International Symposium on Chinese Spoken Language Processing (ISCSLP 2004)
Hong Kong
December 15-18, 2004

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 prosody is the acoustic encoding of a large number of linguistic and paralinguistic features. Two major classes of intonation models have evolved in the past two decades.

Superpositional models interpret prosody as complex patterns resulting from the superposition of more simple 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 tonal and non-tonal languages, it is difficult to simulate tone events by accent components. Besides, the automatic extraction of the phrase and accent commands from observed F0 is not a solved problem. 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, phrases and clauses are all considered.

2 DESCRIPTION OF THE MODEL

Principles. SFC considers that the prosodic contour is the contribution of few basic metalinguistic functions (phonetic such as tonal distinctions, segmentation, salience, hierarchy…) acting on different units at various scopes. We suppose that (1) each function affect prosody by means of a function-specific multiparametric contour called *functional contour* (FC); (2) An FC is co-extensive to the units concerned by the function it implements - this extend is called the *domain* or the *scope* of the FC - and is independent from the other units or functions implied in the discourse structure; (3) the shape of a FC is only a function of its scope (and of course of the metalinguistic function it implements); (4) the predicted/target contour is the superposition of corresponding FC using an appropriate scale (logarithmic for both F0 and syllabic lengthening).

Functional contour generators. All FC implementing a given prosodic function are generated by a unique functional contour generator (FCG). FCGs are now

implemented as neural networks. They produce characteristic prosodic cues for each syllable of the function scope. as a function of characteristics of its scope. So one FCG will output one stable pattern that just varies with different time domains [see an analog approach in 8]. Input parameters characterize the position of the current syllable within the scope i.e. its relative and absolute position from/to landmarks such as beginning and end of the units (see Figure 1). The output of the neural network is a series of 4-dimentional vectors. Each vector expresses the local contribution of the FC to the observed F0 on beginning, middle and end of the syllabic nucleus and to the z-scored syllabic duration of the syllable. 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).

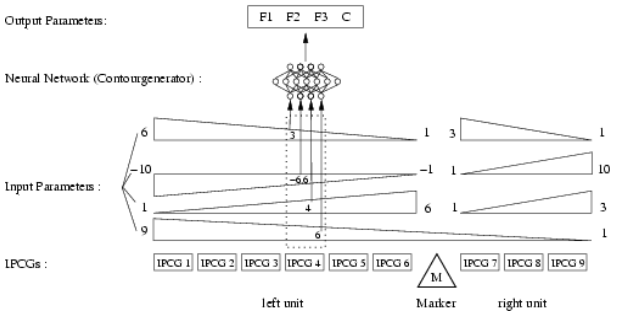

Figure 1: A contour generator (neural network) converts linear ramps anchored on the boundaries of unit A and B into prosodic trajectories(F1..3 and C).

Learning. The predicted/observed contour is the sum of the FCs implementing all metalinguistic functions acting on the different units of the discourse. Since there is no constraints on the FC shapes nor on scope and number of FCs involved, the learning of FCG is not straightforward and faces an ill-posed problem. The learning procedure is iterative and based on an analysis/synthesis loop (see Figure 2). At initialization, untrained FCGs just output zero FCs. Observed prosodic contours are then simply decomposed into elementary contours by considering for each syllable an equal contribution of all FCs contributing to the prosody of that syllable. All elementary contours implementing a given metalinguistic function are then combined in order to train the appropriate FCG. Only part of the observed variance could then be predicted since each FCG receives only phonotactic characterization of each FC. The prediction error is then distributed and these *unexplained* contributions are further added to the predicted FCs to form the next training set of elementary contours. The learning loop stops when the prediction error do not diminish significantly.

Richness. The diversification of the predicted contours mirror the diversity of the discourse structures and metalinguistic functions involved in the training corpus. This diversification is due to the complexity of embedded discourse structures and not by compensating a flat and linear phonological structure with context-dependent information. Since there is no limit in the degree of FC embedding, FCGs can collect dozens exemplars of FCs implementing the same functions within the same utterance. It is the advantage of this model to need few data and rely on the very strong assumptions of the prosodic model.

