On Analysis of Eigenpitch in Mandarin Chinese

Jilei Tian and Jani Nurminen

Multimedia Technologies Laboratory
Nokia Research Center, Tampere
{jilei.tian, jani.k.nurminen}@nokia.com

Abstract

Prosody is an inherent supra-segmental feature of human's speech that is being employed to express e.g. attitude, emotion, intent and attention. Pitch is the most important feature among the prosodic information. For Mandarin Chinese speech, the pitch information is even more crucial because Mandarin is a tonal language in which the tone of each syllable is described by its pitch contour. In this paper, the concept of syllable-based eigenpitch is introduced and investigated using principal component analysis (PCA). The eigenpitch and the related eigen features are analyzed, and it is shown that the tonal patterns are preserved in the eigenpitch representation. Furthermore, we show that the dimension of pitch in the eigen space can be reduced while minimizing the energy loss of the original pitch contour. Finally, we briefly discuss the quantization properties of the eigenpitch representation. We also present experimental results obtained using a Mandarin speech database. They are in line with the theoretical reasoning and further prove the usefulness of the proposed pitch modeling technique.

1 Introduction

The term *prosody* refers to certain properties of a speech signal that are related to audible changes in pitch, loudness and duration. Among these features, pitch usually plays the most important role. Physically, the pitch of an utterance depends on the rate of vibration of the vocal cords; the higher the rate of vibration, the higher the resulting pitch becomes. 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 four basic tones: high level, high rising, dipping/falling and high falling. They are used to distinguish otherwise homophonous words as shown in Table 1.

Word	Intonation	Meaning
ma	[--]	mother
ma	[/]	numbness
ma	[\/]	horse
ma	[\]	curse

Table 1. Examples of different tones in Mandarin Chinese.

The most commonly used representation of tonal pitch contours as numbers is shown in Table 2. It consists of five pitch levels, rather like the use of staves in music scores. They are labeled from the bottom upwards from 1 to 5. The tonal patterns are captured using the reference pitch numbers by observing the start, the middle and the end points of the pitch contour [7].

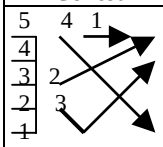
Contour	Type	Pattern	Feature
	Tone 1	5-5	H-H (High)
	Tone 2	3-5	L-H (Rising)
	Tone 3	2-1-4	L-L (Low)
	Tone 4	5-1	H-L (Falling)

Table 2. Tonal patterns and phonological notations of four citation tones in Mandarin Chinese.

Obviously pitch information plays a crucial role in speech synthesis systems, especially for tonal languages [3][8]. Since the pitch contour conveys information about word meaning distinction, prosodic phrase and word boundaries, it has been found in [5] that human beings use the pitch contour information to enhance the speech recognition performance. Various techniques have also been proposed to improve the noise robustness of speech recognition systems by using the pitch information [5]. 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 [2] and the two-stage modeling technique [1]. In [6], 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 data-driven pitch modeling approach based on the concept of eigenpitch and study its properties to verify the above assumption. In addition, we provide results related to tonal classification and pitch compression using the proposed modeling approach.

The remainder of the paper is organized as follows. We first describe the process of eigenpitch extraction and some of the basic properties of the eigenpitch representation. Then, the performance of the proposed modeling approach in the tonal

classification task is discussed in Section 3. In Section 4, we briefly study the quantization properties of the pitch features in the eigen space. Finally, conclusions are drawn in Section 5.

2 Concept of eigenpitch

2.1 Definition

The concept of eigenpitch is derived through the use of the Principal Component Analysis (PCA) [4] technique. PCA is a multivariate procedure that computes a compact and in a way optimal 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 ordered by reducing variability. This process can be viewed as a rotation of the existing axes to new positions in the space defined by the original variables. In this new rotation, the new uncorrelated variables are linear combinations of the original variables. The first new variable is the combination of variables that explains the greatest amount of variation; the second new variable contains the maximum amount of variation unexplained by the first and orthogonal to the first, etc. Thus the last of these variables can be removed with the minimum loss of real data. The main use of PCA is to reduce the dimensionality of a data set while retaining as much information as possible, and also to extract new uncorrelated features from the original data.

