PREDICTING PAIRWISE PITCH CONTOUR RELATIONS BASED ON LINGUISTIC TONE INFORMATION IN BEIJING OPERA SINGING

Shuo Zhang, Rafael Caro Repetto, Xavier Serra
Music Technology Group, Universitat Pompeu Fabra
ssz6@georgetown.edu, {rafael.caro, xavier.serra}@upf.edu

ABSTRACT

The similarity between linguistic tones and melodic pitch contours in Beijing Opera can be captured either by the contour shape of single syllable units, or by the pairwise pitch height relations in adjacent syllable units. In this paper, we investigate the latter problem with a novel machine learning approach, using techniques from time series data mining. Approximately 1300 pairwise contour segments are extracted from a selection of 20 arias. We then formulate the problem as a supervised machine-learning task of predicting types of pairwise melodic relations based on linguistic tone information. The results give a comparative view of fixed and mixed-effects models that achieved around 70% of maximum accuracy. We discuss the superiority of the current method to that of the unsupervised learning in single-syllable-unit contour analysis of similarity in Beijing Opera.

1. INTRODUCTION

One of the most salient aspects of Chinese operas is the role of various dialects and their distinct tone contours. In the musicological study of Beijing opera, the similarity between linguistic tone contours of the lyrics and the vocal melodic contours is a classic problem that arises from the nature of Chinese tone languages. 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 [1]. In order for lyrics to be more intelligible, Beijing opera's melody is traditionally arranged with considerations of lyrics tone information. The degree and manner of this incorporation, however, is only partly known through scholarly work [1-4]. The difficulty of this problem is further complicated by the fact that there are two dialects with distinct tone contours within Beijing opera (Beijing and HuGuang dialects, or BJ and HG in this paper) [3].

Previous works cited above indicate that the similarity between linguistic tones and melodic pitch contours in Beijing Opera can be captured either by the contour shape of single syllable units, or by the pairwise pitch height relations in adjacent syllable units. [1] considered the single-syllable unit contour analysis with a time-series data mining approach. This study concluded that while the Smoothing Spline model's R-squared values are consistent with the expected variance relations between the first tone and other tones, overall there exists a large amount of un-explained variance in melodic contours that cannot be attributed to grouping of tone categories from a single tone system (BJ or HG).

In this paper we investigate the second type of similarity of linguistic tones and melodic contours. Following musical literature [12], we postulate that the perceived similarity of the melody to a tone category is realized by the similar pitch height relations in a pair of adjacent syllable units in singing (to that of the tone in speech). We then formulate this problem as a supervised machine learning problem of predicting the type of pairwise pitch height relations based on features derived from linguistic attributes. First we perform experimentation on the most efficient and cognitively accurate time-series representations for pitch contour vectors and extract the class labels. Following feature extraction and data preprocessing, a series of multinomial, binary and mixed effects regression models are trained. These allow us to progressively achieve our two main goals: First, using linguistic information to predict (with improved accuracy) the melodic pairwise pitch height relations; second, as a consequence, we also obtain a better understanding of the effect of various linguistic and other attributes on the types of pitch height relations observed in Beijing opera.

The remainder of the paper is organized as follows. Section 2 gives the formulation of the pairwise tone-melody similarity as a supervised machine learning problem, followed by the description of data collection and preprocessing in Section 3. The core methodologies of time-series data representation experimentation and model training are described in Section 4. Section 5 and 6 discuss the results, including the comparison of models and interpretation of model parameters.

2. PROBLEM FORMULATION

Recent research revealed that tone identification by humans does not necessarily depend on the availability of full tone contour information [5]. In the light of this finding, pairwise tone-melody similarity is therefore a cognitively plausible way for the melody to convey underlying tone information without being fully similar to the contour of the linguistic tone. For example, a high-level tone

Attribution: First author, Second author, Third author. "Paper Template For ISMIR 2015", 16th International Society for Music Information Retrieval Conference, 2015.

