Novel Eigenpitch-based Prosody Model for Text-to-Speech Synthesis

Jilei Tian [1], *Jani Nurminen* [2] *and Imre Kiss* [1]

[1] Interaction Core Technology Center, Nokia Research Center, Finland
[2] Technology Platforms, Nokia, Finland
{Jilei.tian, jani.k.nurminen, imre.kiss}@nokia.com

10.21437/Interspeech.2007-229

Abstract

Prosody is an inherent supra-segmental feature in speech that human speakers employ to express, for example, attitude, emotion, intent and attention. In text-to-speech (TTS) systems, high naturalness can only be achieved if the prosody of the output is appropriate. The importance of prosody is even more crucial for tonal languages, such as Mandarin Chinese, in which the tone of each syllable is described by its pitch contour. In this paper, we propose a novel prosody modeling approach that uses the concept of syllable-based eigenpitch. The approach has been implemented in our Mandarin TTS system resulting in less than 0.1% error variance. The results obtained in practical experiments have confirmed the good performance of the proposed technique.

Index Terms: prosodic modeling, pitch, eigenpitch, text-to-speech

1. Introduction

The term prosody refers to certain properties of a speech signal that are related to audible changes in pitch, loudness and duration. Physically, the pitch of an utterance depends on the rate of vibration of the vocal cords. The higher the rates of vibration, the higher the resulting pitch frequency [1]. Another concept closely related to pitch is tone that is used to describe pitch variations inside short stretches of syllables. In tonal languages, these relative pitch differences are used either to differentiate between word meanings or to convey grammatical distinctions.

Many of the languages of South-East Asia and Africa are tonal languages. Mandarin Chinese is probably the most widely studied tonal language in which each stressed syllable has a significant contrastive pitch that is an integral part of the syllable. It has one neutral tone and four basic tones: high level, high rising, dipping/falling and high falling. Since the pitch contour conveys information about word meaning distinction, prosodic phrase and word boundaries, pitch information plays a crucial role in speech synthesis and speech recognition systems, especially for tonal languages [2][3]. Due to all of these reasons, pitch modeling is one of the key issues that must be addressed when dealing with tonal languages.

The most popular pitch modeling approaches are mainly using the concept of separating the pitch contour into a global trend and local variation. Two examples following this approach are the superpositional modeling technique [4] and the two-stage modeling technique. In [5], the mean and the shape of the syllable pitch contours are taken as two basic modeling units by using a 3rd order orthogonal polynomial expansion. Since the syllable pitch contour patterns vary dramatically from their canonical form, a reasonable assumption is that some data-driven approach could preserve more precise and more relevant information compared to pure artificial fitting.

In this paper, we propose a novel data-driven pitch modeling approach based on the new concept of eigenpitch. The proposed model has been implemented in our unit selection based Mandarin TTS system. In our implementation, the eigenpitch representation is employed not only in prosody prediction but also in unit selection. We have verified that the eigenpitch based approach can offer both a meaningful and compact representation and a good performance from the viewpoint of quality.

The rest of the paper is organized as follows. The concept of eigenpitch is first described in Section 2. In Section 3, we describe the proposed prosodic modeling approach that is based on the usage of eigenpitch in prosodic templates and in unit selection. The performance of the proposed approach is discussed in Section 4. Finally, some concluding remarks are presented in Section 5.

2. Eigenpitch

The prosody model proposed in this paper is based on the concept of eigenpitch that we have recently introduced and tentatively studied in [6]. Our analysis results showed that the eigenpitch representation offers many beneficial properties. A very important benefit is that the pitch vector dimension can be reduced in the eigen space while minimizing the energy loss and preserving the tonal features and high classification capability.

The concept of eigenpitch is derived through the use of Principal Component Analysis (PCA) [7] technique. PCA is a multivariate procedure that computes a compact description of the data set by rotating the data in such a way that the maximum variabilities are projected onto the axes. Essentially, a set of correlated variables is transformed into a set of uncorrelated variables that are sorted in the order of reducing variability. The main use of PCA is to reduce

the dimension of a data set while retaining as much information as possible, and also to extract new uncorrelated features from the original data.

