A SUPERPOSED PROSODIC MODEL FOR CHINESE TEXT-TO-SPEECH SYNTHESIS

Gao-Peng Chen[*], *Gérard Bailly*[#], *Qing-Feng Liu*[*], *Ren-Hua Wang*[*]

[*] Iflytek Speech Lab, University of Science & Technology of China
[#] Institut de la Communication Parlée - CNRS/INPG/U3

ABSTRACT

The paper presents the application of the trainable SFC superpositional prosodic model to Chinese. Within the SFC model, prosodic parameters (F0, syllabic lengthening) are interpreted as the superposition of overlapping multi-parametric contours. These contours are associated with high-level prosodic features operating at different scopes, such as tones, stress, prosodic boundary, part of speech of words, etc. Each feature label corresponds to a metalinguistic function (morphological, lexical, syntactic, attitudinal…) which is represented by a neural network. The observed contour is the sum of the outputs of the corresponding neural networks. An analysis-by-synthesis scheme is implemented for automatically learning. This model works well in the concatenation of neighbored units. The RMSE of F0 prediction is 2.34st (referenced to 200Hz), correlation is 0.86. Perceptual experiments show that the predicted prosody is quite appropriate and fluent.

1 INTRODUCTION

The fundamental problem for intonation analysis and synthesis is that the variation of F0 as a function of time is the acoustic correlate of a number of linguistic prosodic features. Two major classes of intonation models have evolved in the past two decades.

Superpositional models interpret F0 as complex patterns resulting from the superposition of several components. Fujisaki model [5] is the typical model in this class, which decomposes F0 into phrase component and accent component. The parameters are associated with the mechanism of pronouncing, which is quite relevant to the macro-prosodic features. It has been tried on many languages including Chinese [4, 9]. Due to the different characteristics between tone language and non-tone language, it is difficult to simulate tone events by accent components. Besides, to automatically extract the phrase commands and accent commands from observed F0 is not solved well. Even if the commands are labeled manually, we are not sure to make them consistent in different annotators. Other proposals [1, 6, 11] face also the problem of the ill-posed problem of analysis, i.e. decomposing an observed contours into elementary contributions. The SFC [2, 7] implements a prosodic model initially proposed for French [1] which introduces a new model-constrained, data-driven method to generate prosody contours with very few prototypical movements. The SFC introduces an original training paradigm using an analysis-by-synthesis framework that iteratively decomposes prosodic contours and builds the prosodic model *in the same time* (see §2).

On the other hand, there are models that claim that F0 contours are generated from a sequence of phonologically distinctive tones or categorically different pitch accents, which are determined locally. The typical ones are the Tilt model [10] in English, PENTA [12, 13] in Chinese. These models focus on local events, but they ignore the trait of prosody on a big unit, such as on phrase or clause.

Chinese is a tone language with high-level, low-rising, low-falling, high-falling and neutral tones. The tone events are very important to the prosody of an utterance. Each syllable that is the carrier of a tone and a basic meaningful phonetic unit normally is an individual target of prediction. However, sentence declination and phrasing are important as well. In this paper a superposed model is proposed to model Chinese prosodic contours, and the sequences of tones and phrases are both considered.

2 DESCRIPTION OF THE MODEL

SFC considers that the prosodic contour is the contribution of the prosodic features on different scope. We suppose that each feature effect on the prosody respectively and independently. And the contribution of a given feature is a function of scope or domain. The predicted/target contour is the superposition of corresponding functions using an appropriate scale (logarithmic for F0 and syllabic lengthening).

First of all, we collect the most important prosodic layers and assign the feature labels for each layer. SFC devotes a group of neural networks to represent a layer. Each neural network represents each feature label respectively within the local layer. The output of a layer is the output of the neural network of the corresponding feature occurring in the sentence. The predicted contour is the sum of each layer's contributions. The synthesis and analysis flowcharts are showed in Figure 1: they are combined in order to train contour generators. Each neural network

produces characteristic prosodic cues for each syllable as a function of the scope the function it implements, i.e. the input of a neural network is the scope of the units labeled by the prosodic feature (see Figure 2). The output of a neural network is a series of three-point F0 vectors and a lengthening factor for the current syllable. Each vector is the local neural network's contribution to the observed F0 on beginning, middle and end of the syllabic nucleus together with the z-scored syllabic duration. The reference syllabic duration is computed as a weighted sum of a constant duration (tendency to isochronous syllables) and the sum of the mean durations of its phonemic constituents (long/short segments result in longer/shorter syllables).

SFC uses SNNS [14] to implement these neural networks. So one neural network will output one stable pattern that just varies with different time domains [see an analog approach in 8]. The number of neural networks determinates the diversification of the predicted contours. If we just apply the few most important features in this model, few materials are enough to get the functions of the features. It is the advantage of this model to need less data. All the prosodic features have the same influence on the predicted contours in the same context environment. So all the sentences are decomposed and put into training together. By back-propagation, the global solutions are worked out within the training set.

