STUDY OF THE SIMILARITY BETWEEN LINGUISTIC TONES AND MELODIC PITCH CONTOURS IN BEIJING OPERA SINGING

First author
Affiliation1
author1@ismir.edu

Second author
Retain these fake authors in submission to preserve the formatting

Third author
Affiliation3
author3@ismir.edu

ABSTRACT

Features of linguistic tone contours are important factors that shape the distinct melodic characteristics of different genres of Chinese opera. However, the traditional analysis of complex melodic-tone relationships in Chinese operas is limited by the scope of manual analysis. In this paper, we propose a novel data-driven approach to analyze a corpus of Beijing Opera (381 arias) with statistical modeling and machine learning methods. A total number of 1,993 pitch contour units and attributes were extracted from a selection of 20 arias. We then build Smoothing Spline ANOVA models to compute matrixes of average melodic contour curves by tone category and other attributes. K-means and hierarchical clustering is applied to SAX-based feature vectors on the full dataset and a well-separated subset. The results indicate that with-in category pitch contour is best predicted by artist's style, while the data shows a high degree of variance in all categories. We discuss the key methodological issues in melody-tone analysis and future work on pair-wise contour unit analysis.

1. INTRODUCTION

Recent development in signal processing and cognitive neuroscience, among other fields, has revived the research on the relationship between speech and musical melody [9]. Singing in tone languages offers a particularly convenient entry point to compare musical and speech melodies, allowing us to gain insight into the ways the prosody of a particular language shapes its music. In a tone language, as opposed to an intonation language, the pitch contour of a speech sound (often a syllable) can be used to distinguish lexical meaning. In singing, however, such pitch contour can be overridden by the melody of the music, making the lyrics difficult to decode by listeners.

In such consideration, musicologists have observed that features of the prosody of the local dialect often play an important role in shaping the melodic characteristics of the regional operas in China. For instance, in Beijing opera, Pian [8] discussed the competing constraints of music and tone in Beijing opera lyrics textsetting and composition. Stock [14] examined a single performer's style and concluded an artist's consideration on musical style may outweigh the similarity to linguistic tones. On the other hand, it is generally assumed that Beijing opera had incorporated linguistic tone systems from both the Hu-Guang (HG) dialect and Beijing (BJ) dialect. [1] Xu [19] reviewed 90 years of research on the dialect tone system in Beijing opera, and concluded that there is no agreement as to which system is predominant in shaping the melodic characteristics of the genre.

In sum, previous work indicates that the overall degree and manner of the melody-tone relationship is not entirely clear, partly due to the limitation that music scholars typically were not able to go beyond analyzing a few arias by hand [19]. In this paper, we propose a novel approach to melody-tone similarity by applying statistical modeling and machine learning methods to a set of 20 arias selected from a corpus of 381 arias of Beijing opera audio recording. The research questions are defined as followes: (1) How similar are syllable-sized melodic contours within a given tone category? (2) How similar is the "average" melodic contour to its corresponding "average" contour in speech in the same tone category? (3)Which tone system (BJ or HG) better predicts the shape of melodic contours?

Following preprocessing, we build Smoothing Spline ANOVA Models to compute matrixes of average melodic contour curves by tone category and other attributes. K-means clustering is applied to 30-point feature vectors of pitch contours, as well as dimensionality-reduced feature vectors represented symbolically using the SAX algorithm [7]. Hierarchical clustering is applied to feature vectors of all pitch contours as well as a set of "well-behaved" data vectors as computed by the SAX representation.

The current paper reports the results from a initial phase of a larger study of music and lyrics in Beijing opera in the context of MIR.

2. MELODY AND TONE IN BEIJING OPERA

2.1 Beijing Opera: Performance Practice

Several features of Beijing opera may explain why the melody-tone relationship remains challenging. First, the composition process of Beijing opera assumes no designated com-

[1] A schematic representation of the four tones in these two systems is shown in Figure 1.

