Non-linguistic Vocalisation Recognition Based on Hybrid GMM-SVM Approach

Artur Janicki

Institute of Telecommunications, Warsaw University of Technology
ul. Nowowiejska 15/19, 00-665 Warsaw, Poland
A.Janicki@tele.pw.edu.pl

Abstract

This paper describes an algorithm for detection of non-linguistic vocalisations, such as laughter or fillers, based on acoustic features. The algorithm proposed combines the benefits of Gaussian mixture models (GMM) and the advantages of support vector machines (SVMs). Three GMMs were trained for garbage, laughter, and fillers, and then an SVM model was trained in the GMM score space. Various experiments were run to tune the parameters of the proposed algorithm, using the data sets originating from the SSPNet Vocalisation Corpus (SVC) provided for the Social Signals Sub-Challenge of the INTERSPEECH 2013 Computational Paralinguistics Challenge. The results showed a remarkable growth of the unweighted average of the area under the receiver operating curve (UAAUC) compared to the baseline results (from 87.6% to over 94% for the development set), which confirmed the efficiency of the proposed method.

Index Terms: paralingustics, social signals, laughter detection, filler, support vector machines, Gaussian mixture models, cepstrum

10.21437/Interspeech.2013-57

1. Introduction

Detection and analysis of non-linguistic vocalisations plays an increasing role in speech signal analysis. The Social Signals Sub-Challenge of the INTERSPEECH 2013 Computational Paralinguistics Challenge [1] focuses on the detection and localisation of paralinguistic events, such as laughter or fillers, based on acoustic parameters. The work described in this paper aims to contribute to this challenge.

In the past, many scientists have investigated negative aspects of paralinguistic cues. In [2], the authors aimed to detect and remove breath sounds from recordings containing speech and singing, in order to improve the aesthetics of the recorded voice. In [3], a work was undertaken to make speech recognition robust against disfluencies caused by fillers or lip smacking.

However, in parallel to this, researchers have tried in various ways to take advantage of information contained in non-linguistic vocalisations, treating them more as places that can convey 'social signals' rather than just as a nuisance. Several studies have focused on laughter detection, see [4], [5], or [6]. In [7], Hidden Markov Models (HMMs) were successfully applied for classifying signal frames into words, laughter, vocal noises (breathing, coughing), non-verbal consents ('mhm') and fillers ('um', 'uh'). There were also projects that aimed to detect fillers and disfluencies based on text analysis, e.g., [8]. In [9], the authors detected breath sounds in order to diagnose psychiatric diseases, such as schizophrenia. In [10], non-speech sounds were automatically classified using support vector machines and multivariate adaptive regression splines (MARS) to classify sneezing, screaming, laughter and snoring, based on spectral and mel-frequency cepstral coefficients (MFCCs). In [11], the researchers used information on fillers to help in the detection of basic emotions; however, they used a corpus with the fillers annotated manually. The positive impact of information contained in breath sounds on the recognition of the speaker's identity was shown and analysed in [12].

The ComParE challenge, the evaluation procedure and the challenge data were described in detail in [1]. This paper will concentrate on describing our approach (next section), followed by a description of the experiments and their results. A brief discussion of the results and a summary will conclude the paper.

2. Proposed approach

2.1. Inspirations

A hybrid Gaussian Mixture Models - Support Vector Machine (GMM-SVM) approach has been proposed as the recognition method of non-linguistic vocalisations. This method takes its inspiration from research on speaker recognition.

Speaker recognition often uses GMMs as an efficient method of modelling the cepstral parameters of the speaker [13]. GMMs are used especially for text-independent speaker recognition, where no specific phonetic sequence is expected. Laughter is a somewhat similar case owing to its variety [14], which makes it difficult to model as a sequence of acoustic events. This is why GMMs, which (unlike HMMs) have only one state, were used for laughter detection in [5], where the authors achieved the equal error rate (EER) results ranging between 7.1 and 20% for the tested corpora.

