INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

A Framework for Automated Marmoset Vocalization Detection And Classification

Alan Wisler[1], *Laura J. Brattain*[2], *Rogier Landman*[3], *Thomas F. Quatieri*[2]

[1]Arizona State University, USA
[2]MIT Lincoln Laboratory, USA
[3]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, USA
awisler@asu.edu, brattainl@ll.mit.edu, landman@mit.edu, quatieri@ll.mit.edu

Abstract

This paper describes a novel framework for automated marmoset vocalization detection and classification from within long audio streams recorded in a noisy animal room, where multiple marmosets are housed. To overcome the challenge of limited manually annotated data, we implemented a data augmentation method using only a small number of labeled vocalizations. The feature sets chosen have the desirable property of capturing characteristics of the signals that are useful in both identifying and distinguishing marmoset vocalizations. Unlike many previous methods, feature extraction, call detection, and call classification in our system are completely automated. The system maintains a good performance of 20% equal error detection rate using data with high number of noise events and 15% of classification error. Performance can be further improved with additional labeled training data. Because this extensible system is capable of identifying both positive and negative welfare indicators, it provides a powerful framework for non-human primate welfare monitoring as well as behavior assessment.

Index Terms: Automated detection and classification, marmoset vocalization, primate behavioral analysis, primate welfare monitoring, Teager energy operator

1. Introduction

The common marmoset (Callithrix jacchus) is a small new world primate that is emerging as an important non-human primate model for neuroscience research [1]–[3]. In addition to their small size, fast maturation, high fecundity, low maintenance, and genetic similarity to human [4][5], one distinctive feature of marmosets is their large repertoire of vocal behaviors, making them an attractive model for studying the origins and neural basis of human language. Vocalizations belonging to the same species, or Conspecific Vocalizations (CVs), are crucial for social interactions, reproductive success, and survival [6]. Marmosets employ their vocalizations to contact other group members, indicate submissiveness, aggressiveness, anger, fear and alert other group members to varying degrees and types of threats [7]. In spite of recent efforts to provide a quantitative acoustic analysis [8]–[10], there still remains no consensus as to the vocal repertoire of the common marmoset.

A major challenge in utilizing vocalizations for analyzing animal behavior is the time and skills required to monitor and identify vocalization production by hand. Due to the amount of training required, it is difficult to crowd source this task. The advancements in machine learning have spurred a recent push to automate vocalization monitoring in a range of mammals. Such efforts have been used to classify bird songs [11], African elephants [12], killer whales [13], and marmosets [8]. Recent work on semi-automated marmoset vocalization classification [10] is primarily based on the use of short-time spectral analysis, which requires the explicit estimation of the temporal features derived from this representation.

In this paper we introduce a novel framework for automated detection and classification of positive, negative, and neutral welfare indicators using data recorded by microphone collars on marmosets in home cage with background cage noise. The emphasis here is on a fully automated system for capturing naturalistic vocal behaviors. This is in contrast to more common approaches of recording short testing sessions with manual or semi-automated analysis.

This paper is outlined as follows: Section 2 describes the system architecture, including feature selections. Section 3 provides preliminary results achieved on a semi-synthetic dataset designed to realistically model the actual audio data. Section 4 discusses potential future expansions of the system.

2. System Layout

The proposed system architecture is divided into three main modules. Section 2.1 introduces the set of features used. Section 2.2 describes the detection procedure and Section 2.3 describes the approach for classifying a pre-defined N number of vocalizations ($N = 4$ in this case).

2.1. Features

Figure 1 shows the spectrograms of four marmoset vocalizations, which are the focus of this work. Trill is a positive welfare indicator, while phee and twitter are considered ambiguous, and chatter is considered a negative welfare indicator [14].