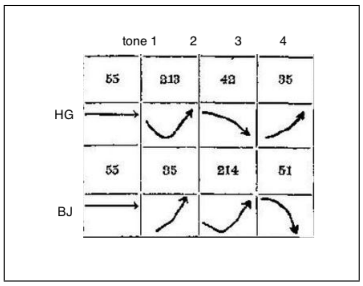

Figure 1. Schematic representation of the BJ and HG tone system

poser for any given opera. Rather, each opera is composed by re-arranging prototype arias from a inventory of arias according to the rhythmic type, role type, tempo, and other factors. Lyrics, taken from libretto, are relatively fixed whereas the specific melody may change according to each performer / performance. While there has been evidence [14] that the performers do take special consideration of tones in their singing, it is not clear in what manner and to what degree they are doing this. As a bridge from previous work to the novel approach proposed in this paper, we discuss several key issues and assumptions in studying the melody-tone relationship with important implications for current and future work.

2.2 Key Issues in Studying Tone-Melodic Similarity

First, the melody-tone relationship as a problem of tone perception (and production). A key assumption underlying previous works is that speech comprehension in tone language crucially depends on the availability of tone contour information. However, recent development in tone perception and modeling has challenged this view and revealed the great capacity of human listeners in identifying words in tone languages, with the minimum amount of tone information and despite a great amount of variances in tone signals. We consider the following aspects of evidence: (1) tone contour shapes in connected speech deviates significantly from its standard (canonical) shapes due to co-articulation, limit of maximum speed of pitch change in vocal folds, and other factors [18], introducing a great amount of variances; (2) Patel et al [10] demonstrated that Mandarin speech in monotone is 90% intelligible to native speakers in a non-noise background, pointing to a low entropy (i.e., high predicability) of the information that is carried by segmental contrast in context; (3) Gating experiments [21] have demonstrated that Mandarin speakers are able to correctly detect tone category based on only the initial fractions of second of a tone signal. From these evidence, we should use caution in making the aforementioned assumption about tone information in music. Similarly, we may also expect to find a even larger amount of variation in the syllable-sized melodic contours in a given tone category. [2]

Second, in the current endeavor, if similarity is not found in a study of reasonably large sample size, there can be two logical conclusions: (1) the similarity does not hold; (2) we haven't been looking at the right dimension where the similarity resides. Currently we have identified two ways that the melody-tone similarity may reside: First, melodic contours may be similar to the pitch contours of a given linguistic tone in a single-syllable-sized unit; and second, in a more abstract manner, the similarity resides in the relative height of a given syllable-sized pitch contour, and its previous syllable unit. In fact, a preliminary hand analysis of a small sample has indicated that the second type of similarity is much more robust than the first. However, given the limited scope of the current study, we leave the second type of similarity to future investigation.

3. DATA COLLECTION

3.1 Beijing Opera Audio Data Collection

The music in Beijing opera is mainly structured according to two basic principles, *shengqiang* and *banshi*, which in a broad sense define respectively its melodic and rhythmic components [17]. On top of these two structural principles, the system of role-types impose particular constrains to the execution of shengqiang and banshi. The interaction of these three components, hence, offers a substantial account of Beijing opera music. Our current collection includes 48 albums, which contain 510 recordings (tracks) featuring 381 arias and over 46 hours of audio [13].

The current study focuses on a small selection of 20 arias from the corpus to serve as a manageable starting point of the melody-tone relationship analysis. This set is selected according to a number of criteria: (1) we selected only *yuanban*, a rhythmic type in which the duration of a syllable sized unit bears the most similarity to that of speech; (2) we selected both types of *shengqiang*, namely *xipi* and *erhuang*; (3) we selected five role types: D(*dan*), J(*jing*), LD(*laodan*), LS(*laosheng*), and XS(*xiaosheng*). For each combination of *shengqiang* and role types, we selected two arias, yielding a total of 20 arias for analysis.

