Non-Linguistic Vocal Event Detection Using Online Random

Abou-Zleikha, Mohamed; Tan, Zheng-Hua; Christensen, Mads Græsbøll; Jensen, Søren Holdt

Published in:
Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014 37th International Convention on

DOI (link to publication from Publisher):
10.1109/MIPRO.2014.6859773

Publication date:
2014

Document Version
Early version, also known as pre-print

Non-Linguistic Vocal Event Detection Using Online Random Forest

M. Abou-Zleikha*, Z.G. Tan *, M. G. Christensen** and S. H. Jensen*
* Department of Electronic Systems, Aalborg University, Denmark
** Audio Analysis Lab, AD:MT, Aalborg University, Denmark
moa,zt,shj@es.aau.dk, mgc@create.aau.dk

Abstract - Accurate detection of non-linguistic vocal events in social signals can have a great impact on the applicability of speech enabled interactive systems. In this paper, we investigate the use of random forest for vocal event detection. Random forest technique has been successfully employed in many areas such as object detection, face recognition, and audio event detection. This paper proposes to use online random forest technique for detecting laughter and filler and for analyzing the importance of various features for non-linguistic vocal event classification through permutation. The results show that ac- cording to the Area Under Curve measure the online random forest achieved 88.1% compared to 82.9% obtained by the baseline support vector machines for laughter classification and 86.8% to 83.6% for filler classification.

I. INTRODUCTION

Social signals processing, a very important aspect of human computer interaction applications, is a challenging task and an essential step towards more efficient interaction with ma- chines. Facial expressions, body postures, gestures and vocal events are all social cues that can be generated during social interactions and contain information about the adjustment of relations and interactions between counterparts [1]. The non-linguistic vocal events are considered one of the main clues that reflect the social interactions and emotions. The impact of vocal events on the meaning of recognized speech has proven to be very significant in HCI systems [2]. Fillers, laughters and silence are some of these events that have a considerable effect on the meaning of the transmitted message as they provide feedback and reflect the level of engagement. Moreover, the detection of non-linguistic vocalization promises a significance improvement in the performance of spontaneous speech recognition systems.

Several studies in the literature have tackled the vocal social event detection problem by investigating various ways of detection. Weninger et. al. [3] used bidirectional short-term memory recurrent neural network combined with a set of features extracted using non-negative matrix factorization for the detection of non-linguistic events such as laughter, filler and breathing. Bayesian information criterion was used for acoustic event detection in podcast [4]. Feed-forward neural net- work were used for laughter detection by Sumi et. al. [4], where the MFCC and pitch features were utilized for training the network. In [5] the frequency analysis was used to detect the filler in speech recognition systems. Automatic Language Independent Speech Processing was also investigated for the generation of acoustic segments and the extraction of acoustic events sequence [6]. Maximum likelihood linear regression and maximum a posterior techniques were then used to detect the laughter [6]. Gaussian mixture models technique was used for laughter detection in [7].

With regard to classifiers random forest (RF) [8] has been successfully employed in several classification tasks in audio domain such as emotion recognition [9, 10, 11] and audio event detection [12].

The aim of this work is to investigate the use of random forest approach for non-linguistic vocal event detection. We present an online RF method with a two phases data balancing approach for this task. We further employ this method as an analysis tool to study the importance of each feature used on the detection accuracy of the chosen events. This is, to the best of the authors' knowledge, the first attempt to use technique as such for non-linguistic vocal event detection. The paper is organized as follows: Section 2 briefly explains the random forest approach and the importance of its variables. Section 3 presents the proposed methodology to detect non- linguistic vocal events, while in Section 4, the evaluation of the proposed method on a paralinguistic challenge corpus is presented.

II. RANDOM FOREST

Random forest is a tree-based non-parametric classification and regression approach. The principle is to grow an assemble of trees on a random selection of samples in a training set. Each tree is a non-pruned classification and regression tree (CART). While constructing the trees, at each tree node, a set of randomly selected features is considered and these features are investigated as a potential predictor that decide the split of the data in the tree. For classification, each tree predicts the target class individually, and the forest predicts the final target class as a majority vote of the individual tree predictions.

Compared to a single decision tree, the random forest assembles several trees that are trained using a randomly selected subset of the data and it usually achieves a higher degree of generalization, stability and accuracy. In this work, the random forest model is used for classifying the frames of the signals into different events.