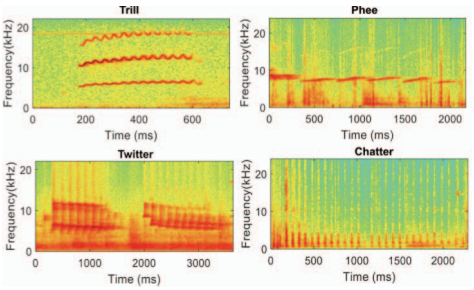

Figure 1: *Spectrograms of four marmoset vocalizations*

A wide variety of features useful in analyzing human speech and other animal vocalizations are explored in this paper. First

http://dx.doi.org/10.21437/Interspeech.2016-1410

is the basic set of six audio features described in [15][16], which measure statistics based on energy entropy, signal energy, zero crossing rate, spectral rolloff, spectral centroid, and spectral flux. This feature set is augmented with their pairwise variability, which is the mean of the absolute value of the derivatives of each feature. In this paper, all the features described above are referred to as the Audio Toolbox features. Next we extract from Mel-Frequency Cepstral Coefficients (MFCC) a feature set that includes the mean of the coefficients along with their first and second derivatives, as well as the variance, skewness, and kurtosis. Finally, in an effort to capture the rapid changes in frequency found in marmoset vocalizations such twitters and trills, we consider the Teager energy operator (TEO) [17]. The TEO has been used in a number of speech applications including automatic speech recognition [18], speech enhancement [19], voice activity detection [20], hyper-nasality detection [21], and emotion recognition [22]. More recently the TEO has been employed in the detection and classification of toothed whale vocalizations [23]–[25]. Despite the effectiveness of the TEO in vocalization analysis for marine life, its effectiveness for analyzing the vocalizations of non-human primate remains largely unexplored. In an effort to capture the temporal variations in the Teager energy over time, we compute the inverse discrete cosine transform of the power spectral density.

All of these features have the desirable property of capturing characteristics of the signal that are useful in both identifying and distinguishing marmoset vocalizations. Furthermore they can be easily extracted in an automated manner unlike the features described in more common approaches [10]. The relative importance of each of these feature sets will be discussed in Section 3.

2.2. Detection

Since the detector must make many decisions for every second of audio data provided, we select features that have low dimensionality and are computationally efficient. We use a set of TEO-based features for our detector. From the framed signals (with frame=$500\,ms$, step=$50\,ms$), we extract the signal energy, the mean Teager energy, and the peak amplitude and frequency of the power spectral density of the Teager energy. Using these features, we train a simple feed-forward neural network containing one hidden layer of 3 neurons to obtain the likelihood that each frame contains a vocalization. These likelihood predictions are then converted to binary predictions using a threshold, which controls the sensitivity of the detector.

Once each frame has been assigned as either vocalized (0) or non-vocalized (1), we merge these decisions in the following manner. Consider that each frame is a candidate vocalization. We first merge any vocalized frames with fewer than K_1 number of non-vocalized frames between them into the same candidate vocalization. This is done in order to prevent strings of vocalizations, such as those found in phees and twitters, from being considered as multiple separate vocalizations. We then reject any candidate vocalizations containing fewer than K_2 number of vocalized frames. These frames are deemed too short in duration to model the types of vocalizations that we are interested in classifying. Increasing K_1 will increase the likelihood of merging separate vocalizations, while decreasing K_1 will raise the likelihood of splitting a single vocalization into multiple predicted vocalizations. K_2 can be adjusted to control the precision and recall of the detector. Lower K_2 will lead to greater sensitivity and the ability to detect shorter duration vocalizations, but will also increase the false alarm rate.

2.3. Classification

The classification module presented here aims to classify four vocalizations (trill, phee, twitter, and chatter) and one additional category for all other acoustic events. We start with a large set of candidate features described in Section 2.1 in order to capture spectral-temporal information helpful in classifying between any pair of vocalizations. While using a large set of features maximizes the chances of identifying useful variables, directly modeling in high-dimensional spaces yields overly complex models that are prone to over fitting. To avoid this problem, we iteratively select the top 20 features using a forward selection algorithm designed to minimize the non-parametric upper bound on the Bayes error described in [26]. This approach outperformed feature selection by the parametrically estimated Bhattacharyya bound. Once the optimal subset of features has been identified, we use error-correcting output codes [27] to generate different multi-class models for standard binary learners: SVMs, naïve bayes classifiers, decision trees, and discriminant analysis. Analsyis of the performance of these different binary learners will be discussed in Section 3.3.

3. Results

A common challenge in automated animal vocalization classification is the limited labeled data. To overcome this limitation, we analyze the system performance on semi-synthetic data generated using the procedure outlined in Section 3.2. The augmented truth data greatly enhanced the system development and validation. While the training and testing data sets for the detector and classifier are generated using the same procedure, the vocalization samples selected for each process are distinct.