In the case of eigenpitch, the M input data vectors are first represented as pitch contour vectors of dimension N, denoted as a $\mathbf{x}_i$. Then, the sample mean is calculated for each element, resulting in an N-dimensional vector $\mathbf{m}$. The sample covariance matrix $\mathbf{R}_{NxN}$ can then be computed by

$$\mathbf{R}_{NxN} = \frac{1}{M} \cdot \sum_{i=1}^{M} (\mathbf{x_i} - \mathbf{m})(\mathbf{x_i} - \mathbf{m})^T \quad (1)$$

The eigen analysis on the covariance matrix $\mathbf{R}_{NxN}$ yields a set of positive eigenvalues $\{\lambda_1, \lambda_2 ..., \lambda_N\}$ in descending order. Their corresponding eigenpitch vectors, $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N\}$, are the principal components. A smaller eigenvalue contributes much less weight to the total variance, hence if the pitch contours are projected onto a subset of principal components, the omission of later components tends to introduce less error than if earlier components are omitted.

3. Eigenpitch Based Prosody

The proposed prosody modeling approach in our Mandarin TTS is based on two main principles: the eigenpitch representation and the use of prosodic templates. Linguistic context information is used in the template selection, whereas the resulting eigenpitch is used for guiding the unit selection that operates on syllable-sized units. These steps are described in more detail in the following subsections.

3.1. Prosodic templates

Chinese syllabic prosodic features have shown close relationship with their linguistic context information [2][3]. Thus, it is reasonable to assume that the prosodic behavior for a syllable can be predicted based on the linguistic context. Our model performs this prediction using syllable-based templates. The syllable-based prosodic templates are built off-line by first collecting all the possible templates from a large training corpus and by pruning them using a statistical algorithm that removes entries that are too similar to some other entry, while still achieving a sufficient coverage on the tonal syllables and the contexts. Each entry in the prosodic template inventory contains information on the linguistic context in which the syllable occurred and information on the syllable itself. The entries also contain data related to the acoustic realizations of the syllables, including the eigenpitch vector representing the pitch contour, the duration and the energy of that particular instance of the syllable.

When using the model, the TTS system must first derive the context parameters for each syllable in the input text using different text processing techniques. Then, a cost function is used to measure the distance between the context parameters extracted from the text and the context parameters in the syllable template data stored in the prosody model. The prosodic template offering the best matching context is selected from the template inventory as described in next subsection. The prosodic data corresponding to the retrieved template is used as the predicted prosody.

More specifically in our Mandarin TTS system, the context features, as categorized below, are extracted to form the context feature vector **c** for an input text:

- Linguistic information for the syllable
  Phonetic entity, syllabic tone, location in prosodic word, preceding and following boundary type, stress, etc;
- Preceding linguistic context information
  1 to 3 preceding phonetic entities of focused syllable and their syllabic tones;
- Following linguistic context information
  1 to 3 following phonetic entities of focused syllable and their syllabic tones;
- Linguistic information in word-level
  POS, number of syllables, location in prosodic phrase;
- Linguistic information in phrase-level
  Number of words, location in sentence;
- Linguistic information in sentence-level
  Number of phrases, sentence type;

Each entry in the prosodic template inventory includes these context features, together with eigenpitch, duration and energy.

3.2. Prosodic template selection

The output of the prosodic model is given by the best prosodic template that optimizes a prosody cost. The prosody cost (*PC*) is defined as combination of the prosody target cost (*PTC*) and prosody concatenation cost (*PCC*). The prosody target cost gives the distance between the current syllable and an entry from the template, and is based on context features described in Section 3.1. The prosodic concatenation cost estimates the prosodic distance between the adjacent units to be concatenated. *PCC* is pitch-based.

The prosodic template selection occurs in three passes: (1) pre-selection of N-best candidates so that *PTCs* are minimized; (2) computation of *PCCs* for the different combinations of candidates; (3) dynamic programming (Viterbi search) to find the best template sequence by optimizing the total *PC*.

PTC is defined as the weighted distance between the context features extracted from the input text and the context features stored in the prosodic template model.

$$PTC_{ij} = \sum_{k=1}^{K} w_k \cdot d(c_{ki}, \hat{c}_{kij}) \quad (2)$$

where d stands for contextual distance, computed here between k-th features of i-th syllable between input context and j-th template in prosodic model. w_k denotes the weight of the k-th contextual features. The N-best template candidates are obtained for each input syllable at the first pass. The same step is repeated for

next syllable until the whole sentence is completely processed.

After computing *PCCs* of joint pitch difference between the different candidates, a Viterbi *N*-best search is carried out on the prosodic template lattice to find the best template path or sequence. The lattice consists of a template net with *PTC* stored at each template candidate (node), and *PCC* defined for each adjacent template pair (arc). A scaling factor λ is used to balance between *PTC* and *PCC*. Given an input sentence consisting of *M* syllables, each syllable has *N*-best prosodic template candidates. Thus, we have the template lattice as shown in Figure 1.