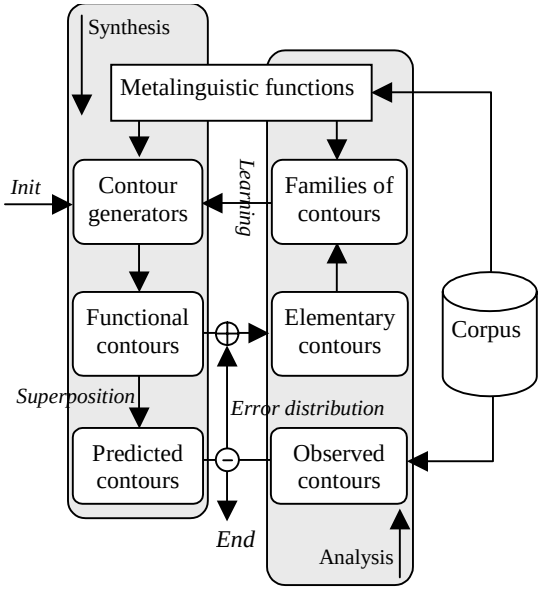

Figure 2: Analysis-by-synthesis loop. Each contour generator implement a given metalinguistic function parameterized by its scope. SFC generators are trained using patterns built by adding to what they already predict a proportion of what they all together do not still predict, i.e. the difference between observed and predicted contours at the iteration considered. The learning loop stops when this difference do not diminish significantly.

3 IMPLEMENTATION AND EVALUATION

Since this model doesn't need much data for training, we design 100 Chinese sentences spoken neutrally by two males and two females. 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.

In the implementation, we design functions intervening at four domains: a tone domain, a word domain, a phrase domain and a clause domain. Figure 3 is an example of the synthesis of prosodic contours.

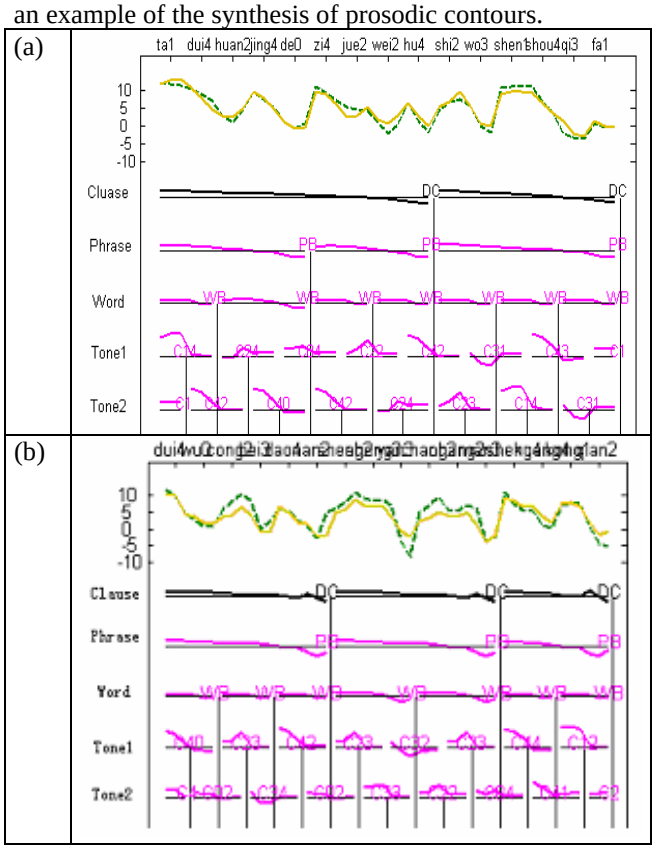

Figure 3: Prediction of the (a) melodic and (b) rhythmic contour as the superposition of the contributions of four domains. The top dashed and solid lines are respectively the observed and synthesized contours.

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 domain. To make the superposition more clear, we display the tone elements in two layers (Tone1 and Tone2) 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 all these tones constitutes the contribution of tone events to pitch contours. Syllables receiving the neutral tone are part of the scope the nearest following non-neutral tone: some scopes of concatenative tones' patterns extend two syllables.

For the word domain, we only considered the potential ability of prosody in signaling word boundaries. But this just improves the prediction a little. An additional distinction between word classes may perhaps be useful. Phrase boundaries are also considered. For instance we did not enrich the phrase-domain functions with any hierarchical information. Clauses are all declarative in the current training corpus: cueing clause boundaries correspond here with cueing declarative clause boundaries but more modal markers in the clause domain. The melodic patterns of Word Boundary, Phrase Boundary and Declarative Clause are illustrated in Figure 5. All of them are declining curves.