Fairly recently, GMMs in speaker recognition were combined with large-margin discriminators, such as SVMs. In [15], the authors proposed using an SVM machine to classify supervectors containing Gaussian mean values of speaker GMM models. The supervectors were constructed by stacking mean values of Gaussian components into one supervector. This means that if, for example, GMMs operated in a 12-dimensional MFCC space and if 32 Gaussians were used, 12 x 32 = 384-elements long supervectors were created and SVMs were supposed to operate in 384-dimensional space. The authors claimed they achieved better recognition results than using the classical GMM approach. In [16], the authors proposed SVMs with the so-called GUMI kernels, which also combined SVMs and GMMs but this time, also exploited the speakers information from the GMM covariance matrices. A hybrid SVM-GMM approach with the Kullback-Leibler kernel was also successfully applied in [17] in order to recognise speakers from coded speech.

2.2. Proposed method

Following encouraging applications in speaker recognition we proposed taking advantage of generative modelling (GMMs) combined with a discriminative classifier (an SVM) for recognition of non-linguistic vocalisations in the Social Signals Sub-Challenge. In our approach, however, no supervectors were constructed, thus the space dimension was not multiplied. In the proposed method the SVM operated in GMM score space, defined by log-likelihood values of garbage, laughter and filler GMMs. This is why the original cepstral space (12-, 24- or 36-dimensional) was shrunk to two up to six dimensions and the SVMs were trained to find decision boundaries between the GMM log-likelihood scores. Using SVMs to find decision boundaries between HMM models had been previously reported in [18].

In the method here proposed the training process consisted of the three stages:

- In the first stage, GMMs were trained for garbage, laughter and filler frames from the training signal, using either basic MFCC parameters, or the basic MFCCs extended by their dynamic derivatives: Δ or Δ^2.
- In the second stage, log-likelihood scores were calculated for the training file, using the three GMMs (for garbage, laughter and fillers) generated in the previous stage and cepstral parameters taken over a sliding window.
- In the third stage, a three-class SVM classifier was trained using the GMM garbage/laughter/filler scores and the data labels.

GMMs with a number of Gaussian components ranging from 8 to 256 were used, with diagonal covariance matrices. They were trained using the Expectation-Maximisation (EM) algorithm, using vector quantisation (VQ) for model initialisation. The use of the three GMMs resulted in achieving three log-likelihood values from each parameter window: LL_g, LL_l and LL_f, for garbage, laughter and fillers, respectively. However, differential scores were calculated as well. In total, six various score combinations were considered, as specified in Table 1, such that the SVMs were tested in six different score spaces.

Table 1: *Tested combinations of GMM scores.*

name	value type	GMM log-likelihood scores
lf	absolute	LL_l, LL_f
glf	absolute	LL_g, LL_l, LL_f
LF	differ.	$LL_l - LL_g, LL_f - LL_g$
LFD	differ.	$LL_l - LL_g, LL_f - LL_g, LL_l - LL_f$
$lfLF$	mixed	$LL_l, LL_f, LL_l - LL_g, LL_f - LL_g$
$glfLFD$	mixed	$LL_g, LL_l, LL_f, LL_l - LL_g,$ $LL_f - LL_g, LL_l - LL_f$

For the classifier, we used SVMs with a polynomial kernel, using Sequential Minimal Optimisation (SMO) as the training algorithm, implemented in the WEKA package [19].

The recognition process consisted of calculating log-likelihood scores for the testing data analogously to the training process and of using the previously trained SVM classifier to assign the GMM scores to the laughter, or filler, or garbage class.

3. Experiment setup

According to the ComParE challenge organisers, the experiments for the Social Signals Sub-Challenge were to be run using the training, development and testing subsets of the SSPNet Vocalisation Corpus (SVC). It was decided to use the 'train.downsampled' data set for training with garbage samples reduced 20 times, in order to maintain a balance between the three recognised classes. The data within the sliding window was labelled after its middle element. The window shift was set to one. During training, it was ensured that the sliding window contained only the data belonging to the same recording, such that there is no overlap between data from adjacent files in the training database. Therefore, the log-likelihood scores for the first $F/2$ and the last $F/2$ data samples of a recording were copied from the scores located at $F/2$ and $L - F/2$ data points, respectively, where L is the length of the recording, and F is the length of the sliding window.