Sentence: syllable_1 …… syllable_m …… syllable_M

Template: p_{11} …… p_{1m} …… p_{1M}

…… …… …… …… ……

p_{N1} …… p_{Nm} …… p_{NM}

Figure 1. Prosodic template lattice.

Let n=path(m) denote the n-th template candidate corresponding to the m-th syllable in the input sentence, and the path indicating a template sequence extracted from template lattice. $PTC_{\text{path}(m),m}$ denotes the target cost for n=path(m)-th template candidate p_{nm} of m-th syllable in the input sentence, and $PCC_{\text{path}(m),\text{path}(m+1)}$ stands for the concatenation cost between adjacent template candidates at m-th and $m+1$-th syllables. Then the best prosodic template path path* is obtained:

$$\begin{aligned} path^* &= \underset{path}{\operatorname{argmin}} \sum_{m=1}^{M} \left(PTC_{path(m),m} + \lambda \cdot PCC_{path(m),path(m+1)} \right) \\ &= \underset{path}{\operatorname{argmin}} \left\{ \sum_{m=1}^{M} \left(PTC_{path(m),m} \right) + \lambda \cdot \sum_{m=1}^{M} \left(PCC_{path(m),path(m+1)} \right) \right\} \end{aligned} \quad (3)$$

The predicted prosodic features of the syllables are determined from the selected templates of the optimal Viterbi template path.

3.3. Acoustic unit selection

The role of the acoustic synthesis is to select the most suitable acoustic units from the large pre-recorded speech corpus and concatenate these units to generate the natural-sounding synthesized speech. Similarly as in our prosodic template selection described in Section 3.2, the acoustic cost is generally composed of the target cost and the concatenation cost, and used for acoustic unit selection. The target cost is the distance measure between the target features and the features ($\mathbf{p}_k$,$\mathbf{c}_k$) of the k-th acoustic unit. The target features include the contextual features **c*** extracted in the text processing and the prosodic features **p*** predicted by the prosodic model as described in Section 3.2. By minimizing the target cost, the best unit k* can be selected as shown in Equation 4.

$$k^* = \arg\min_k \left\{ d_p(\mathbf{p}^*, \mathbf{p}_k) + d_c(\mathbf{c}^*, \mathbf{c}_k) \right\} \quad (4)$$

As can be seen, the target cost is based on the two distance measures. The prosodic distance d_p between the selected template and an acoustic unit is calculated using their durations and the eigenpitch vectors stored. The context distance d_c is also calculated between the context features extracted from the input text and the context features assigned to pre-recorded acoustic units. The target cost is obtained by adding the two parts together. The acoustic unit can then be selected by minimizing the target cost. Intuitively, the concatenation cost could also be defined as the distance between the boundaries of two syllables aiming to minimize the discontinuity between two adjacent acoustic units. In the Mandarin language, however, this is not so critical because there are also natural discontinuities between syllables. Furthermore, in our system the acoustic units are coded in a parametric domain, giving the possibility to efficiently perform smoothing at the concatenation boundaries. Due to these reasons, we do not apply concatenation cost for the acoustic unit selection.

4. Experimental Results

To demonstrate the good performance and properties of the eigenpitch based approach, we carried out practical experiments using our Mandarin TTS system and an internal Mandarin Chinese speech database, consisting of 84,393 syllables from a single female speaker.

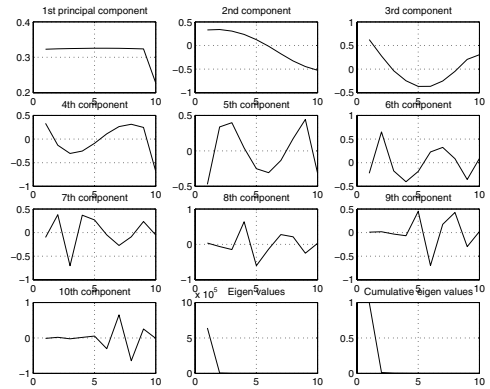

Figure 2. Eigenpitch and eigenvalues from the Mandarin Chinese database.

For each of the syllables, the pitch was first automatically extracted and then manually validated. Furthermore, each variable-length syllable-based pitch contour was converted into a 10-valued pitch contour vector. Figure 2 shows the eigenpitch decomposed from the pitch contour vectors in descending order.

In comparison with the tonal character, the first eigenvector describes the pitch level, one of the key features in the tones, and remarkably matching tone 1 in the shape. The rest of the eigenvectors are used to

model the pitch variation. The second eigenvector is obviously in line with tones 2 and 4, depending on the positive or negative sign. The third (and partially fourth) eigenvectors are the key elements to model tone 3.

