interspeech 2008 incorporating SST 08

Intonation Modeling of Mandarin Chinese Using a Superpositional Approach

Pablo Daniel Agüero[1], *Antonio Bonafonte*[2], *Lu Yu*[2], *Juan Carlos Tulli*[1]

[1]Communications Lab, University of Mar del Plata, Argentina
[2]TALP Research Center, Universitat Politècnica de Catalunya, Spain
pdaguero@fi.mdp.edu.ar

Abstract

The intonation model is an important component in text-to-speech systems to obtain natural and expressive speech synthesis. In this paper we propose a superpositional model for Mandarin Chinese. The intonation model is composed of the syllable and the phrase component. The parameters of the model are estimated using JEMA, a training approach with many advantages related to robustness and precision. Parameter estimation and model training are combined into a loop to progressively refine both the parameterization and the model. The high correlation (0.82) between synthetic and original contours in the test data show the suitability of this approach for modeling Mandarin. Furthermore, the high scores got in subjective evaluation (MOS=4.06) confirm the objective results.

Index Terms: speech synthesis, intonation modeling, Mandarin Chinese

10.21437/Interspeech.2008-553

1. Introduction

Speech synthesis is a field of speech processing with an important development in the past years. One of the key topics to achieve such goals is the modeling of intonation.

Intonation is the melody of the oral language. It conveys many information about the speaker and the message: genre, dialect, style, sentence modality, intention of the speaker, etc. Prosody can even complement or correct the meaning of the words. In Mandarin and other tonal languages, pitch movements have lexical meaning: different tones are used to express different words.

In Mandarin, the tones are associated to syllables. Therefore, the syllable is normally chosen as the prosodic unit to model intonation. Several intonation models are proposed in the literature using this unit and different approaches: RNN, HMM, template selection, etc. Chen et al. [1] proposed the use of a four-layer recurrent neural network (RNN) to generate syllable pitch contours, syllable energy levels, syllable initial and final durations and inter-syllable pause durations. Informal listening tests involving native Chinese speakers confirmed the naturalness of the synthesized speech. The RNN was capable to learn the well-known Sandhi Tone 3 change rule, together with many other human phonological rules. Yoshimura et al. [2] described a HMM-based speech synthesis system. Spectrum, pitch and state duration were modeled using a unified framework of HMM. Chou et al. [3] proposed a set of algorithms for the automatic labeling of the corpora. The system may produce a hierarchical prosodic structure for a desired text sentence. Such structure was used to select prosodic feature sets and appropriate waveform units. Tao et al. proposed a template selection method [4, 5]. Viterbi search was used to find out the best sequence of templates based on a definition of the costs. In this way, the resulting pitch contour is natural in the local and overall sense. This approach is similar to the searching procedure in unit selection [6, 7].

Recently, we have proposed a superpositional intonation model [8] for English, Spanish and Catalan. The f0 is modeled as the sum of phrase and accent components. Each component is modeled using a polynomial which is represented using Bézier coefficients, as proposed by Escudero [9]. Using a superpositional model both local and global aspects of pitch contours are modeled because of the different span of the units. The parameters of the model are estimated using JEMA, *Joint Extraction and Modeling Approach*. JEMA is a new training strategy which can be used for training this and other intonation models [10], as Tilt or Fujisaki. The proposed procedure has some advantages over other methods proposed in the literature: the parameters are optimal for all the training data. Furthermore, interpolation of unvoiced segments is not required. The robustness of the algorithm against missing information (unvoiced segments) and noise (e.g.: extraction errors) was shown in previous papers [10]. Finally, JEMA allows to separate the two components of the superpositional model in an optimal way, without imposing any heuristic. As JEMA analyzes all the corpus at the same time it can solve the ambiguity between accent and phrase components.

In this paper we extend our intonation model and the estimation procedure to Mandarin. We will use a superpositional model with two components: phrase and syllable. The phrase component should represent the declination (if it exists) while the syllable component models syllable tones. Section 2 describes JEMA, the training procedure. The mathematical formulation for the parameterization used in this paper is explained in section 3. The experimental setup and results are shown in Section 4. Finally, a summary and conclusions are presented in section 5.