Formally, a random forest is a set of decision trees:

$$RF=t_1,t_2\ldots t_{ntree} \quad (1)$$

where t_i is the i^{th} individual tree and *ntree* is the number of trees. Suppose D is training data consisting of a set of samples (frames in our the current implementation):

$$D=f_n \quad (2)$$

for $n = 1,..,N$, where each f_n is a frame. In CART, the purity measure maximization is used to build the tree. In a random forest, each tree is grown until its leaves contain one label (100% pure is employed as a purity threshold) or the tree reaches to a maximum depth as in online random forest [13].

A. Variable Importance

The importance of a variable (i.e., feature) is derived from the variable contribution in the classification performance. When changing the value of an important variable with a random value, the classification performance should decrease. On the other hand, if a variable is not important for the classification task, the random replacement of its values should have an unnoticeable effect on the classification accuracy. The random forest algorithm estimates the importance of a variable by looking at the changes in the prediction error when the values of that variable are permuted while all others are left unchanged.

III. VOCAL SOCIAL EVENT DETECTION USING RANDOM FOREST

Suppose we want to classify the three classes: (G)arbage, (F)iller and (L)aughter. The numbers of frames for the classes are x, y, and z respectively, where $x \gg y, z$. Suppose training data $D = fn : n = 1,..,N$,where N is huge. When classifying the data, we need to address two problems:

the unbalancing of the data.

the huge number of samples in the data.

The first problem appears due to the unbalance between the number of frames of each of the social signal events in the corpus and the number of garbage frames. Therefore, a data balancing operation is required. Data unbalancing has been a research question for several studies and several solutions have been proposed such as down-sampling and up-sampling [14]. The down-sampling of the majority class may result an information lose, due to removing a large part of the majority class [14]. The up-sampling is considered a better solution. However this will result in an increase in the data size as well as increase in the training time specially when the dataset size is already huge.

The Balanced Random Forest (BRF) has been proposed as a better alternative for dealing with imbalanced data [14]. For each tree in a random forest, we randomly down-sample the majority class to the same size as the minority class, and use the down-sampled data for training. A more complex approach is required, however, when we are dealing with a multi-classes classification task, where we have several classes with different numbers of samples. Another alternative approach is to use the Weighted Random Forest (WRF) [14], by assigning a higher weight for the minority class samples. Although this approach has been implemented and achieved good results, it faces several problems including its sensitivity to noise.

In the vocal event detection task, the data is huge and could not be read and processed as one big batch and there is one majority class and two minority classes. In addition, the data is read utterance by utterance (call it sub-batch), and these sub-batches contain different numbers of samples for each class, so it is not feasible to select the number of samples that is equal to that of the smallest class. Furthermore, there are two levels of unbalancing: The first is between the majority class and the two minority classes, and the second is between the two minority classes. As a result, it is not possible to balance the data using the BRF only.

Consequently, a combination of the BRF and the WRF methods is proposed. The main idea is that each tree in the forest should be trained on a subset of data and the samples in this subset should be balanced. Specifically, after the subset of data is chosen for training each tree, we select the samples so that each resulted sub-dataset contains a balanced number of samples. Afterwards, WRF is applied.

The data balancing process is done as follows:

For the frames of each utterance i:

- x'_i= the number of frames of the minority classes (filler + laughter).
- randomly select x'_i frames of the garbage class.

As a result of this operation, the number of garbage class frames for training a each tree (x') will be equal to the number of filler frames (y) plus the number of laughter frames (z) ($x' = y+z$). The reason for choosing ($x' = y+z$) is that the distributions of the minority classes in each utterance are unknown and we only know the total number of samples for each class. By selecting $x'_i = y_i + z_i$ at each utterance i, it is guaranteed that by the end of the process, the forest is trained on the maximum number of samples from the majority class possible, taking into account that all of the samples in the minority classes are used for training.

In this case, we reduce the unbalancing problem, and the numbers of training samples for each tree in the forest will be $x' = y+z, y, z$ for classes G, F and L respectively. To make data for the classes fully balanced, the WRF approach is used. The weight of class F samples will be x'/y, the weight of class L samples will be x'/z and the weight of class G samples will be $x'/x' = 1$. In this approach, it is not necessary to know the content of each sub-batch in advance in order to make the data balanced; but it is only required to know the total number of samples from each class in the entire training data (or an approximation of its). As a result, it is possible to train each tree in the random forest using a balanced data without removing many training samples from the majority class since each tree in

the forest is trained using a randomly selected samples from the majority class, which increases the possibility of covering the majority class data.