Mathematically, the principal component analysis involves an eigen analysis on a covariance matrix. In the case of eigenpitch, the M input data vectors are 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 eigenvectors, $\{\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_N}\}$, are the principal components. The first principal component retains the most variance. The second component retains the next highest residual variance, and so on. A smaller eigenvalue contributes much less weight to the total variance, hence if the feature vectors are projected onto a subset of principal components, omission of later components tends to introduce less classification error than if earlier components are omitted. In many cases, the first few components can retain nearly all of the variance, enabling satisfactory classification. If the d most significant principal components are selected for projection of the data, then the variance retained by this approximation is

$$Var_d = \frac{\sum_{i=1}^{d} \lambda_i}{\sum_{i=1}^{N} \lambda_i} . \quad (2)$$

2.2 Basic properties

To demonstrate the properties of the eigenpitch, we carried out practical experiments using an internal Mandarin Chinese speech database, consisting of 84,393 syllables from a single female speaker. For each of the syllables, the pitch is first automatically extracted and then manually validated. Furthermore, each variable-length syllable-based pitch contour is converted into a 10-valued pitch contour vector (N = 10 in the analysis). Figure 1 shows the eigenpitch decomposed from the pitch contour vectors in descending order. In comparison with the tones defined in Table 1, 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. The variance retained by only using the four most significant principal components is 99.9%. 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. The next section will experimentally prove this.

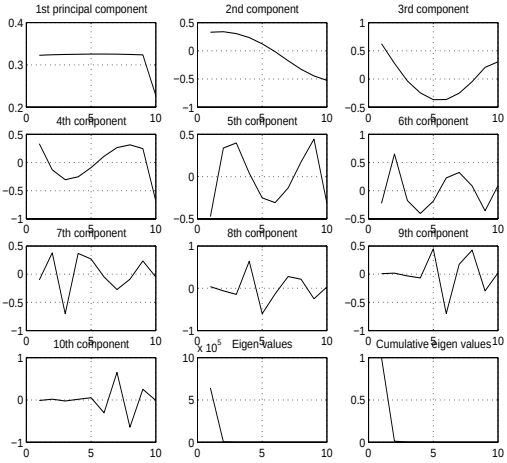


Figure 1. Eigenpitch and eigenvalue of Mandarin Chinese.

3 Tonal classification properties

3.1 Introduction

To demonstrate the classification properties of the eigenpitch representation, we performed a comparison between the proposed approach and an approach based on the linear discriminant analysis (LDA) approach. The objective of LDA is to perform dimensionality reduction while preserving as much of the class discriminatory information as possible. Thus, it is theoretically clear that LDA outperforms PCA in terms of classification capability in cases where the discriminatory information is not aligned with the direction of the maximum variance. However, we will experimentally show later in next subsection that the difference in performance on tonal classification task is marginal. In

addition, the proposed approach has the following clear advantages over LDA:

1. Eigenpitch is linguistically meaningful and in line with the Mandarin tone patterns. The basis patterns of LDA are aiming to maximize the class discriminatory information and thus they are very different than tone patterns.
2. PCA is computed in an unsupervised manner. Therefore, the eigenpitch can be computed without tonal labeling on the database, which in turn makes the process very simple and robust. The LDA approach, on the other hand, requires the tonal labels. Due to the Sandhi effect, the lexical tones are sometimes modified in the realized speech for Mandarin, and thus the labeling task is nontrivial and can lead to errors.
3. In speech synthesis, not only tonal discriminatory information but also the energy of pitch should be maximally preserved because the variation of pitch for the same tone is also very important for speech synthesis. This is particularly true for the unit selection procedure that usually has to measure the distortion or the distance between two pitch contours. In the previous tests discussed in Section 2, it is shown that 99.9% of the energy is preserved by using only the four most significant principal components. The same accuracy cannot be achieved using LDA.

3.2 Discriminative measure

The ability of a feature to distinguish between two classes depends on both the distance between the two classes and the amount of scatter within the classes. A reasonable measure of class discrimination must take into account both the mean and variance of the classes. One such measure of separability between two classes is Fisher's discriminant. The idea is that the overall class separation is increased when the class means are further apart or when the spread of the classes is smaller. For the tasks that have more than just two classes, *F*-ratio provides a measure of separability among multiple classes.

$$F-ratio = \frac{Variance\ of\ means\ (between-class)}{mean\ of\ \text{var}iances\ (within-class)}. \quad (3)$$

The *F*-ratio measures the separability of a single dimension of the feature vector. To evaluate the discrimination of an entire feature set, a multivariate extension of *F*-ratio is the *J*-measure.

$$J = tr(\mathbf{W}^{-1} \cdot \mathbf{B}) \quad (4)$$

The operator *tr(.)* is used to indicate the trace of a matrix. The matrix **B** is the between-class covariance or the covariance of the class means whereas the matrix **W** is the within-class covariance or the average of the class covariances. **B** and **W** measure how close the classes are from each other and how large the classes are. These matrices can be calculated using the formulas.