2. Joint extraction and modeling approach

We can consider the intonation model as a function that maps linguistic and paralinguistic input features onto a fundamental frequency contour at the output:

$$G(F) = \hat{F}_0$$

where F are the set of input linguistic and paralinguistic features, $G()$ is the mapping function (intonation model), and $\hat{F}_0$ is the generated fundamental frequency contour (prediction of the model).

The resulting fundamental frequency contour may be different to the fundamental frequency contour produced by a human given the input text (reference contour). The generated contour has an error (ϵ) compared to the reference contour:

$$G(F) = F_0 + \epsilon$$

Copyright © 2008 ISCA
Accepted after peer review of full paper

The intonation model is built minimizing the error ϵ. The predicted contour must be as close as possible to the fundamental frequency contour produced by a human (reference).

This approach has an intrinsic limitation due to speaker variability and missing input features: the reference contours may be different for the same set of input features. Therefore, the intonation modeling task has an upper limitation due to these facts.

In this paper we use an intonation model training procedure that combines parameter extraction and the generation of the mapping function $G(F) = f_0$ in a loop. It has advantages over approaches proposed in the literature which perform these steps in isolation.

An optimal clustering of the feature space (see Figure 1) lets the contours of the f_0 space to be assigned to classes that may be represented by the same set of parameters, such as Fujisaki, Bézier, Tilt, etc. In our proposal, the parameters are calculated using a global optimization algorithm over all data available for training.

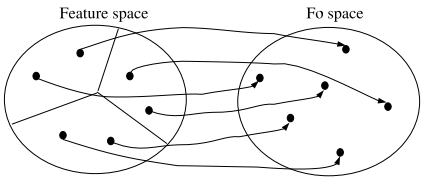

Figure 1: Classes of contours given a clustering of the feature space

One of the most important consequences of this configuration is the avoidance of the need of continuity of the fundamental frequency contour of other approaches. The missing information of some contours that belong to a given class may be compensated by the other contours of the class. The use of this complementary information is very useful to avoid the effects of missing data.

All contours that belong to a given class will share the same set of parameters which will be optimal for the class. This global optimization leads the decisions of the clustering on the feature space to find out the optimal classes of contours.

The process of intonation model training using the joint extraction and modeling approach can be summarized into these steps, using a small example of a corpus of two sentences to illustrate:

- **Initialization.** Initially only one class exists, because the tree has only the root node. In this way, all prosodic units (syllables, phrases) will be represented by the same set of parameters, as shown in Figure 2. These parameters are calculated using a global optimization algorithm over all training data. As will be described in next section, in this paper we use a superpositional model. In this case there are two clustering trees, one for syllables and one for phrases. Therefore, in the initial phase, all the syllables belong to the same syllable class and all the phrases belong to the same phrase class.
- **Splitting.** Linguistic and paralinguistic features are used to do questions in the tree to split training data. After a new question is done, the training data will have two new classes obtained from the splitting of the previous class (Figure 3). This is done twice: for the syllable tree and for the phrase tree.

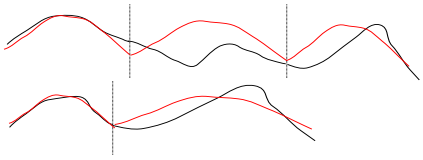

Figure 2: Approximation with class 0 contour.

Prosodic unit contour	F1	F2	F3	F4	class
	B	...	...	...	1
	M	...	...	...	2
	E	...	...	...	2
	B	...	...	...	1
	E	...	...	...	2

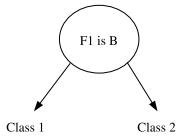

Figure 3: Approximation with classes 1 and 2.

- **Optimization.** When the new classes are obtained, a global optimization algorithm is used to find the new optimal parameters (Figure 4). Depending on the parameterization, this optimization step can be time consuming if the optimal solution has not closed-form (e.g.: Fujisaki's intonation model). In such cases hill-climbing algorithms are used to find the optimal solution.

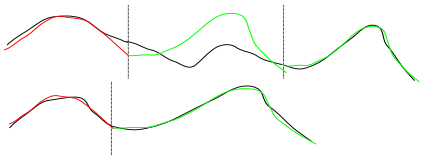

Figure 4: Approximation with two classes.

- **Scoring of the splitting.** The new parameterization is used to measure the improvement of the goodness measure compared to its value previous to the splitting.
- **Selection of the highest improvement.** After all possible splittings were tried, the splitting with the highest improvement is chosen as the best split and the tree is updated for the next iteration.
- **Stopping condition.** The decision of another iteration for an additional splitting is performed taking into account a minimum number of elements on each leaf and a minimum improvement of the goodness score.

