INTERSPEECH 2009 BRIGHTON

Speaker Dependent Emotion Recognition Using Prosodic Supervectors

Ignacio Lopez-Moreno, Carlos Ortego-Resa, Joaquin Gonzalez-Rodriguez and Daniel Ramos

ATVS Biometric Recognition Group, Universidad Autonoma de Madrid, Spain

ignacio.lopez@uam.es

Abstract

This work presents a novel approach for detection of emotions embedded in the speech signal. The proposed approach works at the prosodic level, and models the statistical distribution of the prosodic features with Gaussian Mixture Models (GMM) mean-adapted from a Universal Background Model (UBM). This allows the use of GMM-mean supervectors, which are classified by a Support Vector Machine (SVM). Our proposal is compared to a popular baseline, which classifies with an SVM a set of selected prosodic features from the whole speech signal. In order to measure the speaker inter-variability, which is a factor of degradation in this task, speaker dependent and speaker independent frameworks have been considered. Experiments have been carried out under the SUSAS subcorpus, including real and simulated emotions. Results shows that in a speaker dependent framework our proposed approach achieves a relative improvement greater than 14% in Equal Error Rate (EER) with respect to the baseline approach. The relative improvement is greater than 17% when both approaches are combined together by fusion with respect to the baseline.

Index Terms: emotion recognition, speaker inter-variability, supervectors, SVMs

1. Introduction

Emotion recognition from the speech signal is an increasingly interesting task in human-machine interaction, with diverse applications in the speech technologies field such as call centres, intelligent auto-mobile systems, speaker intra-variability compensation or entertainment industry [1]. Emotion recognition is generally stated as a multiclass classification problem, where a given speech utterance is classified among n emotions (classes). However, it is of interest to detect a given emotion in a speech segment, which justifies the use of a verification or detection approach described as follows: given a speech utterance and a target emotional state e from the whole n emotions set, the objective is to determine whether the dominant emotion that affect the speaker in the utterance is e or not. Thus, emotion detection is essentially a two-class problem, where the *target* class is true when e is the dominant emotion in the test utterance and the *non-target* class is true when it is not. The standard architecture in such scheme is to compute a similarity measure (a *score*) among an emotion model of e and the emotion in the test utteranace, which will be further compared to a threshold for detection.

Recognizing emotions from speech is essentially motivated from their nature: affective states caused by subjective judgements, memories and sensations frequently accompanied of physical and psychological changes of the well-being sensation. Thus humans can recognize emotions by the study of those changes of the neutral states, including the semantic level of the speech, non usual behaviours and decisions, as well as other not so high cognitive levels, commonly more capable to be learned by machines [2].

Unluckily, emotion recognition from speech is a difficult task, mainly because of two reasons. First, emotions does not manifest in the same way in different speakers, and therefore, inter-variability of speakers seriously affects the recognition process. Second, it is difficult to define the target emotions set because the limits among different emotions may not be clear for listeners in general, and several emotions from the considered set can be simultaneously in the same utterance, or even at the same moment in time. Despite the difficulty of the challenge, the research in the area has experimented an increase in the last years, which has motivated the availability of emotional labeled speech corpora. Most popular ones are FAU AIBO Emotion Corpus [3], SUSAS, EMO-DB, ISL meeting corpus, Danish Emotional Speech Database [4] and recently Ahumada III [5].

In this work, we present a novel method for emotions detection based on Gaussian Mixture Models (GMM) of short-term prosodic features, whose supervectors are further classified with Support Vector Machines (SVM). Moreover, we present results of the fusion of the proposed system with a baseline, based on a popular approach of modelling utterance-level prosodic features with SVM. We show that the proposed approach, namely prosodic SVM-GMM, models distances among complete join probability distributions of the prosodic features, and not only with some significant values, as happen with the baseline system. Moreover, the fusion of both systems significantly improves the performance of proposed approach, which indicates uncorrelated information among both methods. We evaluate the proposed system in a speaker-dependent and a speaker-independent scenario. Experiments are presented using the SUSAS database [6].

This work is organised as follows. The role of prosody and the proposed prosodic parametrization is described in Section 2. In Section 3, the proposed system is described in detail, as well as the baseline and the approach for fusion of both systems. Section 4 describes the experimental work which shows the adequacy of the approach. Finally, conclusions are drawn in Section 5.

2. Prosodic features for emotion recogniton

Many works had shown the relation between the variation of speaker prosody and the information of their emotional states [7]. Therefore prosodic features are often considered as input signals in many emotion recognition systems. Frequent prosodic features are the fundamental frequency (*pitch*), the energy and their velocity, also known as Δ features [8].