Numerous experiments were conducted in order to find the best configuration of the proposed GMM-SVM classifier. The evaluation was made based on the unweighted average of the area under the receiver operating curve (UAAUC) for detection of laughter and filler classes, achieved on the development set, as requested by the ComParE challenge organisers [1]. The development set contained 500 utterances with 547 789 frames in total, among them 25 750 frames of laughter and 29 432 filler frames.

During the experiments the following parameters were researched:

- The feature vector used in calculations. Three possibilities were considered: 12 MFCC parameters, the MFCCs + Δ, or the MFCCs + Δ + Δ^2.
- The combination of GMM scores used during the SVM training and testing, see Table 1.
- The length of the sliding window (F) over which the GMM scores were calculated. Windows of the length from 10 frames (i.e., 100 ms) up to 60 frames (i.e., 600 ms) were experimented.
- The number of Gaussian components (M) used in the GMMs. GMMs starting with 8 up to 256 Gaussians were tried.
- The complexity factor (C) used in the SVM training.

In addition it was verified how much (if at all) the remaining features, extracted from the SVC corpus and available to the ComParE challenge participants, could contribute to further improvement of the recognition accuracy. Adding the F0-related parameters, energy-based parameters and zero crossing rate (ZCR) parameters, MFCCs, as well as their statistical derivatives (arithmetic mean and standard deviation) were researched. The complexity factor C was set to 0.1, as it gave the best baseline results described in [1]; however, other values of C were also tried, starting from 0.001 up to 10.

Finally, the results achieved for the optimal configuration, both for the development and the test data sets, were compared against the baseline results provided by the ComParE organisers. The results are presented in the next section.

4. Results

Even though the main evaluation criterion was the UAAUC result (also marked for the sake of shortness as UAA), the tables in this section also separately display the area under the

Table 2: *Recognition results (in percentages) for various feature vectors for the selected GMM-SVM configurations.*

configuration	feature vector	AUC_l	AUC_f	UAA	EER_l	EER_f	Acc	UAR
$M = 8, F = 50, glfLFD$	12 MFCC	88.4	84.5	86.5	18.9	24.4	70.4	69.6
	12 MFCC+Δ	90.7	92.5	91.6	15.9	13.7	81.9	78.3
	12 MFCC+Δ+Δ^2	90.6	93.2	91.9	16.1	12.5	80.7	78.4
$M = 128, F = 40, LF$	12 MFCC	90.0	88.0	89.0	17.0	20.0	80.9	72.1
	12 MFCC+Δ	92.3	93.3	92.8	14.3	13.0	86.3	78.9
	12 MFCC+Δ+Δ^2	92.4	94.2	93.3	14.1	11.7	85.7	80.5

receiver operating curve for laughter and fillers (AUC_l and AUC_f, respectively), as well as the EER values (EER_l and EER_f). Recognition accuracy (Acc) and unweighted average recall (UAR) were also analysed.

A comparison of various feature vectors to be modelled by the GMMs is presented in Table 2. It shows that a full MFCC vector, i.e., the basic MFCCs accompanied by their first and second order regression coefficients: Δ and Δ^2, provided the best results almost in all aspects – they assured the highest UAAUC and UAR values, regardless of the score combination, M or F. Therefore, the subsequent experiments were run using the full, 36-element MFCC vector as the feature vector.

Table 3 shows an excerpt from the recognition results for various GMM score combinations. It turned out that in most of the cases, the $lfLF$ score combination (i.e., differential log-likelihood scores between laughter and garbage GMMs, and the filler and garbage GMMs, extended by the absolute scores of the filler and garbage GMMs) yielded the best AUC and EER results, as shown in Table 3. It was noticed that adding absolute scores lf to the differential LF improved especially the filler detection (see the change from 93.7% up to 94.2% in the AUC_f in the upper part of Table 3). Therefore most of the subsequent experiments were conducted in the 4-dimensional $lfLF$ score space. Occasionally, tests in LF and $glfLFD$ score spaces were run, too.