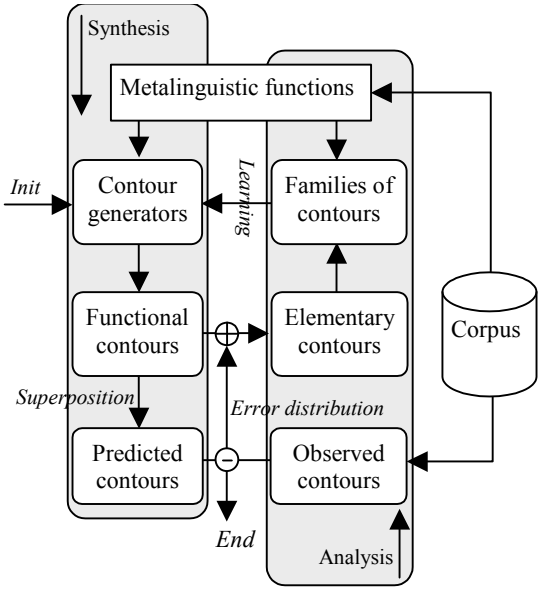

Figure 1: Analysis-by-synthesis loop. Each contour generator implement a given metalinguistic function parameterized by its scope. SFC generators are trained using patterns built by adding to what they already predict a proportion of what they all together do not still predict, i.e. the difference between observed and predicted contours at the iteration considered. The learning loop stops when this difference do not diminish significantly.

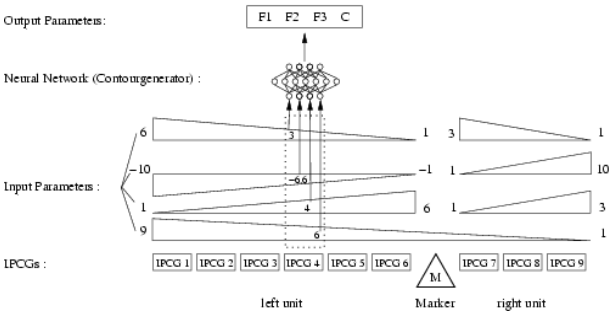

Figure 2: M is a contour generator(neural network) that converts linear ramps anchored on the boundaries of unit A and B into prosodic trajectories(F1..3 C).

3 SPEECH MATERIALS

Since this model doesn't need much data for training. We design 100 Chinese sentences spoke neutrally by a female. The texts are selected by greedy algorithm to cover the most phonetic and prosodic events. The first 20 sentences are around 50 syllables long, whereas the next 80 sentences are only 10 syllables long. The pitch contours are automatically calculated by Praat [3]. Some serious errors are corrected by hands. Segmentation is first determined by an HMM system, then corrected manually. Characteristics of the metalinguistic functions (tone types, phrasing… together with their scopes) are labeled by the professional annotators. 40 of the sentences were picked out stochastically for training, and the rest are for testing.

4 IMPLEMENTATION AND EVALUATION

In the implementation, we design four prosodic layers, a tone layer (or accent layer for English), a word layer, a phrase layer and a clause layer. Figure 3 is an example of the synthesis of prosodic contours.

Chinese has four basic tones and one neutral tone. So there are 4 individual tone markers (C1, C2, C3, C4,) and 16 coarticulated tone markers (C11, C12,…, C44) in the tone layer. To make the superposition more clear, we distribute the tone elements in two layers (Tone1 and Tone2 layers in Figure 3. In Figure 4 the first column shows the single tones' patterns. The others are the coarticulated tones' patterns. We can see that one tone followed by different tones has different patterns. These patterns are automatically generated by 20 small neural networks. In each pattern, the first half is the model of preceding tone influenced by different following tones. It retains the shape of original individual tone pattern but changes a little. The last half is the effect of preceding tone on the following

tone. It is waves up and down about zero. The superposition of two overlapped tone layers is the contribution of tone events to pitch contours. As for the neutral tone, it is too flexible to model it as a fixed pattern. Usually people don't care about it on perception. It can be interpolated by the preceding tone and following tone with a spline function or syllables carrying it just considered as part of the scopes of the adjacent non-neutral tones.