3.2 Data Preprocessing

The vocal segments of the audio recordings of the 20 arias are partially-automatically segmented into syllable sized unit (syllable unit includes the initial consonant and the nucleus (vowel+coda consonant) of the syllable). [3] The segmentation is implemented as timestamps of a TextGrid file

[2] We must bear in mind also that speech tones are generated under a different mechanism than pitch contours in singing. For one thing, the latter has a more planned mechanism of design - the composition of the music. In speech, as the qTA model has demonstrated [11], speakers may have a pitch target (defined by a linear equation) in mind during articulation, but the actual F0 realization is subject to a set of much complex physiological and contextual linguistic factors, which may be modeled by a third-order critically damped system [11]. This complication does not exist in music: in singing, a singer can realize the exact F0 target as planned. Therefore, we propose that approaches that directly compute similarity between melodic and linguistic tone F0 contours should be ruled out.

[3] Automatic segmentation using forced-alignment with machine-readable form of the score is currently being developed. For the current study, we used the result of a trained classifier [3] that is able to separate the pure instrumental frames of the audio signal from those frames that contain both vocal and instrumental parts.

in the speech processing software Praat [2]. The textgrid is later integrated with the metadata labels from the annotation process.

Following segmentation, we annotate the audio with lyrics extracted from the online Beijing opera libretto database `jingju.net`. The Chinese-character lyrics files are converted into romanized pinyin form with tone marks in the end (1,2,3, or 4) using an implementation of Java library `pinyin4j`. A Praat Script is implemented to automatically parse the romanized lyrics files and to annotate the segmented audio files. The metadata attributes (*shengqiang*, role type, artist, and structural positions [4]) are also automatically annotated for each segmented unit.

3.3 Pitch Contour Extraction

We then proceed to the extraction of F0 values for each annotated pitch contours of interest. The F0 is computed using the MELODIA [12] package within the Essentia audio signal processing library in Python [1], in order to minimize the interference of background instrumental ensemble to the computation of F0 of the primary vocal signal. All rows of F0 values associated with a specific pitch contour is automatically assigned a unique `pitch contour id` for the convenience of analysis in later stages. For the sake of analysis (especially clustering), a 30-point F0 vector is computed for each pitch contour.

4. PROPOSED APPROACH

In this section we overview the methodology employed in the analysis of the extracted pitch contour dataset. As discussed above, all methodology are boiled down to addressing the research question (1), which attempts to analyze and describe the variances and clusters found in melodic contours of each tone category and across categories. Research question (2) and (3), both of which involve comparing music with speech melody, can only be addressed indirectly by the average curves computed by the SSANOVA model for each tone category.

4.1 Methodology

4.1.1 Smoothing Spline ANOVA

First, we build a Smoothing Spline ANOVA model with the goal of (1) computing average pitch contours for each tone category, and (2) analyzing the variances accounted for by each predictor variable. Smoothing splines are essentially a piecewise polynomial function that connects discrete data points called knots. It includes a smoothing parameter to find the best fit when the data tend to be noisy, estimated by minimizing the following function:

$$G(x) = \frac{1}{n}\sum_{all\ i}(y_i - f(x_i))^2 + \lambda\int_a^b (f''(u))^2\,du \quad (1)$$

[4] The structural position is included as an attribute in the current study due to its importance in the phrasal structure of Beijing opera, originated from the 7- or 10- syllable poetic phrasal structure. It is hypothesized that melodic contours at important structural positions, such as phrasal boundaries, bears more importance musically and we expect to see less similarity with linguistic tone contours at those positions.

where n is the number of data points, λ is the smoothing parameter, and a and b are the x coordinates of the endpoint of the spline.

The Smoothing Spline ANOVA (SSANOVA) is of the following form, each component of f is estimated with a smoothing spline:

$$f = \mu + \beta x + main\,group\,effect + smooth(x) + smooth(x; group) \quad (2)$$

The SSANOVA does not return an F value. Instead, the smoothing parameters of the components smooth (x)and smooth (x); group are compared to determine their relative contributions to the equation. The SSANOVA has been applied in a number of fields where similarities and differences of curve shapes must be assessed, including plots of circadian rhythms in different group of patients, [16], ultrasound plots of tongue shapes in articulatory phonetics (Davidson 2006). In this paper, we use the implementation of SSANOVA package in statistical computing language *R*.