(5-5)[1] followed by a low-rise tone (2-4) can be reflected in melody as long as the perceived starting pitch of the second syllable is lower than the first. Perceptually, the beginning position of a syllable is the most salient, being a prominent position that carries much phonetic information such as formant transitions [6]. Alternatively, one may propose that this relation can be reflected by the ending pitch of the first syllable and the beginning pitch of the second syllable, being the closest pair in time. Less plausible is the case where this similarity is reflected in the ending region pitch of both syllables.

We therefore formulate this problem of pairwise similarity as a supervised machine learning problem. First, we define three types of relations between the two syllables in proximity (mostly adjacent, but can be separated by a short instrumental interlude), based on the relative pitch height: ascending (A), descending (D), and level (L). These are our target class labels. Second, we define three subtypes of pairwise similarity based on the location of the similarity: Onset-Onset (BB), Offset-Onset (EB), and Offset-Offset (EE). We will train a separate model for each type of similarity. Third, we formulate the research objective: given linguistic tone and other attributes of a pair of syllables in the lyrics, can we correctly predict the type of relations of relative pitch height in vocal melody (A, D, or L as class label)?

3. DATA COLLECTION

3.1 Data Collection

The current study uses about 1300 syllable-sized contours extracted from a selection of 20 arias in a pre-segmented and annotated Beijing opera audio collection corpus [7]. Each syllable in this data set is annotated with linguistic tone, word, artist, role type, melodic type (*shengqiang*), rhythmic type (*banshi*), and relevant metadata information. 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. This set of arias is selected by a music scholar with expertise in Beijing opera music (who is the second author), and is therefore a representative set that is both comprehensive and selective for the task of studying the tone-melody relationship.

3.2 Pitch Contour Extraction

The fundamental frequency of vocal melodic contours is computed using the MELODIA [10] package within the Essentia audio signal-processing library in Python [11], 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 are automatically assigned a unique pitch contour id that encodes the aria, tone, and temporal order information of the syllable. For the sake of analysis, we produce down-sampled 30-point F0 vectors by using equidistant sampling across each pitch contour. A 5-point weighted averaging sliding window is applied to smooth the signal. The single-syllable contour data is then converted into a pairwise-syllable contour data file where each row has 60 pitch points of the two adjacent syllable contours, plus other attributes.

4. METHODOLOGY

4.1 Time-series representation

First we perform automatic extraction of our target class labels (A, D, or L). In order to capture the accurate perceived pitch heights in the beginning and ending regions of each syllable-sized melodic contour, we first convert the 30-point pitch contour into a lower dimension representation using the **S**ymbolic **A**ggregate appro**X**imation (SAX) [8]. SAX transforms the pitch contour into a symbolic representation using Piecewise Aggregate Approximation with a user-designated length (`nseg,` or sometimes referred to as word size, is the desired length of the feature vector) and alphabet size (`alpha`), the latter being used to divide the pitch space of the contour into `alpha` parts assuming a Gaussian distribution of F0 values. [2]

Using this technique, we experiment with the parameter settings of SAX with the goal of yielding the most similar relation types as a human listener would judge it. To perform this experiment, we first have a human listener annotate a selection of 260 sample tone contours extracted from our audio collection[3]. For each contour, the listener would rate the type of pairwise relation (A,D or L) by listening through the contour pairs. The experiment is proctored automatically by a Praat Script program and the presentation of each pair is separated by a white noise of 5 seconds. The listener rates all 260 contours consecutively.

[1] The numbering here follows the relative pitch height from low to high: 1<2<3<4<5.

[2] Here, the Gaussian distribution is used to obtain the break points for vertical pitch space so that each region (represented by a symbol) is equal-probable (probability of that symbol is given by the integration of the area under the Gaussian curve to as defined by the break points). This ensures that the probability of a segment being assigned any symbol is the same[8].

[3] Human rater is used only to train the parameters on a smaller sample so that we can perform automatic label extraction on larger scales of data.

