Probabilistic Classification of Monophonic Instrument Playing Techniques

Akira Maezawa Katsutoshi Itoyama Toru Takahashi Tetsuya Ogata Hiroshi G. Okuno
Graduate School of Informatics, Kyoto University

1. Introduction

Understanding the underlying intentions of a music performer is crucial to enable a machine such as an automated accompaniment system to interact intelligently with a musician. Particularly, understanding the symbol associated with a tone the player generates allows a machine to create response that is in concordance with the symbol. We define *playing technique* as a symbol such as the expression marking that reflects the intention of a human performer in a perceivable change in timbre.

We believe that it is crucial to understand that some playing techniques are inherently ambiguous, and to associate with an input signal the degree of ambiguity along with the estimated class labels. For example, if the class of dynamics are described by "soft" and "loud," it is irrelevant to ask whether a "moderate" sound is "soft" or "loud" - it only makes sense to say that it either belongs to both or to neither.

We shall express ambiguity by modeling a set of playing techniques as a posterior distribution, and using statistics obtained from the distribution to determine the ambiguity. We shall group playing techniques that acts on a same, continuous quality into one set that in turn generates a posterior distribution as shwon in Figure 1. Particularly, we hypothesize that ambiguous sounds are the main cause of misclassification, and such sounds creates a distribution with a high variance.

Existing research in detecting the playing technique involved discretization of techniques involving a continuous factor and converting it into a problem of classification. For example, the position of the bow on a violin was discretized by recording two points and choosing one of the two positions [1]. Other research exclusively dealt with playing techniques of discrete quality such as whether a bass guitar string was slapped or not [2]. Both approaches did not associate any posterior probability with the output label, and thus suffered when recognizing notes that even humans have trouble distinguishing [1]. Another approach involved extraction of "perceptual" features that were used to control another musical instrument [3]. This approach did not attempt to symbolize timbre into playing techniques.

In this paper, we modeled a set of playing techniques that acts on a same continuous factor (position of the bow on a violin) as a posterior distribution given the input signal using a hybrid of Gaussian Mixtures (GMM) and Relevance Vector Machine (RVM). We then rejected data whose variance exceeds a threshold given some input signal, and evaluated the performance of playing technique classification using the following:

- The recognition accuracy after ambiguous frames are rejected
- Ratio of the number of rejected frames to the total number of frames

Receiver Operating Characteristics was evaluated but is omitted for lack of space. We found that the recognition rate increases dramatically by rejecting ambiguous frames, thus hinting our hypothesis that ambiguous sound is the main cause of misclassification.

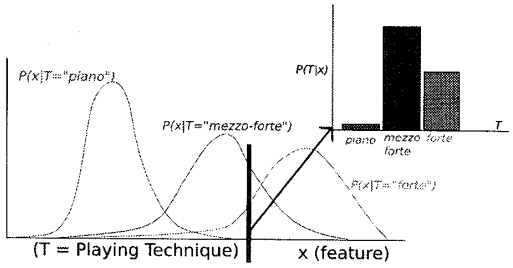

Figure 1: Family of Playing Techniques as Likelihoods

2. Method

Likelihood distributions within a family were generated by training a hybrid of clustering algorithm and multi-class relevance vector machine (RVM) classifier.

2.1 Data Acquisition and Feature Selection

Features as summarized in Table 1 were chosen based on existing research in instrument identification [6] and perceptual synthesis [3]. In addition, we also included the "Median Spectral Roll-off," which estimates the distribution of power in the frequency domain of non-harmonic components. It was determined by applying a median filter along the frequency axis with the width set to twice the width of the main lobe of the windowing function. Frames whose signal power were less than -60dB were discarded. Furthermore, a parameter vector consisting of the fundamental frequency and the negative exponent of the time after onset was associated with each feature vector. Next, the training data was normalized to have a unit variance with zero mean. Outliers, or vectors whose Euclidean norm is greater than 25 ($= 5\sigma$), were removed. This process was iterated until no further data was removed.