4.1.2 K-means Clustering

K-means clustering is performed to the pitch contour dataset with the goal of (1) seeing subgroups of shapes in each tone category; (2) discovering the clusters in the whole dataset of pitch contours and comparing them to the tone category labels; (3) discovering the inter-category differences in the degree of separability (i.e., which categories are more similar and confusable). Due to the known effect that the notion of distance tends to break down in data with high dimensionalities, in this study, two kinds of time series representations are used for k-means clustering: (a) a 30-point pitch contour vector extracted from the raw pitch contours computed by MELODIA; and (b) a series of dimensionality-reduced feature vector (dimension=2,3,4,5). The considerations of time-series representation will be discussed in the next section.

Gauthier et al [4] utilized a dataset of 33-point tone contour vectors in a tone classification task with Self-Organizing Map (SOM). In this study, the original 30-point pitch contour vector is used to perform hierarchical clustering across subsets of pitch contour data to investigate whether any subset of data can be shown to cluster into four categories that correspond to the four tone labels. In this time-series clustering task, the distance matrix is computed using the Dynamic Time Warping (DTW) instead of a standard Euclidian distance measure.

4.2 Time Series Representation

Finding the best measure of time series representation in the current clustering tasks requires careful thought and experimentation. The challenge of this task is that, it is somewhat different from a standard pitch contour similarity task addressed in previous MIR work. In a standard task, such as query-by-humming (QBH), the goal of the task is to match the melody as close as possible. However, in this task, our goal is in a way to model the human

perception of tone. An important capacity of human cognition is its capacity to abstract away the commonalities from groups of pitch contours with much different fine detail variations. In this study, we experiment with the Symbolic Aggregate approXimation (SAX) representation of pitch contour vectors. SAX is the first symbolic representation for time series that allows for dimensionality reduction and indexing with a lower-bounding distance measure. In classic data mining tasks such as clustering, classification, index, etc., SAX is as good as well-known representations such as DWT and DFT, while requiring less storage space. [7].

Even though SAX representation is mostly used outside of MIR, it has been applied to the QBH task [15]. It transforms the pitch contour into a symbolic representation with a user-designated length (nseg=desired length of the feature vector) and alphabet size (m), the latter being used to divide the pitch space of the contour into m parts assuming a gaussian distribution of F0 values. It is in principle very suitable for the current task as discussed above, as it is able to transform the fast-changing time varying signal of pitch contour into a coarse representation of abstract ”shapes”, which mimics the human cognition.

The next sections report the procedures and results of the proposed methodology and experiments with SSANOVA and clustering.

5. SMOOTHING SPLINE ANOVA ANALYSIS

5.1 Procedure

SSANOVA models are built with the dataset from each tone category in the ssanova package in *R*. Predictor variables (artist, role type, shengqiang, position) are both independently and incrementally included in different versions of the model to compare the model's power to account for the variances in the data. The best model is then selected to produce the average curve of each tone category, as well as the Bayesian Confidence Interval computed from standard error. Group effects of the best predictor variable(s) are also analyzed.

model parameter	R-squared
artist	0.062
shengqiang	0.027
position	0.028
role type	0.029
all	0.032

Table 1. SSANOVA Model comparison

5.2 Results

Results of the SSANOVA models comparison (Table 1) indicate that artist is the single best predictor of all the predictor variables (as well as all combinations of predictor variables not shown here), with R squared value being as twice as big as other factors on average. However, it is noticeable that the even the best model only explains less than 10% of the variance among all pitch curves in a given tone category. This indicates a large amount of variation in the shape of the contours. This observation is also confirmed by plotting all pitch curves in a tone category. Further