Next, we permute the single syllable contour unit parameter values[1] within intervals `nseg` ([3,8]) and `alpha` ([3,6]). Each combination of the parameters yields a SAX representation for all contours and then three pairwise relation types (BB, EB, EE) based on the representation is extracted. We then compute the similarity / accuracy of each representation to the human judgments. Here, it is noteworthy that the perceived beginning pitch height of a syllable-sized melodic contour does not necessarily correspond to a predetermined meaningful musical unit such as a note. This is due to the nature of the Beijing opera that has many fine melisma across the melody. In this case, we do not attempt to define a perceptually or musically grounded unit of 'beginning pitch', but rather we will let the experiment results decide which parameter configuration would be the closest to the human judgment. Since SAX is already a dimensionality reduction algorithm, we thereby define the beginning pitch of a syllable as the first symbol in the symbolic time-series representation. After the parameters are chosen, we use the SAX representations to extract pairwise relation types for the entire training set.

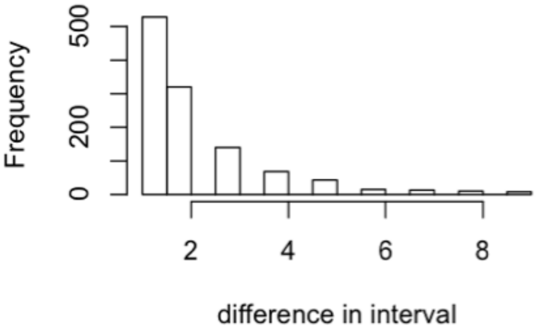

Figure 1 Histogram of interval distances between pairwise syllables

4.2 Regression modeling

4.2.1 Feature Extraction and Data Preprocessing

The basic set of features includes the attributes extracted from the corpus annotations: `ToneFirst`, `ToneSecond`, (of the first and second syllable in a pairwise contour), word, artist, role type, *shengqiang, banshi,* as well as the duration of both syllables. All except the last one are nominal attributes.

We additionally extracted three sets of compound features based on linguistic tones: first, a `toneCombination` feature that encodes the particular tone combinations (e.g., tone1_tone2) of this pairwise contour; second, six other features that encodes the types of linguistic tone pitch height relations of this tone combination (BB, EB, and EE) as well as the two dialectal tone systems (BJ and HG). These features are therefore (BB_BJ, BB_HG, EB_BJ, EB_HG, EE_BJ, EE_HG). These are the only features that directly encode the types of pairwise linguistic tone relations into the feature vectors, using numbered pitch height system from linguistic literature (e.g., tone 3 in BJ is 214 and in HG is 42, 1<2<3<4). Theoretically these features should not be used all at once, since each one of our regression models would only account for one type of output relations (BB, EB, or EE). However, since the previous studies suggest that the BJ and HG tone systems are likely intermixed in affecting the output melodic contour[1], we include both of these two types of features in each model. A third feature encodes the temporal distance/ number of interval segments between the pair of syllable: it is hypothesized that a closer pair of syllable would contribute to the manifestation of linguistic information. We have eliminated those pairs whose distance is greater than 10 intervals (the intervals in between a pair could be due to various reasons, but mostly likely instrumental interlude). Figure 1 shows the distribution of distance in units of time intervals in the entire data set (where an interval is a syllable unit in our data segmentation). From this distribution we can see that most pairs are sung consecutively.

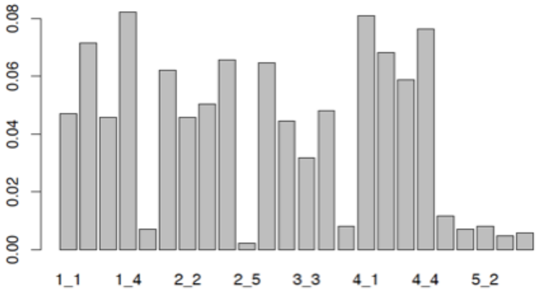

Figure 2 Barplot of the frequency of tone combination types in training data, where x_y indicates tone x to tone y combination

We perform several measures of data preprocessing on the training data. First, we eliminate all contours whose syllable duration is longer than a threshold of 5s(based on the distribution of durations). This is based on the observation that if the syllable is too long, the temporal relations are sparser and it has more chance to musically embellish its contours, further obscuring the linguistic information. Second, we observe that there is an extreme imbalance between tone categories 1-4 and tone 5. Linguistically, tone 5 is a 'neutral' tone that carries different contours according to their context. Figure 2 shows this imbalance in the `toneCombination` feature. We eliminated all examples with tone 5 in order to avoid singularity problems with generalized linear modeling.