$$\mathbf{B} = \frac{1}{M}\sum_{k=1}^{C} M_k (\mathbf{m}^{\mathbf{k}\prime} - \mathbf{m})(\mathbf{m}^{\mathbf{k}\prime} - \mathbf{m})^T, \quad (5)$$

$$\mathbf{W} = \frac{1}{M}\sum_{k=1}^{C}\sum_{i=1}^{M_k} (\mathbf{x}_\mathbf{i}^\mathbf{k} - \mathbf{m}^\mathbf{k})(\mathbf{x}_\mathbf{i}^\mathbf{k} - \mathbf{m}^\mathbf{k})^T, \quad (6)$$

where C is the number of classes, M_k is the number of feature vectors in the *kth* class, M is total number of feature vectors, $\mathbf{x}_\mathbf{i}^\mathbf{k}$ is the *ith* feature vector in the *kth* class, and $\mathbf{m}^\mathbf{k}$ is the mean of class *k*.

3.3 Experimental results

First of all, the normalized F-ratio measuring the separability of each feature in the feature vector is calculated and the results are shown in Figure 2. It can be seen that all the features in the original pitch space have almost equal contribution for the tonal classification task. However, the first four features in the eigenpitch space are the most important for the tonal classification. As mentioned in Section 2, the most tonal features in the eigen space are the projected values on the first four eigenpitches.

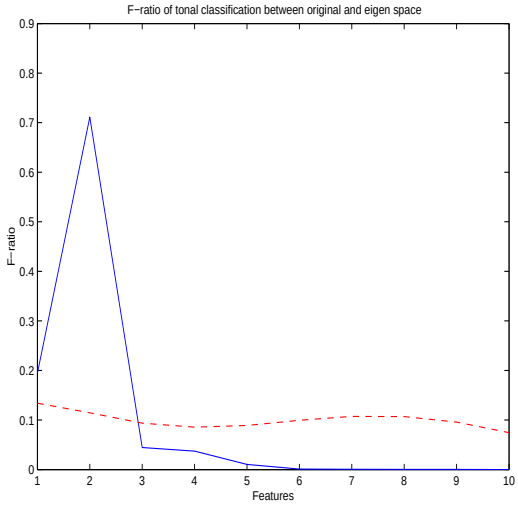


Figure 2. *F*-ratio values of tonal classification between the original space (dotted) and the eigen space (solid).

Next, the *J*-measure is used for evaluating the separability of the entire feature set for the original pitch, the eigenpitch and the LDA transformed space. Table 3 shows that the four principal eigenpitch components preserved the most tonal discriminatory information. The most discriminative LDA transformed features perform slightly better in the classification task but the difference in discrimination capability is marginal. The relative difference of the *J*-measure is 0.28% between PCA and LDA.

It is also found that the eigen feature for each tonal class has approximately Gaussian distribution. This is a very useful property when designing the classifier or modeling the features.

	10 features	First 4 features
Original Pitch	3.1992	2.0160
Eigenpitch (PCA)	3.1992	3.1901
LDA	3.1992	3.1992

Table 3. Separability measure (*J*-measure) for the whole feature sets and for the first 4 features in the sets.

4 Quantization properties

4.1 Motivation

Pitch and tone aspects are becoming increasingly important in speech recognition, synthesis, understanding and dialogue

systems. The memory and computational resources on embedded portable devices are inherently scarce. Though these resources are expected to increase in the future portable devices, the number of applications running simultaneously is also very likely to increase. The memory needed for the pitch contour information is usually rather high, and thus an efficient coding or compression method is needed. Here, we briefly discuss some quantization properties of the eigenpitch representation. We show that in order to minimize the size of the pitch contour while keeping the distortion minimized, an unequal bit allocation is needed for scalar quantization. More thorough investigation on the quantization properties will be done as a part of our future research on this topic.