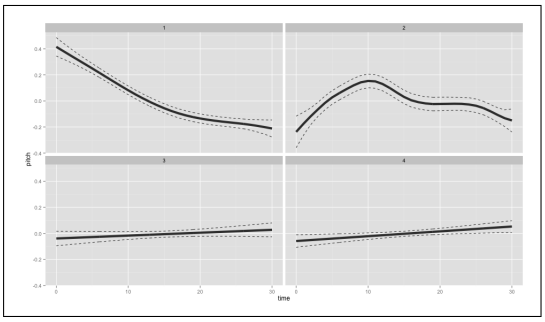

Figure 2. Average curves computed by the time+artist SSANOVA model.

analysis on the pitch contours among different artists indicates that while several artists show a significant group effect (if their confidence intervals contain the zero line, it shows that they are not significantly different in any point along the 1-30 points in the vector, see Figure 3). Meanwhile we must keep in mind that this effect is tiny in the entire dataset considering the low value of R squared and the large amount of variance.

Average curves of four tones are computed based on this model (Figure 2), with confidence intervals shown in dashed lines. The interpretation of these average curves should be done with caution, because of the low R squared value and large standard error in the model. In particular, tone 1 and tone 2 has average contours that differ from both HG and BJ system; the slope of the contour of tone 4 is significantly higher than tone 3, showing a similarity to the HG system [5] .

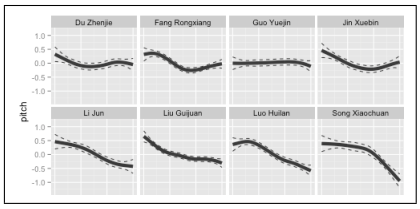

Figure 3. Difference in predicted pitch value by artist variable.

6. CLUSTERING EXPERIMENTS

6.1 K-means Clustering

6.1.1 Procedure

In experiment 1, K-means clustering is performed on the 30-point pitch contour vectors of the full dataset containing all tone category in data mining tool Weka [6] (numOfClust varied, otherwise default setting). Due to the large amount of within-category variance found in the SSANOVA,

[5] The actual realization of tone 3 in connected is very complex, and we do not make a comparison for tone 3 contours

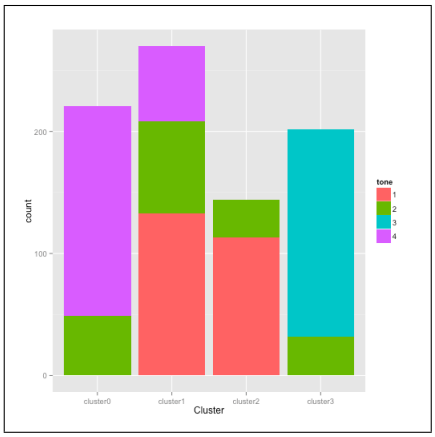

Figure 4. Supervised k-means clustering on 30-point vector

both unsupervised and supervised clustering are performed in order to make observations on the properties of the between-category difference. In experiment 2, time series in the full dataset are converted to SAX representation in Matlab [6] using different parameter settings of nseg and alphabet size. K-means clustering is then performed on the SAX time-series data sets with different parameters.

6.1.2 Results

As expected, unsupervised clustering with 30-point vectors cannot learn any meaningful grouping of tone categories regardless of the value of the number of desired clusters. However, when clustering into four groups using supervised clustering with 30-dimension vectors, it shows that tone 3 and tone 4 categories are more well-separated, whereas tone 1 and tone 2 groups showed a larger amount of variance (see Figure 4).

SAX based clustering results showed more interesting patterns. First of all, experimentation with different values of nseg and alphabet size shows that, in order to capture the abstract nature of tone contours and to be not affected by the large amount of noise in pitch movements, a limit of nseg <=3 must be placed. This is a reasonable limit considering that linguists use only two or three segments to represent tone contours in any tone language [7]. Alphabet size=3 or 4 are shown to be big enough to capture the levels of division in the pitch space, while small enough not to introduce too fine-grained labeling. Second, the result of SAX-based clustering, regardless of choice of parameters, showed that all shapes (level, rising, falling, rising-falling, falling-rising, etc), are present in almost uniform distribution in all tone categories. Finally, even though k-means clustering on SAX-based full dataset did not yield any meaningful, well-separated between-category clusters, the result is utilized in the next section to select a core, 'well-behaved' data set for hierarchical clustering.