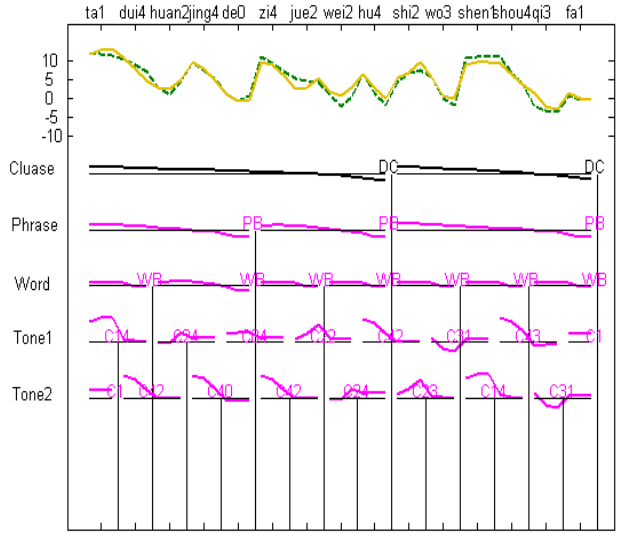

Figure 3: Prediction of a melodic contour as the superposition of the contributions of four layers. The dashed line is the observed F0. The solid line is the synthesized F0.

In word layer, word boundaries are marked. We have considered marking the part of speech of each word and the syntactic relation of neighbored words. But it just improves the prediction a little and makes the layer much more complicated. A single word boundary is enough in a small database. In the same way, there are phrase boundaries for the phrase layer, and declarative clauses constitute the clause layer. If there are some other utterances with different modes in the future, we can think about adding some more modal markers in the clause layer. The melodic patterns of Word Boundary, Phrase Boundary and Declarative Clause are illustrated in Figure 5. All of them are declining curves. It is consistent with the fact that the pitch contour declines within a phrase and clause for a statement.

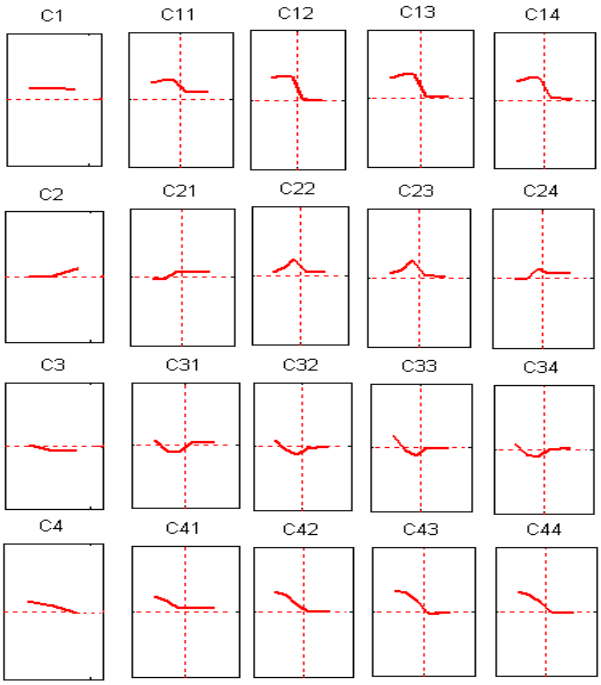

Figure 4: The individual and concatenative tones' patterns. They are the functions of Neural Networks in the Tone layers.

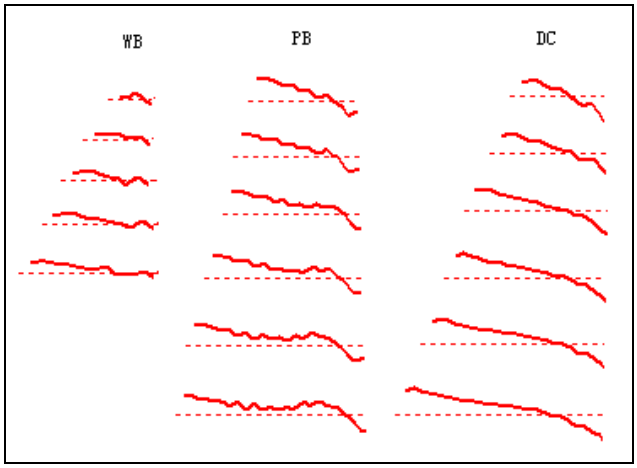

Figure 5: The melodic patterns for Word Boundary, Phrase Boundary and Declarative Clause in time domain.

In tone layer, there are 20 tone functions including 4 individual tones and 16 concatenative tones. There is one function respectively in word layer, phrase layer and clause layer. Each function is a neural network that is a TD-NN with a hidden layer realized by SNNS [14]. So there are 23 networks. They are learning together to get a global minimum error. For the 40 sentences of training set, the RMSE is 2.09st, correlation is 0.90. For the 60 sentences of test set, the RMSE is 2.34st, correlation is

0.86. We use these generated prosody contours to resynthesize sound files by PSOLA. They sound fluent and acceptable, but a little regular since no syntactic cues are implemented for the moment (see Figure 6): if the melodic contours are nicely predicted, rhythmic contours are expected to be more influenced by word/phrase chunks and syntactic hierarchy.