3.1. Experimental setup

We collected vocalizations from two adult marmoset monkeys housed together in their home cage (~1 x 1 x 2 m), which is located in a large animal room with ~10 other marmoset cages. The subjects moved freely inside their home cage. A small voice recorder (PanicTech, 8GB digital recorder, 46 x 5 x 18 mm, 6.9 g, sampling rate 48 kHz) was embedded into a soft silicone-based collar and was worn around each subject's neck. Each recording session lasted about 1 hour, after which the collars were taken off. All animal procedures were performed in accord with National Institute of Health guidelines and were approved by Massachusetts Institute of Technology Committee on Animal Care. The audio files were aligned using Audacity (http:// http://www.audacityteam.org) and further analyzed in MATLAB (Mathworks, Natick, MA).

3.2. Data Augmentation

Labeled data is essential for both the training and evaluation of the proposed model, however because the time and expertise required in this domain, it has been a challenge to obtain large number of labeled vocalizations. Data augmentation is a common approach in machine learning to overcome this constrain [28][29]. We have developed an approach, which takes a small set of sample vocalizations (call dictionary) and augment it to large datasets with background noise and other acoustic events that replicate the acoustic characteristics of a

continuous stream of labeled audio data. The call dictionary used in the experiments contains 24 phee calls, 31 trill calls, 21 twitter calls, 6 chatter calls, and 69 other acoustic events.

To generate augmented audio streams for the detector, we first replicate the background noise found throughout our sample recordings by identifying segments of audio that is free from vocalizations or other acoustic events. To create a new audio noise stream, starting at the 1st second into the file we perform the following:

1. Randomly select 1 second of noise from the sample file.
2. Multiple this noise signal by a triangular window, and add it to the current audio segment.
3. Step forward half a second.
4. Repeat steps 1-3 until reaching the end of the audio file.

The result is a continuous stream of noise of an arbitrary length that closely models that found in the real recordings. Next we populate the noise stream with vocalizations by randomly selecting vocalizations and acoustic events from the call dictionary and adding them at random indices to the background noise. The acoustic events are drawn from a set of sample events such as cage rattling noises and noise from marmosets scratching their necks, found in the original audio streams. CV placement is restricted so that no new vocalizations are placed on top of previous ones. Because samples are drawn directly from the original signal, the SNR of the original signal should be approximately maintained. Once all vocalizations have been placed the resulting audio stream is used to train the detector. Note that for evaluation we partition our call dictionary such that only part of it is used in training and the remainder is used to generate the test data

3.3. Vocalization detection results

Our detection module was tested using the semi-synthetic audio streams described in the previous section. We generate separate 10-minute segments of audio for both training and evaluation, and populate each audio segment with 10 vocalizations from each call type, along with additional acoustic events that represent non-vacal events such as cage rattling or noise from animal scratching their neck. We vary the number of acoustic events in order to better understand the influence of these events on the systems performance. We then evaluate the performance of the detector using true positive rate (TPR), which is the ratio of true positives over the sum of true positives and false negatives, and false positive rate (FPR), which is the ratio of false positives over the sum of true negatives and false positives.

The metrics are calculated by considering each frame as a separate detection problem. Figure 2 is a plot of the receiver-operator characteristics (ROC) curve resulting from each trial of this experiment. The ROC curve clearly illustrates the trade-off between detection rate and false-alarm rate, and shows the impact of acoustic events on the system performance.

3.4. Classification results

We evaulate our classification module from three perspectives: (1) performance of the different classifiers, (2) performance vs. the size of the call dictionary, and (3) feature sets that perform best in discriminating between the various call types.