The proposed GMM-SVM approach in this work uses a prosodic feature extraction scheme in the following way: the audio signal is windowed every 10ms using a 40ms Hamming

window. For every window, energy and log pitch values are extracted (Fig.1) using Praat [9] toolbox. In vocal segments, velocity information is obtained as a difference between two consecutive windows. Using a voice activity detector (VAD), non-voiced segments are erased by accepting only those windows with pitch and energy values higher than a threshold. As a consequence, for every utterance u, the feature vector set consist of a set of $d = 4$ dimensional feature vectors, or streams (energy, pitch and their Δ features). It is possible to normalize each stream by subtracting its mean value. Energy and delta-energy normalization have been applied to the proposed GMM-SVM approach while only energy normalization for the baseline.

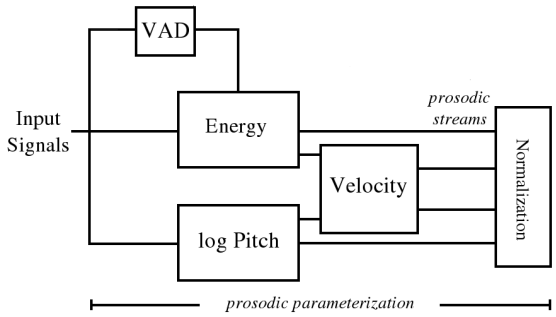

Figure 1: *Block diagram of the prosodic feature extraction module.*

3. A prosodic GMM-SVM approach for emotion detection

This section details the novel prosodic GMM-SVM system proposed in this paper, the baseline modelling scheme and the fusion approach for combining information from both systems.

3.1. Proposed approach

SVM-GMM supervectors have been previously used for emotion recognition at the spectral level of the speech in [10]. This technique also shows an excellent performance in speaker and language recognition. The main advantage of this proposed technique is that it is capable to summarize the whole probability density function (pdf) of the feature vectors in utterance u, into a single high-dimensionality vector known as a GMM supervector. This supervector is obtained by the concatenation of the vectors of means of a d-dimensional GMM model obtained from all the d-dimensional prosodic vectors in the utterance (Figure 2). The M-mixture GMM, is calculated as a Maximum a Posteriori Adaptation (MAP) from a Universal background Model (UBM), which is an standard M-mixtures GMM model, trained with a large amount of development data from all the emotional states available. Thus, the UBM aims at representing the emotion-independent statistical distribution of the features.

The GMM supervector can be considered as a kernel function $sv(u)$ that maps the prosodic features of u in a high-dimensional vector of size $L' = M * d$. This L'-dimensional supervector space is where an SVM is used to obtain a final model $\vec{w_e}$ of the target emotion e. In this case the scoring function $s'(w_e, sv(u_{test}))$ for every testing utterance u_{test} is defined as follows

$$s'(\vec{w_e}, sv(\vec{u_{test}})) = \vec{w_e} * sv(\vec{u_{test}})^T$$

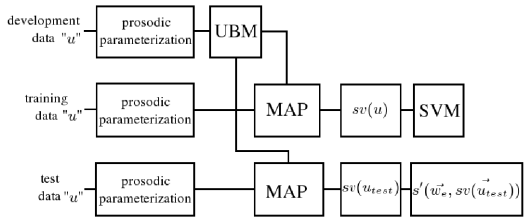

Figure 2: *Block diagram of the GMM Supervector based SVM.*

3.2. Baseline approach

The baseline system is based on a popular scheme presented in [8]. For every utterance u, the statistical distribution of the prosodic vectors is characterized by computing $n = 9$ values for each one of the prosodic streams (table 1). Thus, we obtain a $L = d * n$ fixed-length feature vector per utterance. This new derived L-dimensional feature space is where emotions are modeled by using a one-versus-all linear SVM (Figure 3. Note that this L-dimensional feature vector can be seen as the result of a kernel function $l(u)$, that maps the d-dimensional prosodic vectors of u into a L-dimensional feature space.