If only the L most significant eigenpitch elements are selected for the projection of the pitch contour, the Euclidean distance error in the original time domain between the original and the truncated eigenpitch is limited by the following error variance approximation

$$Var\left\{d(\mathbf{p}^*,\mathbf{p}_k)-d(T(\mathbf{p}^*),T(\mathbf{p}_k))\right\}\le(1-\frac{\sum_{i=1}^{L}\lambda_i}{\sum_{i=1}^{N}\lambda_i}) \quad (5)$$

where $\mathbf{p}^*$ and $\mathbf{p}_k$ denote pitch contours of predicted template and k-th acoustic instance in the unit inventory. *T(.)* denotes the transformation into eigenspace. When the four most significant principal components are used in our system, the error variance approximation gives

$$Var\left\{d(\mathbf{p}^*,\mathbf{p}_k)-d(T(\mathbf{p}^*),T(\mathbf{p}_k))\right\}\le 0.1\%$$

The remaining eigenvectors have the following properties.

1. They contain only a small contribution of energy or variance to the pitch contour.
2. They have more errors due to imperfect pitch extraction. The errors can originate either from the automatic extraction or from the manual validation.
3. They are the least important features for tonal classification. We have experimentally verified this in [6].

Based on the above analysis, we have used 4-dimensional eigenpitch vectors in our prosodic model.

We performed a preliminary internal subjective speech quality evaluation consisting of 11 TTS sentences and 6 testers. In the mean opinion score (MOS) test, a 5-point scale was defined as the MOS value ranged from bad (1) to excellent level (5). The results shown in Table 1 indicate that the speech quality obtained using the proposed model is similar as the quality obtained using a conventional template based prosody model with full pitch contours. The truncation of the eigenpitch representation into 4-dimensional vectors has not only helped in reducing significantly the memory footprint and the computational complexity of the prosody modeling, but also kept the speech quality almost unchanged.

Table 1. *MOS test on pitch representation using original pitch contour and truncated eigenpitch.*

	Pitch Contour	Eigenpitch
MOS	3.57	3.58

5. Discussion and Conclusions

In this paper, we have introduced an eigenpitch based approach for prosody generation in TTS synthesis. The proposed method transforms the pitch contours into a lower-dimensional representation in the eigen space by using the well-known PCA technique. It has superb advantages in terms of compact representation, tonal discriminative capability and Euclidean distance preserving. Since the pitch information is presented in both the prosodic and acoustic data and used for unit selection that is a crucial step for reaching high TTS quality, the distance preserving property of the eigenpitch is of particular importance.

Our experiments have shown that the proposed technique fulfils the requirements for resource-constraint high-quality TTS synthesis. Moreover, we have also verified that the truncation in the eigenpitch domain does not cause distortion in the output quality but it offers significant savings from the viewpoint of memory footprint and computational complexity. It might be possible to also improve the speech quality using the proposed model due to the enhanced robustness.

6. Acknowledgement

This work has partially been funded by the European Union under the integrated project TC-STAR - Technology and Corpora for Speech-to-Speech Translation - (IST-2002-FP6-506738, http://www.tc-star.org).

7. References

[1] Dutoit, T., An Introduction to Text-to-Speech Synthesis, Kluwer Academic Publishers, Dordrecht, 1997.

[2] Li, A., "Chinese prosody and prosodic labeling of spontaneous speech", In Proceedings of International Workshop on Speech Prosody, Aix-en-Provence, France, 2002.

[3] Yu, J., Zhang, W. and Tao, J., "A new pitch generation model based on internal dependence of pitch contour for Mandarin TTS system," In Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toulouse, France, 2006.

[4] Bellegarda, J., Silverman, K., Lenzo, K. and Anderson, V., "Statistical prosodic modeling: from corpus design to parameter estimation", IEEE Trans. Speech and Audio Processing, Vol. 9, No.1, pp. 52-66, 2001.

[5] Lai, W., Wang, Y. and Chen, S., "A new pitch modeling approach for Mandarin speech", In Proceedings of 8th European Conference on Speech Communication and Technology (EUROSPEECH), Geneva, Switzerland, 2003.

[6] Tian, J. and Nurminen, J., "On analysis of eigenpitch in Mandarin Chinese," In Proceedings of 4th International Symposium on Chinese Spoken Language Processing, HongKong, China, 2004.

[7] Fukunaga, K., Introduction to statistical pattern recognition, Academic Press, Dordrecht, 2000.