4.2 Optimal bit allocation

For the original pitch, the variance of each feature is almost equally same. Thus, the same number of bits is assigned to all features. The features in the eigenpitch space have, however, very different variances. Since the eigenvalues are actually the same as the variances of the eigen features, the middle figure at the bottom of Figure 1 shows the variances. A reasonable assumption based on this figure is that unequal bit allocation may be needed to optimally code the pitch. To analyze the effects of quantization, let the *ith* feature be encoded with a fixed number of bits b_i. Now, by assuming that a uniform scalar quantizer is used for the quantization, it is possible to represent the uniformly distributed quantization error e_i as

$$-\frac{\Delta_i}{2} \le e_i \le \frac{\Delta_i}{2}, \quad \Delta_i = \frac{2 \cdot x_{i\max}}{2^{b_i}} \propto \frac{\sigma_i}{2^{b_i}}, \quad i = 1,...,N \tag{7}$$

Therefore, the squared quantization error is

$$\begin{aligned} & e^2 = e_1^2 + e_2^2 + ... + e_N^2 \\ & 0 \le e_i^2 \le \frac{\Delta_i^2}{4} \propto \frac{\sigma_i^2}{2^{2b_i}}, \quad i = 1,...,N \end{aligned} \tag{8}$$

Given a constant number of bits B for the whole feature set, the optimal bit allocation can be estimated by minimizing the following Lagrangian cost function (9). By introducing the Lagrange multiplier μ, we have

$$\begin{aligned} & \min \ f(b_1, b_2, ..., b_N, \mu) = E(e^2) + \mu \cdot (b_1 + b_2 + ... + b_N - B) \\ & s.t. \ \sum_{i=1}^{N} b_i = B \end{aligned} \tag{9}$$

We obtain.

$$\begin{aligned} b_i &= \frac{1}{2} \cdot \log_2 \sigma_i^2 - \frac{1}{2N} \cdot \log_2 (\sigma_1^2 \cdot \sigma_2^2 \cdot ... \cdot \sigma_N^2) + \frac{B}{N} \\ &= \frac{1}{2} \cdot \log_2 \lambda_i - \frac{1}{2N} \cdot \log_2 (\lambda_1 \cdot \lambda_2 \cdot ... \cdot \lambda_N) + \frac{B}{N} \end{aligned}, \tag{10}$$

where λ_i is the *i*th eigenvalue.

By applying Equation (10) for the feature vectors in eigen space obtained from the pitch contours in our internal Mandarin Chinese speech database, we get the optimal bit allocation as

$$b_{i,opt} = b_{i,equal} + \Delta b_{i,opt}; \quad b_{i,equal} = \frac{B}{N} \tag{11}$$

$\mathbf{\Delta b_{opt}}$=[*5.12 1.73 0.40 -0.40 -0.81 -1.06 -1.13 -1.22 -1.30 -1.34*];

With the integer of bit allocation, we approximately have

$\mathbf{Int(\Delta b_{opt})}$=[*5 2 0 -1 -1 -1 -1 -1 -1 -1*];

A gain value is defined to measure the efficiency between equal and optimal bits allocation coding schemes,

$$gain = \log_2 (\frac{E(e_{equal}^2)}{E(e_{opt}^2)}) = 3.47 \ (bits) \tag{12}$$

For having the same distortion, the optimal bit allocation can save 3.47 bits in average, when compared to the equal bit allocation. For example, if 1 byte (8 bits) is assigned to each feature in the eigen space with equal bit allocation, and each feature vector set contains 10 features, then the saving achieved using the optimal bit allocation is 30.1%, for the same distortion level.

5 Conclusions

In this paper, we focused on the pitch modeling and analysis in the Mandarin Chinese language. The proposed method first transforms the pitch contour into a lower-dimensional representation in the eigen space by using the well-known PCA technique. The properties and the advantages of the eigenpitch are studied in terms of discrimination capability and energy preserving. Moreover, we demonstrate the quantization properties of the proposed eigenpitch representation by presenting a bit allocation scheme that efficiently codes the pitch features in the eigen space given a fixed number of bits for every feature vector.

We have carried out experiments with our internal Mandarin Chinese speech database. The experimental results are in line with the theoretical analysis and further prove the efficiency of the proposed pitch modeling approach and optimal bits allocation scheme. Based on the presented analysis and practical results, it can be concluded that this paper provides a very promising method that can be used in many of speech related applications.

6 References

[1] M. Abe and H. Sato, "Two-stage F0 control model using syllable based F0 units", In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, San Francisco, USA, 1992.

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

[3] T. Dutoit, *An introduction to text-to-speech synthesis*. Kluwer, Dordrecht, 2001.

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

[5] K. Iwano, T. Seki and S. Furui, "Noise robust speech recognition using prosodic information", In *Proceedings of International Workshop on DSP in Mobile and Vehicular Systems*, Nagoya, Japan, 2003.

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

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

[8] R. Sproat, *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*. Kluwer, Dordrecht, 1998.