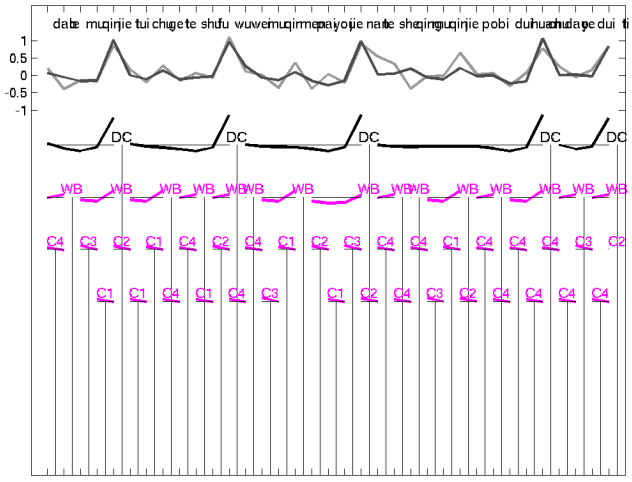

Figure 6: Prediction of a rhythmic contour as the superposition of the contributions of four layers. The light line is the observed F0. The solid line is the synthesized rhythm.

CONLUSIONS

This model can implement the training and prediction automatically. The global optimal solution is obtained within the training set. The experiments showed that the difference of the prediction between training set and test set is very small. The patterns of each prosodic feature give the contribution to the target contour in different scope. These patterns are consistent with our prior knowledge. In a big unit, the F0 contour declines from the beginning to the end in time domain. Since the patterns only depend on the length of scope, the synthesized contours from this model vary in the corresponding scope. They are very stable and fade. The prosody sounds flat and neutral. Because it is not a parameterized model, we can't product new prosodic events as the Fujisaki model by adjusting parameters. If we want to make the output contour more flexible, we must add more layers and more prosodic markers. Signaling syntactic structure is the first work in mind. We also may add some modal markers (Question, Imperative, Exclamatory) in the clause layer to distinguish different moods. We can compare the difference of neutral and emotional utterance and try to add an emotion layer. These are the further research that we will do next.

REFERENCES

[1] Aubergé, V. (1992) *Developing a structured lexicon for synthesis of prosody*, in *Talking Machines: Theories, Models and Designs*, G. Bailly and C. Benoît, Editors. Elsevier B.V. p. 307-321.

[2] Bailly, G. and Holm, B. (2002) *Learning the hidden structure of speech: from communicative functions to prosody*. Cadernos de Estudos Linguisticos, **43**: p. 37-54.

[3] Boersma, P. and Weenink, D. (1996) *Praat, a System for doing Phonetics by Computer, version 3.4*, in *Institute of Phonetic Sciences of the University of Amsterdam, Report 132. 182 pages*.

[4] Chen, G.P., Y.Hu, and Wang, R.H. (2004) *A Concatenative-Tone Model with its parameters' extraction*. in *International Conference on Speech Prosody*. Nara, Japan. p. 455-458.

[5] Fujisaki, H. and Hirose, K. (1985) *Analysis of voice fundamental frequency contours for declarative sentences of Japanese.* Journal of the Acoustical Society of Japon, **5**: p. 233-242.

[6] Gårding, E. (1991) *Intonation parameters in production and perception*. in *Proceedings of the International Congress of Phonetic Sciences*. Aix-en-Provence, France. p. 300-304.

[7] Holm, B. and Bailly, G. (2002) *Learning the hidden structure of intonation: implementing various functions of prosody*. in *Speech Prosody*. Aix-en-Provence, France. p. 399-402.

[8] Kochanski, G. and Shih, C. (2003) *Prosody modeling with soft templates.* Speech Communication, **39**: p. 311-352.

[9] Mixdorff, H., Fujisaki, H., Chen, G.P., and Hu, Y. (2003) *Towards the automatic extraction of Fujisaki model parameters for Mandarin*. in *EuroSpeech*. Geneva, Switzerland. p. 873-876.

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

[11] Thorsen, N.G. (1983) *Standard Danish sentence intonation - Phonetic data and their representation.* Folia Linguistica, **17**: p. 187-220.

[12] Xu, C.X., Xu, Y., and Luo, L.-S. (1999) *A pitch target approximation model for F0 contours In Mandarin*. in *nternational Congress on Phonetic Sciences*. San Francisco, CA. p. 2359-2362.

[13] Xu, Y. and Wang, Q.E. (2001) *Pitch targets and their realization: Evidence from Mandarin Chinese.* Speech Communication, **33**: p. 319-337.

[14] Zell, A., Mache, N., Sommer, T., and Korb, T. (1991) *Design of the SNNS neural network simulator*, in *Östreichische Artificial-Intelligence-Tagung*. Informatik-Fachberichte 287, Springer Verlag: Wien. p. 93-102.