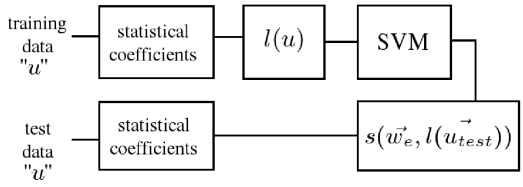

Figure 3: *Block diagram of the Baseline Classifier.*

Given an SVM model$\vec{w_e}$ of an emotion e, the scoring function $s(\vec{w}, l(\vec{u}))$ for every test utterance u_{test} is a simple dot product computed as follows:

$$s(\vec{w_e}, l(\vec{u_{test}})) = \vec{w_e} * l(\vec{u_{test}})^T$$

Table 1: *Statistical coefficients extracted for every prosodic stream in the Baseline approach.*

Coefficients
Maximum
Minimum
Mean
Standard deviation
Median
First quartile
Third quartile
Skewness
Kurtosis

On the one hand, the similarities between the proposed prosodic GMM-SVM system and the baseline are: i) Previous d-dimensional prosodic features vectors are used as inputs, ii) The modeling of their long-term statistical distribution (pdf) of the vectors in u by using linear SVMs and iii) Both cases are an attempt to characterize pdf. Nevertheless, the method used to characterize pdf's differs between both presented subsystem. As a consequence, not only performances differ, also

uncorrelated scores are generated. This fact motivates a posterior subsystem fusion in order to increase the final performance achieved. On the other hand, the baseline only uses a small set of well performing values to characterize the pdf of the vectors in every u, but probably they are not seizing the whole information embedded in it. Note for example that the baseline subsystem compute the n statistical values stream by stream, not using the correlated information among them.

3.3. Subsystem fusion

Final scores generated by the system are combinations of $s'(\vec{w_e}, sv(u_{test}))$ and $s(\vec{w_e}, sv(u_{test}))$. Combination is performed as a sum fusion preceded of a test normalization (Tnorm [ref]) stage, which fosters a similar range of the scores of both subsystems. Tnrom cohort is form by the whole set of emotions models w_e, for $e = 1...N_{emotions}$. The final combined score $S(\vec{w_e}, u_{test})$ is computed as follows

$$S(\vec{w_e}, u_{test}) = \frac{s'(\vec{w_e}, sv(u_{test})) - \mu'}{std'} + \frac{s(\vec{w_e}, sv(u_{test})) - \mu}{std}$$

Where μ' and μ are the means of the cohort scores, and std' and std the standard deviations. Referred to the Proposed and Baseline systems respectively.

4. Experiments

4.1. SUSAS: emotional speech database

The proposed emotion recognition system has been tested over the English SUSAS database (Speech Under Simulated And Actual Stress). SUSAS has been employed frequently in the study of the effects of speech production and recognition, when speaking under stressed conditions [8]. This database was designed originally by John H.L. Hansen, et al. in 1998 for speech recognition under stress. All speech files from SUSAS database were sampled at 8kHz, and 16-bit integers. SUSAS Simulated subcorpora contains speech from 9 speakers and 11 speaking styles. They include 7 simulated styles ($slow$, $fast$, $soft$, $question$, $clear\ enunciation$, $angry$) and four other styles under different workload conditions ($high$, $cond70$, $cond50$, $moderate$). SUSAS Actual speech contains speech from 11 speakers, and 5 different and real stress conditions ($neutral$, $medst$, $hist$, $freefall$, $scream$). Actual and Simulated subcorpora contains 35 spoken words with 2 realisation of each, for every speaker and speaking style. The SUSAS database has been selected for the following reasons: i) presents a large set of target emotions; ii) allows comparisons with previous work in the literature; iii) speaker IDs are available; and iv) there exist simulated and actual emotional states. These two last subcorpora, namely Simulated and Actual, have characteristics different enough to consider them as different databases.

4.2. Results

Speaker inter-variability can cause that different emotions and different speakers may be located in the same region in the feature space. This drawback can be compensated by using speaker independent emotion models. To compare the performance improvement between both scenarios, we carried out speaker dependent and speaker independent experiments. Experiments are performed for both SUSAS subcorporas, Simulated and Actual. Both subcorpus have been divided in three non-overlapped sets with equivalent amount of data: training set, testing set, and a development set used for UBM training.

Any model $w_e(spk)$ or $w'_e(spk)$, for the baseline and the proposed prosodic GMM-SVM subsystems respectively, will be denoted as $w_e(spk)$ for simplicity. Performance results will be measured in terms of equal error rate (EER), which is a popular performance measure for any detection task.

4.2.1. Speaker Independent Experiments

For detection of target emotion e, every model w_e is trained using data belonging to e as the target class, and any other emotion as the non-target class. Therefore we will obtain 11 emotion models for Simulated speech and 5 models for Actual speech. In order to obtain results not affected by speaker overfitting, training, testing, and development sets, each experimental subset of SUSAS will be built with different speakers.