Lastly, the output class label distribution is also highly skewed (Figure 3). This is an interesting property of this musical data set especially when compared with its expected counterpart in language (i.e., the set of six tone features that encodes pairwise tone pitch relations). Figure 3 plots the set of pairwise musical pitch relation labels (BB, EB, EE) alongside its expected counter part (BB_BJ, EB_BJ, etc.) linguistic tone feature distribu-

[1] To ensure the consistency of pitch space division with the Gaussian breakpoints, we convert a pair of syllables at a time, making the `nseg` parameter twice as big.

tions. It is noteworthy that not coincidentally, the linguistic pairwise types have a quite uniform distribution whereas the musical pairwise types have a very skewed distribution, with the "L(evel)" label being the rare class. This is probably a product of music: music, being a play largely about the manipulation of pitch, is intentionally avoiding many of the adjacent syllables (or notes) starting or ending with the same pitch height. Therefore in these cases we may hypothesize that the music is overriding linguistic configurations, thus obscuring our model. For this reason as well as motivations from the machine learning perspective, we created a second data set where all "L" labels are removed from the training data (which is a small portion, ref. Figure 3). Therefore we use this second data set for binary logistic and mixed-effects regression modeling in the latter part of the study.

4.2.2 Multinomial Regression

Our first model approximates this problem with a multinomial logistic regression using the original data set with three output class labels (A, D, L). The multinomial logistic regression is an extension to the binary logistic regression modelling, where we train one-versus-other models for each of the class labels. The model outputs the probability of assigning each label and selects the label with the highest probability as the predicted label.

For all of the algorithms used in this study, as previously discussed, we build three models assuming three different types of relations (BB, EB, EE). We first build baseline multinomial logistic regression models with all available basic features. Then we incrementally drop features whose coefficients are insignificant and having low predicting powers and end up with the best model for this setting. This set of features is used throughout the rest of the models in conjunction with compound features.

4.2.3 Binary (Fixed Effects) Logistic Regression

We then perform all subsequent regression modeling on the binary data set. As a baseline for this data set, we perform classic fixed effects binary logistic regression and compare the result with a number of well known machine learning algorithms such as Support Vector Machine (SVM), decision tress (J48 in Weka), Neural Networks, and Naïve Bayes.

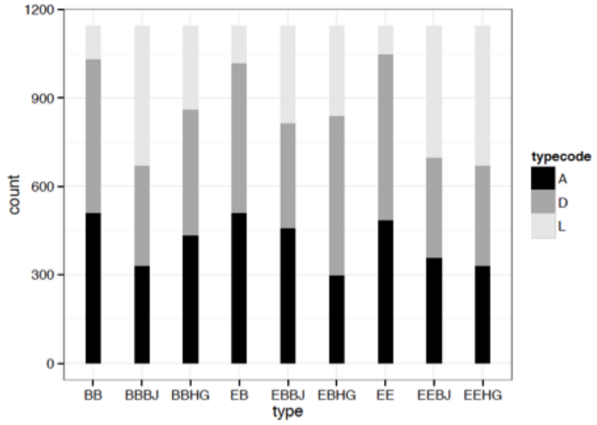

Figure 3 Distribution of pairwise pitch relation types across class labels (BB, EB, EE) and corresponding expected linguistic feature distributions

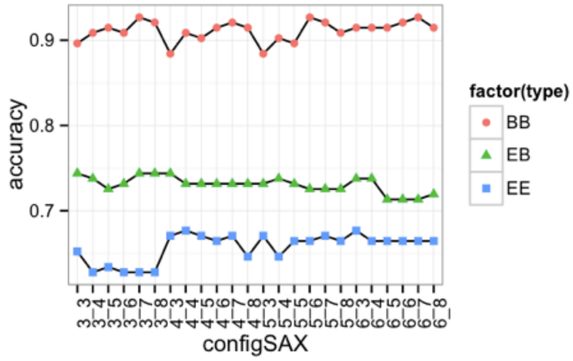

Figure 4 Accuracy of pairwise pitch relation type prediction by different SAX parameters (`alpha_nseg`) with z-transform