Table 3: *Recognition results (in percentages) for various GMM score combinations (upper part for $F = 30$ and $M = 32$, lower part for $F = 40$ and $M = 128$).*

scores	AUC_l	AUC_f	UAA	EER_l	EER_f
lf	85.0	92.5	88.8	21.4	13.0
glf	92.0	94.0	93.0	14.3	11.4
LF	91.9	93.7	92.8	14.4	12.0
LFD	91.8	93.7	92.8	14.4	12.0
$lfLF$	92.0	94.2	93.1	14.3	11.6
$glfLFD$	90.6	93.2	91.9	16.1	12.5
lf	84.9	92.6	88.8	21.9	13.0
glf	92.3	94.3	93.3	14.1	11.2
LF	92.4	94.2	93.3	14.1	11.7
LFD	92.3	94.2	93.3	14.2	11.7
$lfLF$	92.5	94.5	93.5	14.0	11.4
$glfLFD$	92.5	94.5	93.5	14.0	11.4

Figure 1 displays the relationship between the recognition results and the length F of the sliding GMM window. It shows that with the increase of F, the UAAUC result increases constantly; however, the UAAUC increase for the GMM window length exceeding 50 frames (i.e., for windows longer than 0.5 sec) is only minor. It is noteworthy that the UAR result, after initial growth, begins to decrease for $F > 40$.

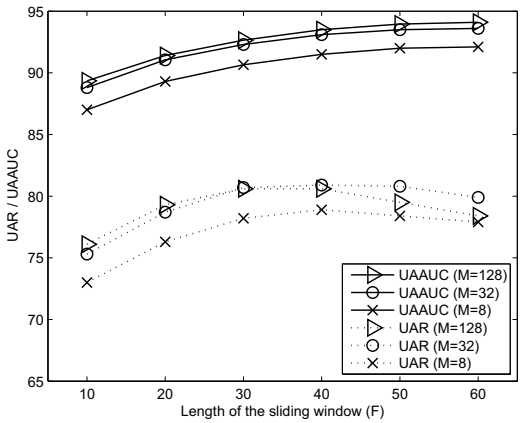

Figure 1: The UAAUC and UAR results (in percentages) against the length of the sliding window (in 10-ms frames) for various M; score combination: $lfLF$.

The increase of the AUC is also visible for an increasing number of Gaussian components (M) – see Table 4, until it reaches 128. Further growth of M shows some decrease of AUC for filler detection. The UAR initially grows, but begins to decrease for the number of Gaussians over 64. Finally, the GMM-SVM classifier with $F = 60$ and $M = 128$, working in the $lfLF$ score space was selected.

Table 4: *The recognition results (in percentages) against the number of Gaussian components used in the GMMs for the $lfLF$ score combination; the upper part is for $F = 40$, lower for $F = 60$.*

M	8	16	32	64	128	256
AUC_l	89.9	91.7	92.0	92.4	92.5	92.6
AUC_f	93.1	93.6	94.2	94.2	94.5	94.4
Acc	79.2	82.2	83.2	84.6	85.8	87.6
UAR	78.9	80.5	80.9	81.2	80.6	79.0
AUC_l	91.2	92.9	93.2	93.6	93.7	93.9
AUC_f	93.0	93.6	94.0	94.1	94.5	94.3
Acc	81.1	84.4	86.0	87.2	88.8	90.4
UAR	77.9	79.2	79.9	80.1	78.4	75.9

It was also verified if adding any features from the provided feature set would improve the recognition results. Various feature groups were sequentially added to the SVM space (in addition to the $lfLF$ scores). Table 5 shows that none of additional features improved either the UAAUC score or recognition accuracy. However, it turned out that the recall result (UAR) in-

creased, among others, when adding the ZCR features or the statistical features of the MFCCs.