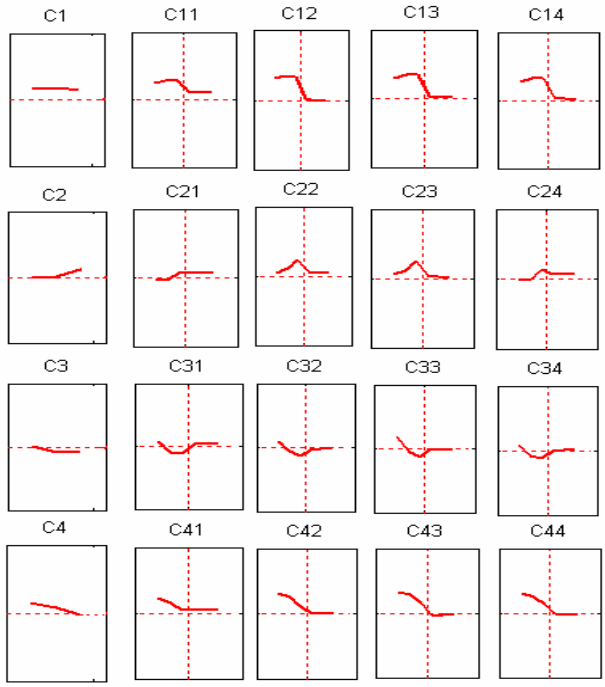

Figure 4: The individual and concatenative tones' patterns. They are the functions of Neural Networks in the Tone layers.

	Male A	Male B	Female A	Female B
	RMSE /Corr	RMSE /Corr	RMSE /Corr	RMSE /Corr
Training Set	1.75 / 0.92	1.60 / 0.94	1.82 / 0.92	1.94 / 0.90
Testing Set	1.78 / 0.84	1.67 / 0.83	1.85 / 0.84	2.10 / 0.86

Table 1: The RMSE (semitone relative to 200hz) and correlation, the prediction error of training set and testing set for four speakers.

In the tone domain, 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 simple NN with one hidden layer comprising 10 units. So there are 23 networks. They are implemented and trained using SNNS [14].

The experiments were done for all of the four speakers. The result presented in Table 1 shows that the correlations are all above 0.8 and the maximum RMSE is 2.10st. Male sets has better result than female sets.

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 3b): if the melodic contours are nicely predicted, rhythmic contours are expected to be more influenced by word/phrase chunks and syntactic hierarchy.

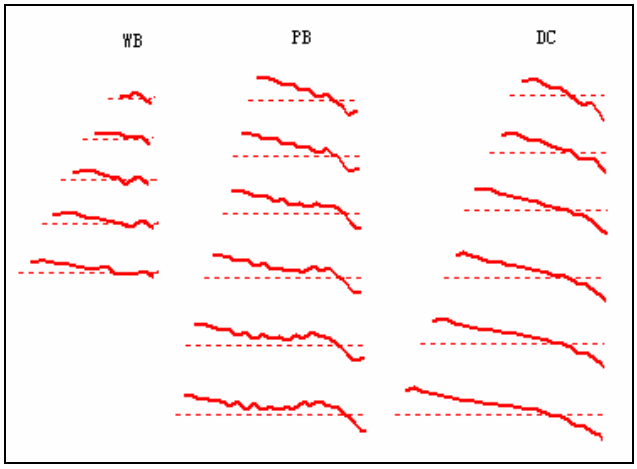

Figure 5: The melodic patterns for Word Boundary, Phrase Boundary and Declarative Clause in time domain.

4. 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. These FC patterns are consistent with our prior knowledge. Within utterances, the F0 contour tend to decline over time. Since the FC patterns only depend on the length of scope, the synthesized contours are very stable. The prosody may sound 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 functions and/or modify the characterization of the scope. 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.

5. REFERENCES

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

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

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

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

[5] Fujisaki, H. and Hirose, K. "Analysis of voice fundamental frequency contours for declarative sentences of Japanese". *Journal of the Acoustical Society of Japon*, p. 233-242.1984.

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

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

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

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

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

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

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

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

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