[6] The matlab code is modeled after the SAX demo code available from Jessica Lin at http://www.cs.gmu.edu/ jessica/sax.htm.

[7] In linguistics convention, high tone=H, low tone= L, rising=LH, falling=HL, falling rising=HLH, etc

6.2 Hierarchical Clustering with DTW Distance

6.2.1 Procedure

Hierarchical Clustering with DTW distance matrix is performed in *R* on the whole dataset of pitch contours, as well as the dataset from individual artists. Finally, a 'well-behaved' dataset containing all four tone categories is chosen by the guidance of the k-means clustering results using the SAX representation [8]. In particular, all pitch contours (identified by their unique pitch contour id assigned in preprocessing) in a given tone category [9] whose SAX representation shows similarity to the linguistic tone contour in the same category are chosen to form the subset of 'well-behaved' data of the current category [10]. The purpose of this hierarchical clustering task is two-fold: first, it is a proof-of-concept task that demonstrates the performance of hierarchical clustering on a set of pitch contour data that we know are well separated into four categories; second, since we're not able to select a subset of well separated data based on any attributes (artist, role type, etc.), this is the best way to find a subset of such data by an alternative path, and we can in reverse look at its composition of attributes: is there a good combination of attributes that define this well behaved data?

6.2.2 Results

As expected, hierarchical clustering on the whole data set, as well as the individual artists data sets, did not yield any meaningful clusters based on tone category. The clustering with the core subset, however, yielded a most promising result of clustering assignments (see Figure 5) that are indeed well separated, yet have no particular correlation with any attribute factor levels. This result offers (1) a validation of the clustering methodology; (2) a validation of SAX-based representation and parameter setting in capturing the shape of pitch contours in a similar manner to linguistic representation of tones in human perception; (3) in reverse, a validation that the variance in the original dataset (or subsets by artists) is indeed too big for any clustering algorithm we have tested to successfully separate.

7. DISCUSSION

Overall, we have shown with different methods that the full dataset of single-syllable tone contours in the 20 arias of Beijing opera bears a large amount of variance in pitch contour shapes in all tone categories. They are thus not well separated contour shapes that show significant demonstrable similarity with prototypes of linguistic tone contours. This result is consistent with the SSANOVA modeling. We have anticipated the possibility of this result in the discussion of key issues in studying melody-tone similarity. Meanwhile, many valuable observations are made from the results of the analysis, including primarily the effectiveness of SAX representation in modeling human per-

[8] This includes clusters derived with SAX parameters (2,3) and (3,3).

[9] Here we used the tone contours from BJ dialect.

[10] The data vectors used in the actual clustering is still the 30-point numeric pitch contour vector.

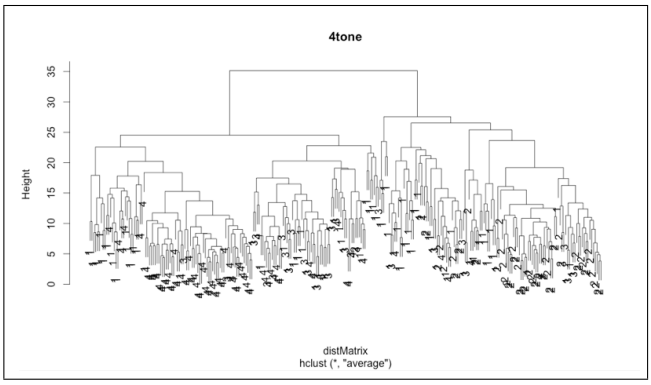

Figure 5. Hierarchical clustering on a core dataset

ception of pitch contour shapes, and the validation of hierarchical clustering with DTW distance on the current task. Given the discussion and methodology proposed here, we expect future research on the pair-wise syllable correspondence analysis to yield more promising results.