To solve the second problem of huge data, an online random forest is used instead of the traditional (one-batch learning) approach [13]. This approach showed efficiency in using the memory and the processing time with very comparable results with the one-batch approach.

The trained forest is then used for event detection. Furthermore, a post-processing smoothing operation has been performed to improve the continuity of the detected events. The smoothing is performed by taking the average of a 13 frames window (6 before and 6 after). This operation is applied on the confident membership of the frame for each class. Then a scaling operation is applied on the confident member- ship of the frame to make the sum of them equal to one.

IV. Model Evaluation

A. Data Description

The SSPNet Vocalization Corpus (SVC) [15] is used in this work for the model evaluation. The corpus is composed of 2763 audio clips extracted from a collection of 60 phone calls collected from 120 participants (63 female, 57 male). The audio clips are annotated according to laughter and fillers with 11 seconds length each. Table 1 presents the distribution of the data over training, development and testing sets. More information about data description could be found in [15].

The data presented in Table 1 clearly shows the unbalance in the data, where the percentages of the laughter, filler and garbage classes in the training dataset are 0.37%, 0.53%, and 99.10%, respectively. It is also noticed that the filler class is about 30% larger than the laughter class.

The features used for classification are:

• MFCCs 1-12 and logarithmic energy + their first and second order delta

• voicing probability + first order delta

• harmonics-to-noise ratio (HNR) + first order delta

• f0 + first order delta

• zero-crossing rate (zcr) + first order delta

• the arithmetic mean and standard deviation across the frame itself and eight of its neighboring frames (four from each side) for each of the listed above features.

As a result, 140 features per frame are used for classification.

B. Online Random Forest Configuration

A java version of the online random forest [13] is implemented and used. Several experiments with different configurations have been explored. In this paper we focus on the configuration that gives the best classification accuracy.

TABLE I. The Content Of The Corpus Used For Training And Testing The Model.

	Database content		
	Training	***Development***	***Testing***
	Utterances		
	1583	500	680
	Segments		
Laughter	649	225	284
Filler	1710	556	722
	Frames		
Laughter	59294	25750	23994
Filler	85034	29432	35459
Garbage	15914421	492607	684937

The maximum depth of forest trees is fixed to 141, and the number of randomly examined variables per node is *number of features /3 = 47*. The number of frames seen before 3 splitting each node is 300 frames, and the forest consists of 100 trees.

C. Experiment and Results

TABLE II. The AUC and UAAUC values for the baseline SVM and the proposed random forest approach using the development and testing sets

Development Set			
	SVM	Random Forest	Smoothed Random Forest
Laughter AUC	86.2%	**88.2**%	**90.4**%
Filler AUC	89.0%	**89.6**%	**90.7**%
UAAUC	87.6%	**88.9**%	**90.6**%
Test Set			
	SVM	Random Forest	Smoothed Random Forest
Laughter AUC	82.9%	**86.2**%	**88.1**%
Filler AUC	83.6%	**85.8**%	**86.8**%
UAAUC	83.3%	**86.0**%	**87.5**%

After constructing the random forest using the training set, an evaluation is performed using the development and test sets. For evaluation purposes, the unweighted average recall (UAR) and the Area Under the Curve measure (AUC) for the laughter and filler classes on frame level (100 frames per second) are used as main performance measures. In addition, the unweighted average of AUC (UAAUC) is used.

Note that the test labels are not available to the authors since the corpus is the official data for INTERSPEECH

2013 Computational Paralinguistics Challenge. The accuracy results of the model were calculated by the challenge website based on the class probability for each frame that we submitted.

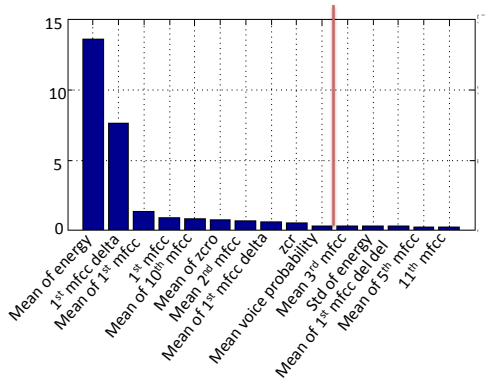

Fig. 1. *The 15 most important variables for laughter classification. The read line is after the* 10^{th} *most important variable.*

