Random Forests for Laughter Detection

Heysem Kaya, Ali Mehdi Erçetin, Albert Ali Salah, Sadık Fikret Gürgen

Department of Computer Engineering, Bogazici University, Istanbul, Turkey
{heysem,mehdi.ercetin,salah,gurgen}@boun.edu.tr

Abstract

In this study, we investigate several methods on the Interspeech 2013 Paralinguistic Challenge - Social Signals Sub-Challenge dataset. The task of this sub-challenge is to detect laughter and fillers per frame. We apply Random Forests with varying number of trees and randomly selected features. We then proceed with minimum Redundancy Maximum Relevance (mRMR) ranking of features. We employ SVM with linear kernel to form a relative baseline for comparability to baseline provided in the challenge paper. The results indicate the relative superiority of Random Forests to SVMs in terms of sub-challenge performance measure, namely UAAUC. We also observe that using mRMR based feature selection, it is possible to reduce the number of features to half with negligible loss of performance. Furthermore, the performance loss due to feature reduction is found to be less in Random Forests compared to SVMs. We also make use of neighboring frames to smooth the posteriors. On the overall, we attain an increase of 5.1% (absolute) in UAAUC in challenge test set.

Index Terms: Laughter/Filler Detection, Interspeech 2013 Social Signals Sub-Challenge, Social Signal Processing, Human-Computer Interaction, Random Forest Classifier

1. Introduction

The maturity of the state-of-the art in audio-visual recognition technology lets the scientists focus their attention on other aspects such as emotional and social cues. The best results are obtained with decision fusion of models trained with different modalities e.g. speech and vision [1-7].

In the Social Signal Sub-Challenge of Interspeech 2013 [8], the detection of non-linguistic events (i.e. laughter and fillers of speaker) is aimed. The fillers contain vocalizations when the speaker tries to hold floor, such as “ah”, “uhm”, “eh”. In speech, laughter is stronger representation of feelings like amusement, embarrassment, and joy. So even in a speaker dependent setting utterances associated with different emotions might reflect diverse values of acoustic features, such as fundamental frequency (F0 a.k.a. pitch) and energy.

The challenge requires the detection of laughter and fillers as well as their discrimination against the 'garbage' class by frame leading to a localization task. The frames are of 10ms length and non-overlapping (see [8] for details). One challenging problem is the dominance of the garbage class instances, as most of the speech consists of proper conversation and silence. The number of instances to classify poses yet another challenge where the incurred memory/computation problem must be handled by elaborate feature and algorithm selection.

For two reasons, we opt to use the TUM baseline feature set provided along with the challenge [8]. First, the provided low level descriptors such as Mel Frequency Cepstral Coefficients (MFCC), F0, Harmonics to Noise Ratio (HNR), and their derivatives are known to be appropriate for these tasks [2,4,9,10]. Second, a standard baseline feature set allows repeatability and comparability.

The classifier under interest, Random Forest (RF), introduces two sources of randomness: one in feature space and the other in sample space making use of an established machine learning notion, namely ensemble learning. RFs are powerful tools to divide-and-conquer large datasets where every tree imposes an information based ranking of its random feature set. Moreover, when tree learning is parallelized, the training takes much less time than any other global learning algorithm.

Layout of the paper is as follows: in Section 2 we briefly review literature. In Section 3, we introduce the corpus and the challenge features; then in section 4, we provide the methodology. Finally, section 5 concludes the study.

2. Literature Review

In this section, we review some of the literature related to the problem of laughter detection.

In Ito et al. [2], Gaussian Mixture Models (GMM) are used for audio features (MFCC and delta-MFCC) and a linear discriminant for visual (facial) features. Multi-modal fusion is carried out with AND operator of single-modal decisions. Despite its simplicity, the fusion method is found to increase precision.

Thruong and Leeuven extracted PLP and prosody features at frame level; pitch, voicing and modulation spectrum utterance level [3]. GMM and SVM were used as base classifiers and the decisions are fused by SUM rule, MLP or SVM. Results indicate the fusion of classifiers with diverse algorithms (i.e. SVM and GMM) outperform fusion of classifiers of the same algorithm both with different features.