This approach can be applied to several parametric intonation models, because it is a general technique to train intonation models, as was already shown for Bézier [8], Fujisaki [11] and Tilt [12].

In this approach the parameters are extracted using a global optimization algorithm. As a consequence, the interpolation is not needed and the parameters are more consistent and not biased by interpolation. The improvements are important in models that have the intrinsic problem that several sets of parameters optimally approximate a given contour, mainly due to the superpositional formulation. This is the case of Bézier and Fujisaki's intonation models.

The clustering in the example is done using trees. However, this approach may be applied to other clustering techniques of the feature space, as shown in Agüero et al. [13].

3. Intonation model

The parametric representation of the model is based on Bézier curves. The polynomial formulation is shown in equation 1. The shape of the base polynomials for a fourth order curve are shown in Figure 5. Bézier coefficients allow a meaningful representation compared with the final polynomial coefficients, which are more sensitive.

$$P(t) = \sum_{n=0}^{N} \alpha_n \begin{pmatrix} N \\ n \end{pmatrix} t^n (1-t)^{(N-n)} = \sum_{n=0}^{N} \alpha_n g_n(t) \tag{1}$$

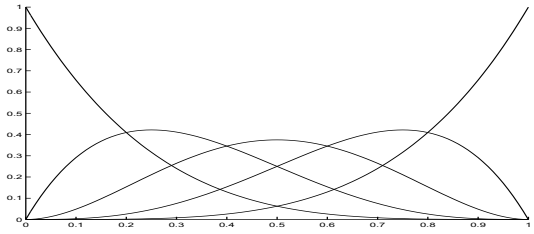

Figure 5: Bézier polynomials

The approach in this paper is superpositional: each syllable and phrase component is represented using a Bézier polynomial. The clustering procedure defined in previous section defines, for each syllable and phrase, the *cluster class* of the contour. The mathematical formulation of the model may be stated using matrix notation as:

$$\hat{f}_0 = G_p a_p + G_s a_s \tag{2}$$

where $\hat{f}_0$ is the approximation to the real contour of all the sentences in the training data. The f_0 contour is sampled every 5 msec., (only in the voiced segments).

G_s is the matrix for the intonation contours g for each order of the polynomial(0 to N) and each class (0 to M) for all concatenated contours from the first f_0 measure (0) to the last (T) in training data:

$$G_s = \begin{bmatrix} g^{0,0}(0) & g^{0,1}(0) & ... & g^{1,0}(0) & ... & g^{M,N}(0) \\ g^{0,0}(1) & g^{0,1}(1) & ... & g^{1,0}(1) & ... & g^{M,N}(1) \\ ... & ... & ... & ... & ... & ... \\ g^{0,0}(T) & g^{0,1}(T) & ... & g^{1,0}(T) & ... & g^{M,N}(T) \end{bmatrix}$$

To match the time definition of equation 1, the time instants of each syllable are scaled to fit the [0 1] range.

a_s is the vector of the Bézier coefficients (α) for each polynomial order (0 a N) and each class (0 a M). It is the solution vector:

$$a_s^T = \left[a_s^{0,0} a_s^{0,1} \dots a_s^{0,N} a_s^{1,0} \dots a_s^{1,N} \dots a_s^{M,N} \right]$$

G_p and a_p have the same role that G_s and a_s, but applied to phrase component.

The goal is to minimize the approximation error e using the real contours (f_0) in the training data as reference. f_0 is the vector with the concatenation of all intonation contours in the training data.

$$e = f_0 - \hat{f}_0 \tag{3}$$

$$e^2 = e^T e = (f_0 - \hat{f}_0)^T (f_0 - \hat{f}_0) \tag{4}$$

The minimization is performed taking the first derivative of the quadratic error function e^2 with respect to the Bézier coefficients a_p and a_s, and setting it equal to zero.

$$\frac{\partial e^2}{\partial a_p} = \frac{\partial}{\partial a_p} (f_0 - G_p a_p - G_s a_s)^T (f_0 - G_p a_p - G_s a_s) = 0 \tag{5}$$

$$\frac{\partial e^2}{\partial a_s} = \frac{\partial}{\partial a_s} (f_0 - G_p a_p - G_s a_s)^T (f_0 - G_p a_p - G_s a_s) = 0 \tag{6}$$