Table 5: *Recognition results (in percentages) for various groups of features added from the main feature set to the proposed $lfLF$ score space. The number of features given in brackets. The best results are highlighted.*

added parameters	$UAAUC$	Acc	UAR
none	**94.1**	**88.8**	78.4
all (141)	93.4	79.3	**82.0**
energy-based (3)	93.2	83.1	80.4
F0/HNR and derived (6)	93.8	88.0	78.2
ZCR (2)	93.2	82.0	81.3
MFCCs (36)	93.1	82.4	80.8
statist. energy-based (6)	93.3	82.0	81.2
statist. F0/HNR-derived (12)	93.5	87.0	78.0
statist. ZCR (4)	93.0	80.0	80.3
statist. MFCCs (72)	93.1	80.7	81.0

When experimenting with the complexity parameter C it was observed that changing it to $C = 1.0$ improved slightly the classification performance of the SVM, so this value was finally selected. Table 6 summarises the results and compares them with the baseline ones. The proposed classifier working with the test set, compared to the baseline result presented in [1], allowed the increase of the UAAUC from 83.3% to 89.8%. The UAAUC for the development set increased from 87.6% to 94.2%. The EER values for the development set improved significantly, too. The EER values for the test set were not available to the authors of this article.

Table 6: *Comparison between the results achieved for the proposed method for the optimal configuration and the baseline results (in percentages), for the development and the test sets. The result of the official Social Signals Sub-Challenge competition measure is highlighted.*

metrics	devel base	devel proposed	test base	test proposed
AUC_l	86.2	93.8	82.9	90.7
AUC_f	89.0	94.5	83.6	89.0
$UAAUC$	87.6	94.2	83.3	**89.8**
EER_l	21.2	12.3	N/A	15.3
EER_f	16.9	11.0	N/A	18.4

5. Discussion

We found the achieved improvement of recognition results as remarkable, considering a relatively simple recognition algorithm proposed. The UAAUC for both the development and the test set increased by more than six percentage points. It is believed that this improvement was caused by the advantageous combination of the three GMMs working in the 36-dimensional MFCC space and the discriminative SVM working in the 4-dimensional log-likelihood space, constituted by the $lfLF$ scores. The novel combination of differential (LF) and absolute (lf) log-likelihood scores turned out to be successful, too.

Interestingly, the fillers, as suggested in [1], seemed indeed slightly easier to be recognised in the development set, however, they turned out to be slightly more difficult to be detected in the test set. It could be probably explained if more information on the test set was available. It was also observed that laughter is sometimes confused with breathing, probably due to a somewhat similar nature.

The results showed a positive impact of dynamic parameters (Δ and Δ^2) on the recognition of non-linguistic vocalisations. It is supposed that these regression coefficients help in taking into account the temporal structure of the recognised events. In particular, fillers show a tendency to exhibit typical dynamics; many of the English fillers consist of a vowel (usually 'eh' or 'uh'), followed by a nasal (usually 'm'). Therefore, it is not surprising that the AUC for fillers benefited the most from adding Δ and Δ^2 parameters to the basic MFCCs (see Table 2).

The phenomenon of UAR decreasing for larger M is similar to the one described in [7]; however, in our case, it took place for much higher values of M. In the authors' opinion, this is caused because GMMs with higher number of Gaussian components begin to lose their ability to generalise and become overfitted to the training data. A similar behaviour for the increasing length of the GMM parameter window (F) is probably caused by the fact that shorter laughter/filler events simply become unnoticed in a longer (over 0.5 sec) time span.

Adding any features from the provided feature set did not cause any improvement of either the UAAUC or the detection accuracy. However, if an increase of recall was requested, adding, e.g., statistical features of the MFCCs to the SVM classifier space can be considered.

It was observed that the original speech signal in some recordings severely overloaded the input analogue-to-digital converter. It is suspected that improved acquisition of the audio signal could have a positive impact on the extraction of the parameters and as a consequence, on the detection of various acoustic events.