Knox and Mirghafori carried out a single-modal study for frame-by-frame laughter detection [4]. They utilized MFCC and prosody features (pitch and energy) along with first and second order derivatives. A Neural Network (NN) is trained with Low Level Descriptors (LLD) of 75-frame context window. Every LLD is coupled with delta features to be trained in an independent NN. Then the class probabilities of LLD-wise NNs are stacked to another NN. The study reaches a state-of-the-art on Bmr subset of ICSI Meeting Recorder corpus at the time. A notable finding is that alone delta-MFCC performed better than alone raw MFCC. Highest normalized cross correlation value was found the second best LLD after MFCC.

In two of their studies, Petridis and Pantic investigate decision level fusion of audio and video features with a simple sum rule and stacking to a NN [5,6]. The visual cues are found to be superior to PLP, Pitch and energy based audio features. As expected, multi-modal fusion is found to outperform single-modal model. Moreover, the best results

are obtained when single-modal decisions are fused with NN (i.e. stacked).

In their 2009 survey, Vinciarelli at al. highlight the novelty of the research area since it requires collaboration of several disciplines such as sociology, psychology and computer engineering as well as several study branches within machine learning and signal processing [7]. Also the study points out the multi-modality of natural understanding of social signals which should be reflected on automatic detection. Multi-modal fusion, or even single-modal classifier fusion systems are suggested, as the combination of diverse classifiers are known to be more robust and accurate.

To our knowledge, Random Forests, which have inherent classifier combination, are not employed in laughter detection.

3. Social Signals Sub-Challenge: Corpus and Features

The INTERSPEECH 2013 Social Signals Sub-Challenge is held on the "SSPNet Vocalization Corpus" (SVC). To make the reader more familiar with the corpus, we review the corpus properties and features extracted as all the details can be found in Schuller et al. [8]. Table 1 summarizes some statistics of the SVC corpus: In total there are 2763 clips each with 11 seconds length, 1.2 k laughter events and 3.0 k filler events. Fillers are vocalizations to hold floor (e.g. "uhm", "eh", "ah") [8]. The corpus is collected over 60 phone calls[1] among 120 subjects 63 of whom are female. To provide speaker independence in the challenge [8], the first 35 calls amounting to 70 speakers are used for training, calls from 36 to 45 for development and the remaining (46-60) for test set.

Table 1. Corpus Summary Statistics

Property	Statistic
# of Clips	2763
Clip Duration	11 sec.
# of Phone Calls	60
# of Subjects	120
# Male Subjects	57
# Female Subjects	63
# of Filler Events	3.0 k
# of Laughter Events	1.2 k

The task is to classify each non-overlapping frame of length 10 ms into one of filler, laughter and garbage classes. The total number of 10 ms-frames in the data set is 3 027 949 (posing a high cardinality challenge). In the training set there are 59 294 frames of laughter, 85 034 frames of filler and 1 591 442 frames of garbage class. As a sub-sampling measure, the challenge paper propose taking one every 20 'garbage' frame in training set reducing the number to 79 572 [8].

Taking memory limitations incurred from high cardinality into account, a small set of affectively potent features [2,4,9,10] are extracted using frame-wise LLDs and functionals [8]. Frame-wise LLDs are composed of MFCC 1-12, logarithmic energy, voicing probability, HNR, F0, and zero crossing rate together with their first order delta regression coefficients. For MFCC and logarithmic energy, second order delta coefficients are also extracted. These frame-wise LLDs are augmented with the arithmetic mean and standard deviation of frame itself and 8 of its neighbors (4 before and 4 after) totaling 141 features [8].

[1]microphone recordings of Nokia N900 model phones at both ends

4. Methodology

For classification, we used Random Forests (RF) [11]. Random forests are a combination of decision tree predictors (for details of decision trees see [12] ch. 9). Each decision tree is grown with a set of randomly selected instances (sampled with replacement) with a subset of features which are also randomly selected. The instance selection with replacement leaves on the average one third of the instances 'out of the bag' for each tree. These out of bag instances are used to provide 'an unbiased estimate of generalization error' [11]. The forest outputs the class that is the mode of the classes found by individual trees. We think Random Forests fit the problem well, since they are shown to be superior to current algorithms in accuracy and run efficiently on large databases [11].