By using the following matrix identities

$$\frac{\partial x^T B x}{\partial x} = (B + B^t) x \tag{7}$$

$$\frac{\partial a^T x}{\partial x} = \frac{\partial x^T a}{\partial x} = a \tag{8}$$

$$(AB)^T = B^T A^T \tag{9}$$

we obtain the expression that minimizes the error e:

$$\begin{bmatrix} G_p^T f_0 \\ G_a^T f_0 \end{bmatrix} = \begin{bmatrix} G_p^T G_p & G_p^T G_a \\ G_a^T G_p & G_a^T G_a \end{bmatrix} \begin{bmatrix} a_p \\ a_a \end{bmatrix} \tag{10}$$

As already mentioned, this expression is used in each iteration and each possible question to obtain the best tree.

4. Experiments

The experiments in this paper are performed using the Mandarin Chinese database provided by the The National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing. It consists in 6.5 hours of utterances in Mandarin Chinese (Beijing dialect) of a female speaker. The utterances are transcribed using Chinese Traditional Characters and Pinyin. Part-of-Speech tags are provided as well as complementary information to help in the task of prosody modeling in the Blizzard Challenge 2008 [14]. The goal of the experiments is to study the naturalness of the intonation compared with natural speech. The database is splitted in two sets, training (70%) and test (30%).

The intonation model was trained using a set of features extracted from the transcription of the utterances. Due to the superpositional approach, two prosodic units are used: syllable and phrase.

The syllable intonation contours are predicted based on: position of the syllable relative to the word and phrase, syllable tone, preceding and succeeding tones (in order to deduce rules such as tone Sandhi), prepausal and the pinyin transcription of the syllable. The phrase component is predicted using the following features: number of words and syllables in the phrase, punctuation, POS preceding and succeeding the break, and POS sequence in the phrase.

Two objective measures are used to compare real and predicted contours: root mean squared error (RMSE) and correlation coefficient.

The RMSE for training and test data was $37.4Hz$ and $37.5Hz$, and the correlation was 0.824 and 0.833, respectively. Such high values for correlation show that the intonation model achieves a high resemblance in trajectory with a natural pitch contour.

The difference between natural (reference) and predicted pitch contours shown by RMSE are due to two facts. On the one hand, speakers with higher pitch range will show higher RMSE in the linear scale of frequency, as shown in many papers in the literature. On the other hand, although the high correlation shows an appropriate trajectory in pitch contour, the predicted contours may not have the full pitch range of the original speaker due to the inherent smoothing of this kind of clustering approaches. The representative pitch contour of the cluster may not have enough excursion to match the natural contour, as shown in Figure 6.

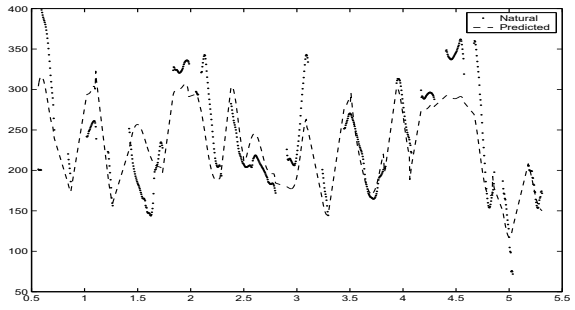

Figure 6: Difference in excursion between natural and predicted pitch contours.

The analysis of relevance of the features show that the most important component is the syllable tone. The most relevant features are syllable tone, and preceding and succeeding tones. The relevant features for phrase component are the number of syllables and words in the phrase.

A subjective experiment was also performed to study the naturalness of the intonation from the point of view of perception. The listening test was done by 88 native mandarin Chinese speakers. They were asked to score natural and synthetic utterances in a five-point scale. The synthetic audios were obtained using resynthesis with Praat [15]. In order to control the distortion introduced by resynthesis, the listeners also evaluated the quality of the synthetic (or natural) voice.

The naturalness of original and synthetic utterances was 4.72 and 4.06 respectively. It can be seen that the score is very high. However, there is a degradation with respect to the original contours. We think that the main reason is the smaller f_0 range of the synthetic contours. On the other hand, the quality of the voice is 4.85 and 4.33 for natural and synthetic sentences. These figures show that a very small distortion is introduced by the PSOLA manipulation.