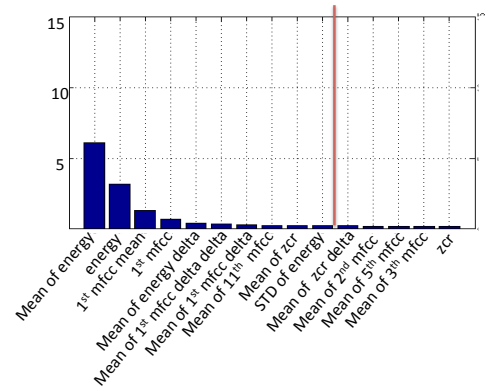

Fig. 2. *The 15 most important variables for filler classification. The read line is after the* 10^{th} *most important variable.*

Table 2 represents the AUC and UAAUC results obtained from baseline support vector machine (SVM) and the pro- posed model using the development and test sets with and without smoothing.

Table 2 shows that the random forest model has outperformed the baseline SVM model for both the development and testing sets using the AUC and UAAUC measures. The UAAUC performance difference between the proposed model and baseline one is 1.3% without post-processing smoothing and 2.0% with post-processing smoothing using development set, and 2.7% without post-processing smoothing and 4.2% with post-processing smoothing using test set. The best performance is achieved using the smoothing operation. This is due to the remove of the small discontinuity errors that might occur during the frame classification.

In the next section, the feature importance analysis for the non-linguistic event detection is explained and the experiments that assert the use the feature importance as a feature selection mechanism are presented.

TABLE III. THE AUC AND UAAUC VALUES FOR THE BASELINE SVM AND THE PROPOSED RANDOM FOREST APPROACH USING THE WHOLE FEATURE SET AND USING THE MOST IMPORTANT 10 FEATURES FOR THE DEVELOPMENT AND TESTING SETS

Development Set		
	Random Forest	Feature Selected Random Forest
Laughter AUC	88.2%	88.3%
Filler AUC	89.6%	89.0%
UAAUC	88.9%	88.7%
UAC	75.2%	74.8%
	Smoothed Random Forest	Smoothed Feature Selected Random Forest
Laughter AUC	90.4%	90.0%
Filler AUC	90.7%	90.7%
UAAUC	90.6%	90.4%
UAC	77.4%	76.9%
Test Set		
	Random Forest	Feature Selected Random Forest
Laughter AUC	86.2%	86.7%
Filler AUC	85.8%	84.7%
UAAUC	86.0%	85.7%

D. *Feature Importance Analysis*

To study the importance of each feature used in the classification, the variable importance is calculated. The *AUC* is used as an accuracy measure as proposed in [16]. After permuting the values of a feature f_i, the *AUC* is calculated ($AUC_p(f_i)$) for each tree in the forest. Then, the distance between this *AUC* and the *AUC* of the model without permutation AUC_{fi} is calculated. Finally, the average distance over all trees is calculated as:

$$VI(f_i) = \frac{1}{ntree} \sum_{j=1}^{ntree} (AUC_{j,f_i} - AUC_{j,p(f_i)}) \qquad (3)$$

where *ntree* is the number of trees in the forest. The value of $VI(f_i)$ represents the feature importance.

The feature importance analysis approach is applied to discover the important features for classifying each of laughter and filler classes only since they are the important events to detect. Figures 1 and 2 illustrate the 15 most important features for these two classes, respectively. The x−axis is the features ordered by importance and y−axis is the importance of the feature calculated using equation 3. The figures show that the energy, zcr, low-order mfccs, and their mean values across the frame and its neighboring frames have a consider- able effect on the classification accuracy.

To further justify the variable importance, the proposed random forest was rebuilt using only the union of the 10 most important features from each class (as a result

15 features are used since 5 features are common between the two classes), and the UAR, AUC and UAAUC are calculated for the new forest. Table 3 presents the results obtained using the development and the testing sets.

The results presented in Table 3 show that the model built using only the 10 most important features for each class generate a very comparable classification accuracies with the model that uses the full set of features, where the AUC difference between the results obtained from model built using the whole feature set and the selected features are 0.0% and 1.1% for development and testing sets respectively. Moreover, constructing models based on a smaller subset of features results in less complex models that are easier to analyze and this also has a positive impact on processing time. These comparable results between the two models show that we gain simpler and faster to built models without losing a lot at the accuracy level. However, the effect of combining the important features for a certain class with the important features of the other class requires more investigation.

V. CONLCUSION

This paper presented the use of the random forest approach for non-linguistic vocal event detection task. In order to employ the random forest for classifying non-linguistic vocal events and to deal with large corpus, two main problems re- quire special attention: the first one is the unbalance of the data used for training, the second issue relates to the huge size of audio data.