The dimensionality is reduced with minimum Redundancy Maximum Relevance (mRMR) feature selection method [12]. mRMR is a feature ranking method which suggests incrementally selecting the maximally relevant variables while avoiding the redundant ones with the aim of selecting a minimal subset of variables that represents the problem. Let $\{S_i\}$ denote the set of already selected features and y the target variable and $I(a,b)$ the mutual information between a and b vectors; mRMR selects the feature f with maximal difference (or ratio) between $I(f,y)$ and $\sum_i I(f, S_i)$.

Our baseline is a linear SVM classifier, also used in the Challenge paper [1] (for details of SVM you may consult [12] ch. 13). We replicate the experiments of [8] to provide an independent baseline .

4.1. Experimental Results

We have utilized the Weka [13] implementation of SVM and Random Forests. For mRMR we used Peng et al.'s original implementation [14]. In all our experiments we set the seed parameter to its default value of 1 to ease repeatability.

For Social Signal Sub-Challenge, we consider the Area Under the Curve (AUC) measure of laughter and filler classes and their unweighted average (UAAUC) as suggested in [8].

4.1.1. Reproducing Baseline with Linear SVM

The development set UAAUC results reported in the paper of the challenge (for C=0.1 with linear kernel) are 86.2% for laughter and 89.0% for filler. As used in the challenge paper, we experimented the down-sampled dataset which was attained by taking every 20th frame (5%) of 'garbage' class. However, the repeated experiments with the same setting did not result in the same AUC values (see Table 1). We attributed this to the possibility that the reported performance is that of original set rather than the down-sampled set. To provide a baseline for our study we issued an experiment with the full set of features using SVM with linear kernel and complexity parameter C = {0.1, 1, 10}. As can be seen in Table 2, all three settings of C resulted in almost the same AUC.

Table 2. *SVM Baseline on Development Set (% AUC).*

C	Laughter	Filler	UAAUC
0.1	81.3	83.6	82.5
1	81.2	83.7	82.5
10	81.2	83.7	82.5

We further evaluated the effect of mRMR filter with highest ranking 50, 70 and 90 features. In order to use the mRMR feature selection, we discretized the features at the preprocessing step as suggested by Peng et al. [14]. For this, we computed the z-score, then rounded *z-score/10* to nearest integer (i.e. 10 bins for 1 standard deviation) for each feature independently.
We observed no difference with respect to the hyper-parameter C, so we provide the results for C=0.1. The results indicate that as the number of selected features decrease, the development set detection rate goes down slightly for SVM with linear kernel (see Table 3). When 50 out of 141 original features are used, the decrease in UAAUC is around 2% keeping all other parameters the same with the challenge baseline setting.

Table 3. *SVM Performance with Linear Kernel & mRMR Ranked Features on Development Set (% AUC).*

# of mRMR Features	Laughter	Filler	UAAUC
90	80.7	83.3	82.0
70	79.9	83.0	81.5
50	78.8	82.4	80.6

Further studies with the SVM including partitioning the feature set and training a separate model for late fusion did not provide higher AUC results than those reported in Table 1. When we tried to train SVM models with quadratic and RBF kernels, the training took much longer than with linear kernel so we aborted further work with other SVM kernels. We have observed that the training time increases more than ten times when C is increased ten fold. When C is set to 10, training with 50 features took more than $100x10^3$ seconds.

4.1.2. Experiments with Random Forests

Random Forests are thought to suit the problem better than SVMs, as they perform relatively better than current algorithms and run efficiently on large databases [11]. Hyper-parameters of an RF are the number of (randomly selected) features per decision tree (denoted *d*) and the number of trees (denoted *T*) to form the forest. We also investigated the effect of mRMR ranked features (denoted *D*), as in SVM. The tested values used for the hyper-parameters are d = {8,16,32}, T= {10,20,30} and D = {50,70,90, All}. Note that since $d \leq D$, we did not investigate higher values of d.

Table 4. *Random Forest Performance with T=20, varying d and D on Development Set (% AUC).*

d	D	Laughter	Filler	UAAUC
8	All	88.9	90.7	89.8
	90	89.3	90.8	90.1
	70	88.3	90.3	89.3
	50	88.0	89.7	88.9
16	All	89.0	90.6	89.8
	90	89.3	90.7	90.0
	70	88.8	90.7	89.8
	50	88.3	90.0	89.2
32	All	89.5	90.9	90.2
	90	89.6	90.9	90.3
	70	89.1	90.8	90.0
	50	88.2	90.0	89.1

The UAAUC results of RFs were found better than those found with SVMs. While best baseline SVM/UAAUC performance was 82.5%, the poorest performing RF model (trained with *d*=8, *D*=50, *T*=10) resulted in an UAAUC performance of 86.3%. Table 4 lists the simulation results with varying local dimensionality (d) and global dimensionality (D) parameters, where the number of trees (T) is set to 20. As can be seen in the table, the average performance of Random Forests is generally better than SVMs' performance. Moreover with d=16 we see that reducing the number of global features via mRMR to half (D=70) does not trade-off UAAUC performance.