Table 1: Preliminary Features Used. F_k denotes the frequency of the k-th overtone, and N ranges from 1 to 19.

Feature	*Dimension*
F_k-to-F_0 Power Ratio	12
Spectral Spread about F_k	13
Spectral Kurtosis about F_k	13
Spectral Flatness	1
Spectral Centroid	1
Spectral Spread	1
Signal Power	1
MFCC	14
ΔMFCC	14
Δ^2MFCC	14
Spectral Roll-off with threshold $= 4N\%$ of total power	19
Median Spectral Roll-off with threshold $= 4N\%$ of total	19
Zero-Cross Rate	1
Total	123

2.2 Supervised Training of the Playing Styles

A hybrid of Gaussian mixture and multiclass relevance vector machine (RVM) classifier was used for generating the

training distribution. RVM was chosen because it could provide probabilistic interpretation to the output, it could support highly non-linear decision boundary, and it tends to produce few relevance vectors (akin to the support vector in SVM) [4]. These were all important characteristics since the shape of the decision boundary is not well-known and a quick response was desired for our application.

Training time for a RVM with number of classes and training vectors used in our application was too impractical and therefore, we generated a mixture of 512 Gaussians based on the parameter vector described previously. Training data was associated to four Gaussians with the greatest likelihood. A RVM using Gaussian kernel of width 8 and 2 was trained for each Gaussian in the mixture. No bias was incorporated because that caused "ambiguous" sound to be classified into one with the greatest bias term.

Let v_p be the parameter vector, θ the playing technique, v_x the feature vector, ϕ_k the k-th Gaussian of the mixture of $K = 512$ Gaussians, each with associated weight, mean and covariance matrix, w_k, μ_k and Σ_k. That is, $P(v_p|\phi_k) \sim N(v_p;\mu_k,\Sigma_k)$ and $P(\phi_k) = w_k$. The parameter vector v_p and the likelihood θ are conditionally independent given the k-th cluster, ϕ_k because the calculation of the likelihood at the k-th RVM depends only on the feature vector v_x. Also, it is assumed that v_p and v_x are independent. Since cluster is assigned by the parameter vector, ϕ_k is only dependent on v_p. Then, the posterior is estimated by the following:

$$P(\theta|v_x,v_p) \approx \sum_{k=1}^{k_{Max}} w_k(v_p)o_k(\theta,v_x)N(v_p|\mu_k,\Sigma_k)Z$$

where $k_{Max} = 4$, summation index k sorted such that $P(v_p|\phi_k) \geq P(v_p|\phi_{k+1})$, o_k the output of the RVM associated with the k-th cluster and Z a normalization constant; uniform priors $P(\theta)$ and $P(\phi_k)$ were assumed.

3. Experiments

Samples of the violin were obtained from the RWC Music Database (RWC-MDB-I-2001 No. 15) [5]. Playing styles "VNPO (Ponticelli)," "VNNOM (Normal)" and "VNTA (Tasto)" were used as samples for describing the family of playing techniques *sul tasto (S.T.), ordinaire (Ord.)* and *sul ponticelli (S.P.)*, which describes position of the bow on the string in a bowed instrument, where S.P. indicates playing *closer* to the bridge than the usual, Ord. normally, and S.T. *farther* from the bridge. Each playing style consists of a 64-note chromatic scale starting at G2, the lowest possible sound generated on a violin.

After extracting the features, 5-fold Cross-Validation was performed by using every fifth frame as the validation and all other for the training. To generate a mapping from playing technique to real number that is consistent with the meaning of the playing techniques, S.T. was mapped to -1, Ord. to 0, and S.P. to 1. To determine the effectiveness of rejecting "unconfident" samples, different data rejection criteria were imposed. Namely, if the variance of the distribution σ^2 was above some threshold τ, the data was rejected as being too ambiguous. The recognition rates with different rejection criteria were then compared against 5-Nearest Neighbor (5-NN).