5. Conclusions

In this paper we have proposed a superpositional approach for modeling Mandarin intonation. The model is the sum of two components, one for syllables and one for phrases.

The model is estimated using JEMA. JEMA has already been applied to English, Catalan and Spanish and several intonation models (Fujisaki, Bézier, Tilt) with very good results. JEMA has some advantages over other methods proposed in the literature: global optimal parameters and no need of interpolation of unvoiced segments. As a consequence, the algorithm is robust against missing information (unvoiced segments) and noise (e.g.: extraction errors). The advantages of the proposal are manifested in objective and subjective tests. High correlation (0.82) and MOS (4.06) for testing data support our approach. What's more, many listeners indicated that it was almost impossible to distinguish the natural and synthetic contours.

This model has been included in Ogmios, the UPC text-to-speech system, in Blizzard Challenge 2008 [14].

6. References

[1] Chen, S., Hwang, S. and Wang, Y., "An RNN-Based Prosodic Information Synthesizer for Mandarin Text-to-Speech", IEEE Transactions on Speech and Audio Processing, 6(3), 1998.

[2] Yoshimura, T., Yokuda, K., Masuko, T., Kobayashi, T. and Kitemura, T., "Simultaneous Modeling of Spectrum, Pitch and Duration in HMM-Based Speech Synthesis", Proceedings of Eurospeech, Budapest, Hungary, 1999.

[3] Chou, F., Tseng, C. and Lee L., "A Set of Corpus-Based Text-to-Speech Synthesis Technologies for Mandarin Chinese", IEEE Transactions on Speech and Audio Processing, 10(7), 2002.

[4] Tao, J., "F0 Prediction Model of Speech Synthesis Based on Template and Statistical Method", Lecture Notes of Artificial Intelligence, Springer, 2004.

[5] Tao, J., Yu, J. and Zhang, W., "Internal Dependence Based F0 model for Mandarin TTS system", TC-STAR Workshop on Speech-to-Speech Translation, 2006.

[6] Wang, R., Ma, Z., Li, W. and Zhu, D., "F0 Prediction Model of Speech Synthesis Based on Template and Statistical Method", Proceedings of ICSLP, Beijing, China, 2000.

[7] Chu, M., Peng, H. and Chang, E., "A Concatenative Mandarin TTS System without Prosody Model and Prosody Modification", Proceedings of the 4th ISCA Workshop on Speech Synthesis, 2001.

[8] Agüero, P. D. and Bonafonte, A., "Intonation Modeling for TTS Using a Joint Extraction and Prediction Approach,", Proceedings of the International Workshop on Speech Synthesis, Pittsburgh, USA, 67-72, 2004.

[9] Escudero, D., Cardeñoso, V. and Bonafonte, A., "Corpus based extraction of Quantitative Prosodic Parameters of Stress Groups in Spanish", Proc. of ICASSP'2002, Orlando, USA, May 2002.

[10] Agüero, P. D., Tulli, J. C., and Bonafonte, A., "A Study of JEMA for Intonation Modeling", Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, USA, 2008.

[11] Agüero, P. D., Wimmer, K. and Bonafonte, A., "Joint Extraction and Prediction of Fujisaki's Intonation Model Parameters,", Proceedings of International Conference on Spoken Language Processing, Jeju Island, South Korea, 757-760, 2004.

[12] Rojc, M., Agüero, P. D., Bonafonte, A. and Kacic, Z., "Training the Tilt Intonation Model Using the JEMA Methodology,", Proceedings of Eurospeech 2005, Lisboa, Portugal, 3273-3276, 2005.

[13] Agüero, P. D. and Bonafonte, A., "Facing Data Scarcity Using Variable Feature Vector Dimension,", Proceedings of the International Conference on Speech Prosody, Dresden, Germany, 1-4, 2006.

[14] Black, A. W., King, S., and Tokuda, K., "The Blizzard Challenge 2008 - Evaluating Corpus-based Speech Synthesis on Common Databases", http://www.synsig.org/index.php/Blizzard_Challenge_2008.

[15] Boersma, P. and Weenink, D., "Praat: Doing Phonetics by Computer", http://www.fon.hum.uva.nl/praat/.