Table 5. *Random Forest Performance with varying T, d and D on Development Set (% UAAUC).*

d	D	10 Trees	20 Trees	30 Trees
8	All	87.4	89.8	89.7
	90	**88.0**	**90.1**	**90.1**
	70	87.9	89.3	89.8
	50	87.5	88.9	89.4
16	All	88.2	89.8	90.4
	90	**88.6**	**90.0**	**90.5**
	70	88.5	89.8	90.2
	50	87.8	89.2	89.6
32	All	88.8	90.2	90.4
	90	**89.0**	**90.3**	**90.8**
	70	88.7	90.0	90.4
	50	87.9	89.1	89.6
Avg		88.2	89.7	90.0

Table 5 summarizes all simulations showing the UAAUC measure for all tested values of d, D and T. UAAUC measure was found to increase with increasing values of *d, D* and *T*. The average performance increase from *T*=10 to *T*=20 was found to be higher (1.5%) than the difference between average *T*=30 and *T*=20 (0.3%) UAAUC. Moreover, the average performance decrease with respect to D (from 141 to best mRMR 50) is observed to be less than 1%, whereas in SVM simulations it was found to be around 2%. The average difference between UAAUC measure of models with full set of features and models with first 50 mRMR features is less than 1%. We also see that using first 90 mRMR features (in bold), it is possible to attain a slightly better performance than the full feature set. We further investigated the effect of Gaussian smoothing on posteriors making use of frame neighborhood information. For this, the class posteriors of a frame were re-calculated as a weighted sum of itself and its 2K neighbors (K before and K after) where K ranged from 1 to 10. The Gaussian weight function used to smooth frame *i* with a neighboring frame j, $i-K \leq j \leq i+K$ is given as

$$w_{i,j} = (2 * pi * B)^{-(1/2)} * \exp(-|i-j|/(2*B)) \quad (1)$$

where |i-j| is the absolute value of the difference of corresponding frame indices. For simplicity we set B=1 in our tests. To relatively assess the performance increase with respect to K, we computed the development set accuracy instead of UAAUC. Since accuracy increase from K=8 to K=9 was found to be less than 0.05% (see Figure 1), we chose K to be 8. AUC measures with Gaussian smoothed posteriors were found to be 92.2 and 92.4 for laughter and filler, respectively. This corresponds to an increase of 1.5% absolute in UAAUC for the development set.

To probe the performance of best model setting, we trained RF with 30 trees each with 32 features randomly drawn from best 90 mRMR ranked features using training and development set together. Then, we applied Gaussian smoothing with K=8 (16 neighboring frames). We attained 89.6% and 87.3% AUC for laughter and filler respectively. The resulting test set UAAUC (88.4%) outdid the challenge baseline (83.3%) by 5.1% absolute.

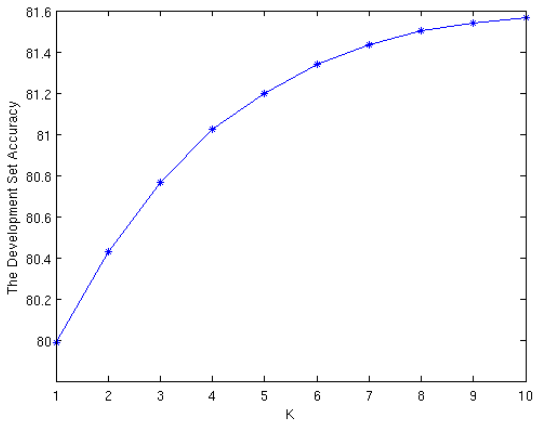

Figure 1: Effect of Gaussian Smoothing on The Development Set Accuracy with K Neighbors Before and After

5. Conclusions and Future Work

In this study, we proposed the use of random forest classification for the problem of laughter and filler detection. Our results suggest that RFs provide superior performance in terms of accuracy and efficiency (especially when tree learning is parallelized). Due to their learning structure, RFs are capable of providing an unbiased estimate of generalization error at training time. The core problem of this sub-challenge is the localization i.e. frame-wise recognition of laughter and fillers. This limits the number of features to use as it proliferates the number of instances.