4. Results and Discussion

Figure 2 shows the recognition rate and the percentage of data rejected as the rejection criterion is varied. Average performance of the 5-NN yields a recognition accuracy of S.T.=87%, Ord.=93% and S.P.=96%. Our method exceeds the baseline given the rejection criterion $\sigma^2 > \tau = 0.25$, which discards 93.4% of Ord., 98.2% of S.P., and 81.8% of S.T., but recognizes the respective playing techniques at 93.8%, 99.8%, and 100.0%. While the ratio of analyzed frames to the number of total frames is drastically lowered, we have shown that imposing stringent criterion increases the recognition rate. This result also suggests that ambiguous sounds are the primary cause of misclassification in playing technique recognition. Noting that the recognition without a criterion performs considerably poorly than the baseline, the result hints that a better set of features and hyperparameters could increase the lower bound, and the rejection criterion may be relaxed in order to attain similar performance (thus increasing the effective rate of analysis).

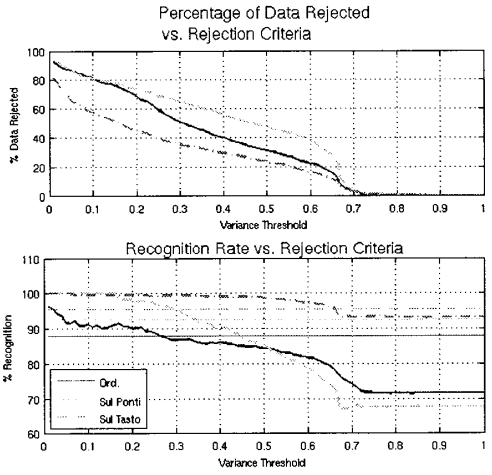

Figure 2: Recognition Rate and Data Rejection Rate as a Function of Rejection Criteria. Light horizonatal line in the bottom figure shows the recognition result from 5-NN.

5. Conclusion and Future Work

In this paper, we treated a class of playing techniques as a probability distribution to exploit statistics that could detect data that might decrease recognition accuracy. We found that rejecting data whose distribution has a large variance could bring the recognition accuracy substantially at the cost of reduced rate of recognition. Though only one type of playing technique was tested due to time constraints, it is of interest to test our method on multiple instruments and various playing techniques. Furthermore, we would like to adapt the trained machine for other instruments or performers with minimal retraining process.

6. Acknowlegements

This research is partially supported by Kakenhi (S), GCOE and CREST-Muse.

References

[1] Krishnaswamy, A. et al. "Inferring control inputs to an acoustic violin from audio spectra," ICME '03. Proc. 2003 Int'l Conf. on, vol.1, no., pp. I-733-6 vol.1, 6-9 July 2003.

[2] Kikuchi, Y. et al. "An Automatic Transcription System for Bass Guitar." Proc. 52nd National Convention of IPSJ, 5Z-7, March 1996.

[3] Jehan, T. et al. "An Audio-Driven Perceptually Meaningful Timbre Synthesizer," Proc. ICMC '01. Havana, Cuba, 2001.

[4] Tipping, M. E. et al. "Fast marginal likelihood maximisation for sparse Bayesian models." In C. M. Bishop and B. J. Frey (Eds.), Proc. of the Ninth Int'l Workshop on Artificial Intelligence and Statistics, Key West, FL, Jan 2003.

[5] Goto, M. et al. "RWC Music Database: Music Genre Database and Musical Instrument Sound Database." ISMIR '03 Proc. of 4th, pp.229-230, October 2003.

[6] Kitahara, T. et al. "Pitch-dependent identification of musical instrument sounds." Appied Intelligence, Vol.23, No.3, Springer-Verlag. pp. 267-275, December 2005.