To evaluate the classifiers, we generate a synthetic training and test vocalizations via the procedure outlined in Section 3.2. To analyze the dependency of the system on the size of the call dictionary, we vary the fraction of vocalizations used for training vs. testing from 20% to 50%, and then generate a total of 2000 instances (400 per vocalization) each for the training and test data. Once the training and test vocalizations are generated, we iteratively select the top 20 features using a forward selection algorithm designed to minimize the non-parametric upper bound on the Bayes error described in [26]. We then use error-correcting output codes [27] to generate different multi-class models for standard binary learners including SVMs, naïve bayes classifiers, decision trees, and

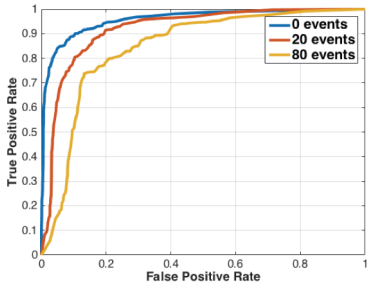

Figure 2: *Detection/false alarm tradeoffs with increasing number of noise events.*

discriminant analysis. We evaluate the performance of each of these classifiers on the test data for each partition of the call dictionary at every feature subset. These results are then averaged across a 25 iteration Monte Carlo simulation, and the average and standard error of the classification error rates are displayed in Figure 3. Though we tested smaller feature subsets, we observed the performance of most classifiers asymptote to the optimal performance by 20 features, thus we present only the results of classifiers constructed on 20 features.

From Figure 3, we see that the performance of the classifier is dependant upon the size of the call dictionary.

Due to the dramatic improvements in performance at each

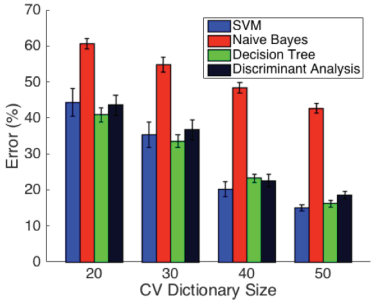

Figure 3: *Comparison of the classification errors (%) from four different methods given different CV dictionary sizes. Error bars are standard errors.*

increment of dictionary sizes tested, we hypothesize that the performance with respect to the dictioanry size is not close to asymptote, however we are unable to test this hypothesis at any larger sizes as attributing any more than 50% of the CV dictionary impairs our ability to estimate the out-of-sample performance of each classifier. Additionally, while none of the binary learners showed a statistically significant advantage over other classifiers, we found that the decision trees performed best for smaller dictionary sizes (20% and 30%), while the SVM learner yielded the highest performance for larger dictionaries (40% and 50%).

To better understand the cause for these errors, we look at the confusion matrix in Table 1, which is drawn from a single trial of this classification experiment. This matrix shows that

the majority of the mistakes made by the proposed model come from confusion between twitters and chatters and confusion between chatters and other events. Because both twitters and chatters are calls containing periodic bursts of energy, the confusion between them is not surpising and indicates a need for features that better capture the short-term spectral structure in the twitter. Confusion between the chatters and other acoustic events likely stems from the difficulty in distinguishing chatters from the noise resulted from the marmosets scratching their collars, as the two are similar. This dificulty can be alleviated by the integration of data from additional microphones located outside of the cage. Increasing the number of chatters in the call dictionary could also result in a more robust representation of them.

Table 1. Confusion Matrix

True\Predicted	Phee	Trill	Twitter	Chatter	Other
Phee	367	4	29	0	0
Trill	6	385	5	0	4
Twitter	3	1	337	47	12
Chatter	0	0	0	289	111
Other	10	6	7	25	352

To better understand the relative significance of each grouping of features, a second experiment is conducted where the feature set is limited to specific group of features (Figure 4). This experiment is identical to the previous one with a few exceptions. The size of the training dictionary is held constant at 50% and we instead vary the base feature set. Only 5 or 10 features are selected rather than 20, because the Audio Toolbox only contains 10 features total. We find from this experiment that the features from the Audio Toolbox yield the

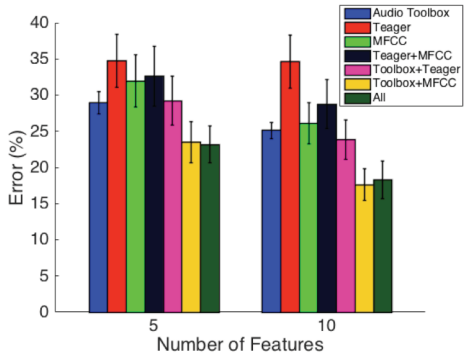

Figure 4: *Performance comparison of individual feature sets. Error bars are standard errors.*

highest individual performance among the 3 feature sets, though they only slightly outperform the MFCC grouping. When we look at combinations of feature sets, we find that the performance of the Audio Toolbox and MFCC features significantly improves when grouped together, and while the Teager features don't improve the performance when added to either of the other sets, they yield a small boost when added to their combination.