To solve the first problem, a data balancing approach was proposed and implemented. The methodology allows efficient use of most of the data samples while preserving the data balance for generating the trees. To deal with a corpus of a large size, an online RF approach was employed that allows efficient use of memory resources.

The model proposed was tested using Social Signals Sub-Challenge in the INTERSPEECH 2013 Computational Paralinguistics Challenge corpus. The proposed model was com- pared with a baseline SVM model using the UAAUC. The results obtained show that the proposed model outperforms the baseline model in absolute value by 2.7% without post- processing and by 4.2% with post-processing. In addition, a variable importance analysis was performed. The results obtained show that it is possible to identify most important features for classification using the random forest and the AUC measurement, and by using only these most important features (15 out of 140 in the experiment) it is possible to rebuild the random forest with a very comparable result to the one using the full feature set. This feature selection process improves the computational time of the model building stage.

ACKNOWLEDGMENT

This work was supported in part by the Danish Council for Strategic Research of the Danish Agency for Science Technology and In- novation under the CoSound project, case number 11-115328. This publication only reflects the authors' views.

REFERENCES

[1] M. Pantic, R. Cowie, F. D'Errico, D. Heylen, M. Mehu, C. Pelachaud, I. Poggi, M. Schroeder, and A. Vinciarelli, "Social signal processing: The research agenda," Visual Analysis of Humans, pp. 511–538, 2011.

[2] D. Crystal, Prosodic systems and intonation in English. Cambridge University Press, 1976.

[3] F. Weninger, B. Schuller, M. Wollmer, and G. Rigoll, "Localization of non-linguistic events in spontaneous speech by non-negative matrix factorization and long short-term memory," in In Proceedings of the IEEE In- ternational Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2011, pp. 5840–5843.

[4] K. Sumi, T. Kawahara, J. Ogata, and M. Goto, "Acous- tic event detection for spotting hot spots in podcasts," in In Proceedings of the 10th Annual Conference of the In- ternational Speech Communication Association (Inter- speech 2009). Brighton, UK, 2009, p. 11431146.

[5] M. Goto, K. Itou, and S. Hayamizu, "A real-time filled pause detection system for spontaneous speech recog- nition," in *In Proceedings of the 6th European Confer- ence on Speech Communication and Technology (Eu- rospeech'99), Budapest, Hungary*, 1999, p. 11431146.

[6] S. Pammi, H. Khemiri, and G. Chollet, "Detection of nonlinguistic vocalization using alisp sequencing," in *In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.

[7] K. P. Truong, "Measuring affective and social signals in vocal interaction," *In Proceedings of the Measuring behaviour 2010,Eindhoven, Netherlands*, vol. 27, 2010.

[8] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[9] B. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous *et al.*, "The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals," in *In Proceed- ings of the 8th Annual Conference of the International Speech Communication Association (Interspeech 2007). Antwerp, Belgium*, 2007.

[10] S.Casale,A.Russo,G.Scebba,andS.Serrano,"Speech emotion classification using machine learning algo- rithms," in *In Proceedings of the IEEE International Conference on Semantic Computing, 2008*. IEEE, 2008, pp. 158–165.

[11] J. Rong, G. Li, and Y.-P. P. Chen, "Acoustic feature se- lection for automatic emotion recognition from speech," *Information processing & management*, vol. 45, no. 3, p. 315328, 2009.

[12] A.Kumar,P.Dighe,R.Singh,S.Chaudhuri,andB.Raj, "Audio event detection from acoustic unit occurrence patterns," in *In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 489–492.

[13] A. Saffari, C. Leistner, J. Santner, M. Godec, and H. Bischof, "On-line random forests," in *In Proceedings of the 3rd IEEE ICCV Workshop on On-line Computer Vision*, 2009.

[14] C. Chen, A. Liaw, and L. Breiman, "Using random for- est to learn imbalanced data," *Technical Report, Univer- sity of California, Berkeley*, 2004.

[15] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The inter- speech 2013 computational paralinguistics challenge: Social signals, confict, emotion, autism," in In Proceed- ings of the 14th Annual Conference of the International Speech Communication Association (Interspeech 2013). Lyon, France, 2013.

[16] S. Janitza, C. Strobl, and A.-L. Boulesteix, An AUC- based Permutation Variable Importance Measure for Random Forests. Technical Report, University of Munich, 2012.