% in relative penalty reduction. However, we only used the V domain analysis for the ***Wt*** setting experiment. One reason was that the W domain enumeration has a larger potential number of the combinations of 5 wavelet bands (Wi) which makes it more difficult for the transcriber to make a consistent decision. On the other hand the low pass V domain provides more intuitive reconstruction curves for the transcriber to match against. More importantly, the purpose of ***Wt*** setting is just to determine the weights for optimal tradeoff between mean square error and line number.
To keep this work general, we did not involve any specific application. Incorporating the results of the pitch stylization within a spoken language understanding application, however, is a topic of our ongoing work.

5. References

[1] J. Hirschberg, "Communication and prosody: functional aspects of prosody", *Speech Communication*, Vol. 36: 31-43, 2002

[2] J-C. Lee, Y. Kim, S-H. Hahn, Minsoo, "Intonation processing for TTS using stylization and neural network learning method", *ICSLP*, 1996

[3] A.I.C Monaghan, "Extracting Microprosodic Information from Diphones -- a Simple Way to model Segmental Effects on Prosody for Synthetic Speech", *ICSLP*, 1992

[4] Hiroya Fujisaki, "Dynamic characteristics of voiced fundamental frequency in speech and singing," *in The production of speech*, Ed. Springer, Berlin, 1983.

[5] Hirst, D., "Prediction of prosody: An overview", In Bailey, G. and Benoit, C., editors, *Talking Machines*, North Holland, 1992

[6] Taylor, P. A. "The rise/fall/connection model of intonation", *Speech Communication*, 15:169–186, 1995

[7] Pierrehumbert, J. B, "The Phonology and Phonetics of English Intonation". *PhD thesis, MIT. Published by Indiana University Linguistics Club*, 1990

[8] Ladd, D. R, "Intonational Phonology", *Cambridge Studies in Linguistics*. Cambridge University Press, 1996

[9] t'Hart, J. and Collier, R. "Integrating different levels of intonation analysis", *Journal of Phonetics*, 3:235–255, 1975

[10] Silverman, et al. "ToBI: a standard for labelling English prosody", *In Proceedings of ICSLP*, 1992

[11] P. Taylor. "Analysis and synthesis of intonation using the tilt model", *Journal of the Acoustical Society of America*, 107(3):1697-1714, 2000

[12] Wightman, Colin W, "ToBI or not toBI? ", *In Speech Prosody*, 25-29, 2002

[13] C. Wightman and M. Ostendorf, "Automatic Labeling of Prosodic Patterns", *IEEE Trans. Speech and Audio Proc.*, vol. 2, no. 4, pp. 469-481, 1994.

[14] S. Ananthakrishnan and S. Narayanan. An Automatic Prosody Recognizer using a Coupled Multi-Stream Acoustic Model and a Syntactic-Prosodic Language Model. In Proc. ICASSP, Philadelphia, PA, March 2005.

[15] D. Baron, E. Shriberg, and A. Stolcke, "Automatic punctuation and disfluency detection in multi-party meetings using prosodic and lexical cues," *ICSLP*, 2002

[16] K. Sonmez, E. Shriberg, L. Heck, and M. Weintraub, "Modeling dynamic prosodic variation for speaker verification", *ICSLP*, Sydney, Australia, August 1998.

[17] E. Shriberg, R. Bates, and A. Stolcke, "A prosody-only decision-tree model for disfluency detection, " in Proc. EUROSPEECH, vol. 5, pp. 2383--2386, September 1997

[18] E. Shriberg, A. Stolcke, D. Hakkani-Tur, & G. Tur, "Prosody-based automatic segmentation of speech into sentences and topics", *Speech Communication* 32 (2000), pp. 127-154. Special Issue on Accessing Information in Spoken Audio.

[19] A. Stolcke, et al. "Dialog act modeling for conversational speech," in *Applying Machine Learning to Discourse Processing. Papers from the 1998 AAAI Spring Symposium*. Technical Report SS-98-01 (J. Chu-Carroll and N. Green, eds.), (Stanford, CA), pp. 98--105, AAAI Press, Menlo Park, CA, March 1998

[20] 't Hart J, "F0 stylization in speech: straight lines versus parabolas", *J Acoust Soc Am*. 1991 Dec;90(6):3368-70

[21] J. Godfrey and E. Holliman, "SWITCHBOARD-1 Release 2", *LDC97S62*, 1997