4. Discussion

This paper represents the preliminary effort in developing a system to automatically process noisy audio stream for marmoset vocal behavior classification. We focus primarily on evaluating and tuning the classification model, since it can be used to mitigate deficiencies in the detection system by operating the detector in a high detection region and using the classifier to filter out the large number of false positives. While the proposed system exhibits relatively high performance in our evaluations thus far, there remains significant work in refining the design of the proposed model. Many aspects may be improved with the availability of additional data, which will allow the use of more sophisticated models for both the detection and classification.

Furthermore, while the spectral plots based on the Teager energy shown in Figure 5 provide a representation that is visually distinctive for each vocalization type, the features extracted based on this representation have not positively influenced performance with significance in our evaluations thus far. Further research is necessary for more effective use of TEO in this domain.

It is also worth noting that we only consider four categories of vocalizations in this paper, which represents a small subset of the marmoset's entire vocal repertoire. Since the architecture is modular, we can easily extend the system to include a broader set of discrete vocalizations. More work is also needed for compound and overlapping calls.

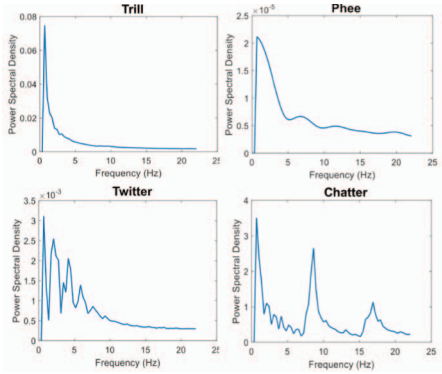

Figure 5: *Power spectral density of the Teager energy extracted from the four vocalizations shown in Figure 1.*

5. Conclusions

This paper presents a novel framework for automated marmoset vocalization detection and classification. Three major components are described: automated feature extraction from audio data collected in home cage, the detection module for identifying vocalizations from noisy background, and the classification module for discriminating between four different vocalizations. The proposed system performs well experimentally with 20% of equal error rate in detection on data with high number of noise events and 15% of classification error. The architecture is flexible and can be extended to a larger number of vocalizations. We believe that such an automated system has the potential to greatly improve primate welfare monitoring and behavioral analysis.

6. Acknowledgements

This material is based upon work supported by the Assistant Secretary of Defense for Research and Engineering under Air Force Contract No. FA8721-05-C-0002 and/or FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Assistant Secretary of Defense for Research and Engineering. It is also sponsored by the MIT McGovern Institute Neurotechnology Program.

7. References

[1] J. F. Mitchell, J. H. Reynolds, and C. T. Miller, "Active Vision in Marmosets: A Model System for Visual Neuroscience," *J. Neurosci.*, vol. 34, no. 4, pp. 1183–1194, Jan. 2014.

[2] N. Kishi, K. Sato, E. Sasaki, and H. Okano, "Common marmoset as a new model animal for neuroscience research and genome editing technology," *Dev. Growth Differ.*, vol. 56, no. 1, pp. 53–62, 2014.

[3] E. Sasaki, "Prospects for genetically modified non-human primate models, including the common marmoset," *Neurosci. Res.*, vol. 93, pp. 110–115, Apr. 2015.

[4] D. H. Abbott, D. K. Barnett, R. J. Colman, M. E. Yamamoto, and N. J. Schultz-Darken, "Aspects of common marmoset basic biology and life history important for biomedical research," *Comp. Med.*, vol. 53, no. 4, pp. 339–350, 2003.

[5] J. Heam, "Reproduction in new world primates."

[6] X. Wang, M. M. Merzenich, R. Beitel, and C. E. Schreiner, "Representation of a species-specific vocalization in the primary auditory cortex of the common marmoset: temporal and spectral characteristics," *J. Neurophysiol.*, vol. 74, no. 6, pp. 2685–2706, 1995.

[7] G. Epple, "Comparative Studies on Vocalization in Marmoset Monkeys," *Folia Primatol. (Basel)*, vol. 8, no. 1, pp. 1–40, 1968.

[8] C.-J. Chang, "Automated classification of marmoset vocalizations and their representations in the auditory cortex," 2014.