8. REFERENCES

[1] Bogdanov, D., Wack N., Gmez E., Gulati S., Herrera P., Mayor O., et al.: ESSENTIA: an Audio Analysis Library for Music Information Retrieval. International Society for Music Information Retrieval Conference (ISMIR'13). 493-498.(2013).

[2] Boersma, Paul.: Praat, a system for doing phonetics by computer. Glot International 5:9/10, 341-345.2001.

[3] Chen,K.: Characterization of Pitch Intonation of Beijing Opera. Master Thesis, Music Technology Group, Universitat Pompeu Fabra, Barcelona, 2013.

[4] Gautheir,B., Shi,R, Xu,Y.: Learning phonetic categories by tracking movements. Cognition, 103, (2007),80-106.

[5] Gu,C.: Smoothing Spline ANOVA Models. New York: Springer- Verlag.2002.

[6] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten: The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.2009.

[7] Lin,J., Keogh,E., Wei,L.,and Lonardi,S.: Experiencing SAX: a novel symbolic representation of time series. Data Mining and Knowledge Discovery. Oct.2007, Vol.15, Issue.2, pp107-144.2007.

[8] Pian,R.C.: Text Setting with the Shipyi Animated Aria. In Words and Music: The Scholars View,edited by Laurence Berman, 237270. Cambridge: Harvard University Press,1972.

[9] Patel,A.: *Music, Language, and the Brain*, Oxford Press, 2008.

[10] Patel, A. D., Xu, Y. and Wang, B.: The role of F0 variation in the intelligibility of Mandarin sentences. In Proceedings of Speech Prosody 2010, Chicago.(2010).

[11] Prom-on, S., Xu, Y. and Thipakorn, B.: Modeling tone and intonation in Mandarin and English as a process of target approximation. Journal of the Acoustical Society of America 125: 405-424.(2009).

[12] Salamon, J and Gmez E: "Melody Extraction from Polyphonic Music Signals using Pitch Contour Characteristics", IEEE Transactions on Audio, Speech and Language Processing, 20(6):1759-1770.2012.

[13] Serra,X.: "Creating Research Corpora for the Computational Study of Music: the case of the CompMusic Project", AES 53rd International Conference, January 27-29th 2014, London (UK).

[14] Stock, J: A Reassessment of the Relationship Between Text, Speech Tone, Melody, and Aria Structure in Beijing Opera. Journal of Musicological Research (18:3): 183206. 1999.

[15] Valero,J: Measuring similarity of automatically extracted melodic pitch contours for audio-based query by humming of polyphonic music collections. Master's Thesis, MTG, DTIC, UPF, 2013.

[16] Wang, Y., Ke, C., and Brown, M.: Shape-invariant modeling of circadian rhythms with random effects and smoothing spline ANOVA decompositions, Biometrics 59, 804812.2003.

[17] Wichmann, E.: Listening to the theatre: the aural dimension of Beijing opera. Honolulu: University of Hawaii Press. 1991.

[18] Xu, Y. and Sun X.: Maximum speed of pitch change and how it may relate to speech. Journal of the Acoustical Society of America 111: 1399-1413.2002.

[19] Xu,Z. 2007: Jiu shi nian lai jingju de sheng diao yan jiu zhi hui gu. (Overview of Ninety Years of Research of Tone Relations in Peking Opera). Nankai Journal of Linguistics, 10(2):39-50.

[20] Yu,H.: Qiang ci guan xi yan jiu. (Study on the relationship between tone and melody). Beijing: Central Conservatory Publishing House.2009.

[21] Lai, Yuwen and Jie Zhang. : Mandarin lexical tone recognition: the gating paradigm. In Emily Tummons and Stephanie Lux (eds.), Proceedings of the 2007 Mid-America Linguistics Conference, Kansas Working Papers in Linguistics 30. 183-194.2008.