4.2.4 Mixed Effects Logistic Regression

A mixed-effects regression model performs prediction by combining the contributions from fixed effects and random effects. Parameters associated (coefficients) with the particular levels of a covariate are known as the "effects" of the levels. Essentially, if the set of possible levels of the covariate is fixed and reproducible we model the covariate using fixed-effects parameters. If the levels that we observed represent a random sample from the set of all possible levels we incorporate random effects in the model [9].

We extend from generalized linear models (GLMs) to multilevel GLMs by adding a stochastic component **Z** to the linear predictor (see (1)), where the random effects vector **b** is normally distributed with mean 0 and variance-covariance matrix Σ. In a mixed-effects logistic regression model, we plug the stochastic linear predictor in the binomial (logistic) linking function.

$$\eta = \alpha + \beta_1 X_1 + \cdots + \beta_n X_n + b_0 + b_1 Z_1 + \cdots + b_m Z_m \qquad (1)$$

In the current setting, the random effects of our feature correspond to the variable `words`. Any general model would usually exclude the particular words of the two syllables as a fixed-effect feature; however, the problem with the fixed effects model (like the one described above) is that it is not capturing the variances in the output label caused by being different words. The mixed effects model corrects this by estimating the conditional mode of the random effect term coefficients **B**. Strictly speaking, we don't estimate the random effects in the same sense that we estimate model parameters. Instead, we consider the conditional distribution of B given the observed data, (B|Y =y), where Y is the output class label [9].

All modeling in this study are done with 10-fold cross validation in the training phase.

5. RESULT

5.1 Time-series Representation Experiment Results

Overall the time-series representation experiments using SAX technique yielded informative results. First, as shown in Figure 4, the differences in the accuracy are mostly the effect of varying `nseg,` whereas the effect of `alpha` parameter is not apparent (except for the EE type). Second, somewhat surprisingly, the average type prediction accuracy vary significantly across the three types: BB>>EB>EE, with Onset-to-Onset relation types performing almost perfectly at peak accuracy of 93%. This contrast is surprising considering that the human listener (who is a skilled musician) did not rate these three types directly; instead, the listener only had to rate the beginning and end of a single syllable on a numbered fixed scale[1] (almost like transcribing the end notes) and the values for the three types of relations were extracted using that reference scale automatically. This systematic difference in the accuracy of these three types of relations indicates that the Offset annotation / judgment has a much lower correlation with the acoustic signal than the Onset annotation. One possible explanation for this could be that the ending of a syllable is embellished with more melismas than the beginning portion of the syllable, making the correlation lower. However, that would not predict the systematic lower performance of SAX EB and EE, which is at a lower resolution than the original acoustic signal. An alternative explanation is that the offset position is not as salient as the onset position, making it less appropriate for the location of carrying linguistic tone information.

Due to the differences in performance, we choose the SAX parameter settings for each of the three types: {BB:6_7, EB: 4_3, BB: 6_3}, where the parameter combinations stand for `alpha_nseg`.

5.2 Regression Modeling Results[2]

	Coef1(D)	p-value	Coef2 (L)	p-value
Intercept(A)	-0.0761	6e-01	-1.2447	*2e-05
Duration1	0.2149	0.0437	0.2594	0.1048
T_1_2	-0.8454	*6e-06	-0.8159	*7e-03
T_1_3	-0.6952	*0.0004	-0.4844	0.1148
T_1_4	-1.2471	*6e-12	-0.9295	*1e-03
T_2_2	0.5765	*0.0018	-0.1137	0.6986
T_2_3	0.6709	*0.0007	0.2909	0.3265
T_2_4	1.1505	*2e-10	0.2175	4e-01

Table 1. Significant (basic) predictors overview from multinomial regression for BB type, with asterisks indicating coefficients significant at 0.05 level. Coef1 and coef2 represent the regression coefficient associated with features. T_i_j is the coefficient associated with a i-th tone being a tone j.