6. Conclusions

In this paper we have proposed a simple yet effective method for recognising non-linguistic vocalisations, such as laughter and fillers. A hybrid method was elaborated, as it was based on a an SVM classifier working in the space of GMM scores. The best results were achieved when mixed (differential and absolute) log-likelihood scores were used. The experiments showed that GMMs with a high number of Gaussians (128 or more) and with a relatively long parameter window (50 frames or more) yielded the best results for area under the curve; however, at the expense of decreasing recall. The proper balance between UAAUC and UAR should be set depending on the actual application of the algorithm.

As the Social Signals Sub-Challenge official measure was set to UAAUC, a GMM-SVM classifier with 128 Gaussians and 60-frames-long window was selected; thus, enabling laughter and filler detection with UAAUC of 89.8% for the test set and 94.2% for the development set, which is a remarkable improvement compared with the baseline of 83.3% and 87.6%, respectively. Future work could involve combining HMMs and SVMs and comparing them with the GMM-SVM approach proposed in this work.

7. References

[1] 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 Interspeech 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism," in *Proc. Interspeech 2013*, Lyon, France, August 2013.

[2] D. Ruinskiy and Y. Lavner, "An effective algorithm for automatic detection and exact demarcation of breath sounds in speech and song signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 838–850, 2007.

[3] J. Rajnoha, "Speaker non-speech event recognition with standard speech datasets," *Acta Polytechnica*, vol. 47, no. 4-5/2007, pp. 107–111, 2008.

[4] N. Campbell, "Perception of affect in speech - towards an automatic processing of paralinguistic information in spoken conversation," in *Proc. Interspeech 2004*, Jeju Island, Korea, October 2004.

[5] K. P. Truong and D. A. van Leeuwen, "Automatic detection of laughter," in *Proc. Interspeech 2005*, Lisbon, Portugal, 2005, pp. 485–488.

[6] K. P. Truong and J. Trouvain, "On the acoustics of overlapping laughter in conversational speech," in *Proc. Interspeech 2012*, 2012.

[7] F. Weninger and B. Schuller, "Discrimination of linguistic and non-linguistic vocalizations in spontaneous speech: Intra- and inter-corpus perspectives," in *Proc. Interspeech 2012*, 2012.

[8] M. Asahara and Y. Matsumoto, "Filler and disfluency identification based on morphological analysis and chunking," in *Proceedings ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, Tokyo, Japan, 2003.

[9] V. Rapcan, S. D'Arcy, and R. B. Reilly, "Automatic breath sound detection and removal for cognitive studies of speech and language," in *IET Irish Signals and Systems Conference (ISSC 2009)*, 2009, pp. 1–6.

[10] W.-H. Liao and Y.-K. Lin, "Classification of non-speech human sounds: Feature selection and snoring sound analysis," in *IEEE International Conference on Systems, Man and Cybernetics - SMC 2009.*, 2009, pp. 2695–2700.

[11] L. Devillers and L. Vidrascu, "Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs," in *Proc. Interspeech 2006*, 2006, pp. 801–804.

[12] A. Janicki, "On the impact of non-speech sounds on speaker recognition," *LNAI*, vol. 7499, pp. 566–572, 2012.

[13] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," in *Digital Signal Processing*, 2000.

[14] N. Campbell, H. Kashioka, and R. Ohara, "No laughing matter," in *Proc. Interspeech 2005*, Lisbon, Portugal, 2005.

[15] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using gmm supervectors for speaker verification," *IEEE Signal Processing Letters*, vol. 13, pp. 308–311, 2006.

[16] C. H. You, K.-A. Lee, and H. Li, "GMM-SVM kernel with a Bhattacharyya-based distance for speaker recognition," *IEEE Transactions on Audio, Speech & Language Processing*, pp. 1300–1312, 2010.

[17] A. Janicki and T. Staroszczyk, "Speaker recognition from coded speech using support vector machines," *LNAI*, vol. 6836, pp. 291–298, 2011.

[18] M. J. F. Gales and M. Layton, "SVMs, score-spaces and maximum margin statistical models," in *Proc. Beyond HMM workshop*, 2004.

[19] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The Weka data mining software: An update," *SIGKDD Explorations*, vol. 11, no. 1, 2009.