[9] X. Wang, "The harmonic organization of auditory cortex," *Front. Syst. Neurosci.*, vol. 7, 2013.

[10] J. A. Agamaite, C.-J. Chang, M. S. Osmanski, and X. Wang, "A quantitative acoustic analysis of the vocal repertoire of the common marmoset (Callithrix jacchus)," *J. Acoust. Soc. Am.*, vol. 138, no. 5, pp. 2906–2928, 2015.

[11] S. E. Anderson, A. S. Dave, and D. Margoliash, "Template-based automatic recognition of birdsong syllables from continuous recordings," vol. 100, no. 2, pp. 1209–1219, 1996.

[12] P. J. Clemins, M. T. Johnson, K. Leong, and A. Savage, "Automatic Classification and Speaker Identification of African Elephant (Loxodonta africana) Vocalizations," vol. 117, no. 2, pp. 956–963, 2005.

[13] J. C. Brown, "Automatic classification of killer whale vocalizations using," no. August, pp. 1201–1207, 2007.

[14] "Common Marmoset Care - Understanding Behaviour - Calls." [Online]. Available: http://www.marmosetcare.com/understanding-behaviour/calls.html. [Accessed: 22-Jun-2016].

[15] S. Theodoridis and K. Koutroumbas, "Pattern recognition," 2003.

[16] T. Giannakopoulos, D. Kosmopoulos, A. Aristidou, and S. Theodoridis, "Violence content classification using audio features," in *Advances in Artificial Intelligence*, Springer, 2006, pp. 502–507.

[17] H. M. Teager, "Some Observations on Oral Air Flow During Phonation," no. 5, pp. 599–601, 1980.

[18] D. Dimitriadis, P. Maragos, and A. Potamianos, "Auditory Teager Energy Cepstrum Coefficients for Robust Speech Recognition," in *INTERSPEECH*, 2005, pp. 3013–3016.

[19] M. Bahoura and J. Rouat, "Wavelet Speech Enhancement based on the Teager Energy Operator," *Signal Process. Lett. IEEE*, vol. 8, no. 1, pp. 10–12, 2001.

[20] B. Wu and K. Wang, "Voice Activity Detection Based on Auto-Correlation Function Using Wavelet Transform and Teager Energy Operator," vol. 11, no. 1, pp. 87–100, 2006.

[21] D. A. Cairns, J. H. L. Hansen, and J. F. Kaiser, "Recent Advances in Hypernasal Speech Detection using the Nonlinear Teager Energy Operator," in *ICSLP*, 1996, vol. 2.

[22] D. Ververidis and C. Kotropoulos, "Emotional speech recognition : Resources , features , and methods," vol. 48, pp. 1162–1181, 2006.

[23] V. Kandia and Y. Stylianou, "Detection of sperm whale clicks based on the Teager–Kaiser energy operator," *Appl. Acoust.*, vol. 67, no. 11–12, pp. 1144–1163, Nov. 2006.

[24] M. A. Roch, A. Širović, and S. Baumann-Pickering, "Detection, Classification, and Localization of Cetaceans by groups at the Scripps Institution of Oceanography and San Diego State University (2003-2013)."

[25] M. A. Roch, H. Klinck, S. Baumann-Pickering, D. K. Mellinger, S. Qui, M. S. Soldevilla, and J. A. Hildebrand, "Classification of echolocation clicks from odontocetes in the Southern California Bight," *J. Acoust. Soc. Am.*, vol. 129, no. 1, pp. 467–475, 2011.

[26] V. Berisha, A. Wisler, A. O. Hero III, and A. Spanias, "Empirically estimable classification bounds based on a nonparametric divergence measure," *Signal Process. IEEE Trans. On*, vol. 64, no. 3, pp. 580–591, 2016.

[27] T. G. Dietterich and G. Bakiri, "Solving multiclass learning problems via error-correcting output codes," *J. Artif. Intell. Res.*, pp. 263–286, 1995.

[28] D. A. Van Dyk and X.-L. Meng, "The art of data augmentation," *J. Comput. Graph. Stat.*, 2012.

[29] N. G. Polson and S. L. Scott, "Data augmentation for support vector machines," *Bayesian Anal.*, vol. 6, no. 1, pp. 1–23, Mar. 2011.