First, results (Table 1) of multinomial logistic regression reveal that tone information and the duration of the first syllable are among the most significant predictors of the probability of a pairwise contour being one of the three output classes (A, D or L). Concretely, for example, being a Tone 2 in the first syllable would significantly lower the probability of being a "D" by log-odds -0.8454 or odds `exp(-0.8434)`, and being a tone 4 for the second syllable would significantly increase the probability of being a "L" by log-odds 0.2175 or odds of `exp(0.2175)`, where the `exp()` is the exponentiation function. Overall the multinomial regression has a mean classification accuracy of 56.7% for all types of models on the 10-fold cross validation on the entire data set. This is a lower baseline for the subsequent models. Due to the skewed output distribution of class labels, the model consistently assigns the lowest probability to "L" in all predictions.

Despite this finding, further analysis shows that the basic tone features (`ToneFirst` and `ToneSecond`) have limited predictive power compared to other compound tone features. Therefore in the subsequent analysis we drop these two basic features and keep the other 1+6 types of compound features.

Our binary logistic model on the binary class data set (class label A and D) improves the accuracy by about 9%. This result is comparable across different classification algorithms (Table 2).

Algorithm	**mean Accuracy**
Binary Logistic Regression	65.12%
Decision Tree (J48)	61.57%
SVM	62.44%
NaiveBayes	61.56%
NeuralNetwork	60.07%

Table 2. Average performance of different algorithms on the binary classification data set with a 10-fold cross validation

[1] Here, all the contours are extracted sequentially from the same aria so the judgment and extraction of consecutive relations are accurate.

[2] In this table, class A is treated as base/default level, and the `Intercepts` represent the base probabilities of D and L with regard to A without any knowledge about features.

The mixed-effect model further improves the prediction accuracy to around and above 70%. This set of models has two variations. The first set is built with `ToneCombination` and the six other compound features (two per model) as well as the `duration` of the first and second syllable as fixed effects features, and the `Word` as a simple scalar random effect feature `(1|Word)`. The second set includes more complex random effects features `(1+duration1|Word)`, which takes into account the interaction between the `duration` of the first syllable (fixed effect) and the `Word` (random effect) feature. The performance of these two sets varies between the three types of models. Table 3 gives a comprehensive overview of the evaluation of the models.

Overall, all models have shown that the prediction accuracy decreases from BB to EB to EE. This is in accordance with our initial SAX representation accuracy rank, therefore is expected. However, the underlying reason for this is unclear, as discussed previously.

Figure 5 gives an overview of the model performances based on average accuracy.

	AIC	BIC	LogLik	Sc.Residual	Accuracy
BB1	1059.3	1153.1	-509.7	-0.438	70.77%
BB2	1063.1	1166.3	-509.5	-0.437	69.99%
EB1	1086.8	1180.6	-523.4	0.526	65.30%
EB2	1096.4	1193.6	-523.2	0.512	65.8%
EE1	1104.3	1198.1	-532.2	0.685	60.95%
EE2	1107.2	1210.4	-531.6	0.649	64.80%

Table 3 Model comparison for mixed effect models, where AIC and BIC are commonly used Akaike Information Criterion and Bayesian Information Criterion

6. DISCUSSION AND CONCLUTION

The current study has considered the problem of the similarity between linguistic tones and melodic contours in Beijing opera in the form of pairwise pitch height relations. We formulate this similarity problem as a supervised machine learning problem of predicting the type of relations based on linguistic tone information. We have shown that using a set of linguistic features alone (tone and duration information), the model is able to achieve an accuracy of 65% to 70% based on the types of hypothesized relations, after we have reduced the output class labels to binary (for reasons discussed above). Here we discuss several aspects of the interpretation and evaluation of the current results.