In this paper we used the standard feature set delivered along with the challenge data. In order to provide a baseline similar to that provided on the challenge paper, we used SVM with linear kernel. We investigated the possibility and the effects of feature reduction via mRMR - a recently popular feature selection method. The relatively poor performance and long training time of SVMs lead us analyze the dataset with Random Forests where we attained better results. When first 90 mRMR features were used, the performance was found to be slightly higher than full set. We have also observed that post-processing using the frame neighborhood information contributed to performance. Using Gaussian weighted posteriors of the 16 neighboring frames, it was possible to increase UAAUC by 1.5%. In the development set, we attained an increase of 9.8% from the baseline setting provided in the challenge paper (SVM with linear kernel, C=0.1). Finally, we applied the best hyper-parameter setting to train a RF from training and development set together, followed by Gaussian smoothing. We attained an UAAUC of 88.4% in test set, improving the challenge baseline by 5.1% (absolute).

Future work of our study include introducing other features e.g. relative spectra, and utilizing generative modeling (GMM variants) to cope with several types of laughter. Moreover, the effect of B in Gaussian smoothing need to be investigated.

6. Acknowledgments

H. Kaya is supported by Faculty Member Training Program (ÖYP) lead by Turkish Higher Education Council (YÖK) and partly funded by PhD scholarship of Scientific and Technical Research Council of Turkey (TÜBİTAK). A. A. Salah is partially supported by Bogazici University BAP 6531 project.

7. References

[1] Dupont, S., & Luettin, J., “Audio-visual speech modeling for continuous speech recognition”, IEEE Transactions on Multimedia, 2(3), 141-151, 2000

[2] Ito, A., Xinyue, W., Suzuki, M., and Makino, S., “Smile and Laughter Recognition using Speech Processing and Face Recognition from Conversation Video”, In Proceedings of the 2005 International Conference on Cyberworlds, Washington, DC, USA, 437-444, 2005

[3] Truong, K. P., and Van Leeuwen, D. A., “Automatic discrimination between laughter and speech”, Speech Communication, 49(2), 144-158, 2007

[4] Knox, M., and Mirghafori, N., “Automatic laughter detection using neural networks”, Proc. Interspeech 2007, 2973-2976, 2007

[5] Petridis, S., and Pantic, M., “Audiovisual discrimination between laughter and speech”, In Acoustics, Speech and Signal Processing ICASSP 2008. IEEE International Conference on, 5117-5120, 2008

[6] Petridis, S., and Pantic, M., “Fusion of audio and visual cues for laughter detection”, In Proc. 2008 International conference on Content-based image and video retrieval (CIVR '08), ACM, New York, NY, USA, 329-338, 2008

[7] Vinciarelli, A., Pantic, M., Bourlard, H., “Social signal processing: Survey of an emerging domain”, Image and Vision Computing 27.12 1743-1759, 2009

[8] Schuller, B., Steidl S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., Salamin, H., Polychroniou, A., Valente F. and Kim S., “The Interspeech 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism”, Proc. Interspeech 2013, ISCA, Lyon, France, 2013

[9] Schuller B., In Salah, A. A. and Gevers, T. (eds) “Computer Analysis of Human Behavior” chap. Voice and Speech Analysis in Search of States and Traits, 227–253, Springer, 2011.

[10] El Ayadi, M., Kamel , M. S., and F. Karray, “Survey on speech emotion recognition: Features, classification schemes, and databases”, Pattern Recognition, 44(3), 572–587, Mar. 2011

[11] L. Breiman, “Random Forests”, University of California, Berkeley,USA, 2001

[12] Alpaydın E., “Introduction to Machine Learning”. Massachusetts, USA: The MIT Press, 2010

[13] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, “The WEKA Data Mining Software: An Update; SIGKDD Explorations”, 11(1), 2009.

[14] Peng, H., Long, F., and Ding, C., “Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(8), 1226-1238, 2005