First, the performance of the models has shown consistently that Onset-Onset is a more robust pairwise relation type compared to Offset-Onset and Offset-Offset. However, this result may be dependent upon the initial performance rank of SAX representation accuracy for these three types. To better understand this phenomenon, we performed a post-hoc re-analysis of the SAX conversion using the original fundamental frequency data without down-sampling and evaluate its accuracy. The result showed a more balanced yet overall lower accuracy on extracted class labels as compared to human annotation (75%, 72%, and 78% peak accuracy values for BB, EB, and EE). When using this set of class labels, we obtained generally lower performance on the best mixed effects models in the classification task (57%, 68%, and 58% for BB, EB, and EE). Noticeably, the BB type model has a lower prediction accuracy than the EB type, making the Offset-to-onset relation more robust. Meanwhile, there is less confidence in this result due to the general lower accuracy in the representation of the class labels.

Second, we should also bear in mind that in the current problem, the class labels of pitch height relations are dependent upon the musical considerations on top of the linguistic considerations. For all practical and theoretical reasons we believe that Beijing opera music has its own rules that at many times take precedence over linguistic rules, and that should give us a large proportion of unexplained variances when predicting pairwise pitch relations. Considering this factor, it is fair to conclude that the current models have shown effectively the high degree of pairwise similarity between linguistic tones and melodic contours in Beijing opera. For the same reason discussed above, we have justified our decision to take the “L” class out from our model because of its likely irrelevance to linguistic information (and should be explained by musical considerations).

Third, comparing the current study with previous works on the single-syllable contour similarity [1], we observe that the current approach yields higher explanatory power than the previous approach, while requiring significantly less computing resources. [1] Specifically, it is worth noting that while the contour-shape-based SSANOVA models in [1] suffers from the lack of knowledge on the exact weights of the two dialects (BJ and HG), the current approach is able to encode expected pairwise pitch relations from both dialects into the features, thus making it more effective in a supervised learning task.

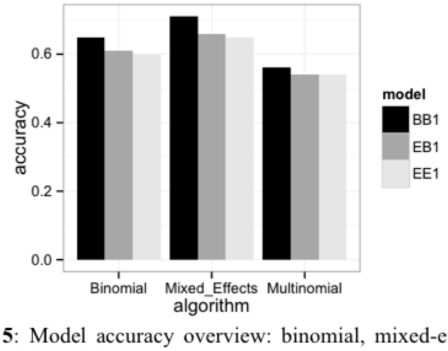

Figure 5: Model accuracy overview: binomial, mixed-effects, multinomial models

[4] This comment is in reference to the expensive computations of the DTW distance matrixes in the time-series data mining of the pitch contours in a large dataset. In addition, the SSANOVA model in [1] was only able to achieve a R-squared value of around 0.2, suggesting low explanatory power of the variance present in the data.

7. REFERENCES

[1] S. Zhang, R.Caro Repetto, S.Xavier: “Study of the similarity between linguistic tones and melodic pitch contours in Beijing Opera singing”. *Proceedings of The 15th International Society for Music Information Retrieval (ISMIR) Conference*, pp.345-348. Taiwan, October 27-31 2014.

[2] 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.

[3] Xu,Z. 2007: *Jiu shi nian lai jingju de sheng diao yan jiu zhi hui gu.* (Review of Ninety Years of Research of tones in Beijing Opera). *Nankai Journal of Linguistics*, 10(2):39-50.

[4] 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. [16]

[5] 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.

[6] Steriade, Donca. (2001). Directional asymmetries in place assimilation. In *The role of speech perception in phonology*, eds. Elizabeth Hume and Keith Johnson, 219–250. San Diego: Academic Press.

[7] C. Repetto and X. Serra X. “Creating a Corpus of Jingju (Beijing Opera) Music and Possibilities for Melodic Analysis”. *Proceedings of the 15th International Society for Music Information Retrieval Conference,* Oct. 27th-31st 2014, Taipei (Taiwan).

[8] 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.

[9] D.Bates. *Mixed-Effects Modeling with R.* Springer: 2010.

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

[11] 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).

[12] Lian, B. *Xiqu zuoqu jiaocheng* (Textbook on Chinese opera composition). Shanghai Conservatory Publications, 1999.