Table 2: *EER(%) in Speaker Independent experiments for SUSAS Simulated speech. R.I. denotes the relative improvement of $Combine$ in respect of $Baseline$.*

Emotion	Baseline	Proposed	Combined	R.I. %
angry	18.16	20.47	16.73	+7.87
clear	42.68	31.04	31.99	+25.05
cond50	40.76	39.84	38.22	+6.23
cond70	42.28	40.21	40.43	+4.37
fast	24.31	27.23	20.63	+15.13
lombard	51.24	42.06	42.55	+16.96
loud	23.03	24.57	21.03	+8.68
neutral	36.29	35.33	34.38	+5.26
question	12.44	4.38	4.38	+64.79
slow	19.60	26.10	22.46	-14.59
soft	20.65	38.19	22.26	-7.79
Avg. EER	30.13	29.94	26.82	+10.37

Table 3: *EER(%) in Speaker Independent experiments for SUSAS Actual speech.*

Emotion	Baseline	Proposed	Combined	R.I. %
neutral	35.12	34.61	33.31	+5.15
medst	40.99	42.21	41.51	-1.26
hist	36.82	38.97	35.75	+2.9
freefall	25.07	54.75	31.29	-24.81
scream	6.46	11.68	7.6	-17.64
Avg. EER	28.89	36.04	29.78	-3.08

Results in tables 2 and 3 shows better performance for Actual subcorpus than for Simulated one. This fact is probably cuased by the less number of target classes, which makes the performance of the detection of a target emotion with respect to the rest easier. Also note that the EER for similar classes such as $cond50$, $cond70$ and $lombard$ is higher than for other more differentiable emotions such as $question$ and $angry$. This emphasizes the strong dependence of the performance on the emotion set.

4.2.2. Speaker Dependent Experiments

For a speaker spk and a target emotion e, every model $w_e(spk)$ is trained using all the utterances belonging to simultaneously spk and e for the target model. Non-target model is trained in this scenario using data from all speakers and emotions except those included in the target model training set.

Table 4: *EER(%) in Speaker Dependent experiments for SUSAS Simulated speech.*

Emotion	Baseline	Proposed	Combined	R.I. %
angry	11.07	12.00	9.04	+18.33
clear	37.51	26.31	26.34	+29.77
cond50	37.40	33.61	32.38	+13.42
cond70	37.17	33.52	33.14	+10.84
fast	20.18	19.71	15.62	+22.59
lombard	31.14	29.02	26.63	+14.48
loud	15.56	11.27	10.17	+34.64
neutral	32.22	27.31	26.04	+19.18
question	5.80	3.19	1.98	+65.86
slow	16.66	15.08	13.17	+20.94
soft	10.13	15.67	10.18	-0.49
Avg. EER	23.16	19.70	18.60	+19.68

Table 5: *EER(%) in Speaker Dependent experiments for SUSAS Actual speech.*

Emotion	Baseline	Proposed	Combined	R.I. %
neutral	18.23	17.21	15.23	+16.45
medst	27.06	24.29	22.79	+15.77
hist	23.35	21.53	19.85	+14.98
freefall	25.40	19.27	20.97	+17.44
scream	8.31	5.72	5.72	+31.16
Avg. EER	20.47	17.60	16.91	+17.39

Results in tables 4 and 5 shows that by combining individual classifiers in a speaker dependent framework, we can achieve better performance than for any of them separately. Relative improvements of the combined approach respect to the baseline are about 17.4% or 19.7% in Actual and Simulated speech respectively. Table 6 also shows that class overlapping is remarkable reduced between speaker dependent and independent schemes. Note that the Combined system achieves a relative improvement about 30.64% when it is evaluated in Actual subcorpus. Relative improvement is about 43.21% for Simulated subcorpus.

5. Conclusions

This work introduces a novel approach for emotion recognition using prosodic features. The porposed approaches models the statistical distribution of short-term pitch, energy and their velocities by a GMM, and the a SVM classification of in the mean-supervector space of the models gives the final score for detection. We compare this prosodic GMM-SVM system with a baseline implementing a popular approach also at the prosodic level. Moreover, we explore a combination (fusion) approach with a baseline system, which further increases performance. The task is presented as a verification or detection problem measured in terms of EER. The experimental set-up is based on two subcorpus of the SUSAS database, as well as in two different experimental frameworks: speaker-independent and speaker-dependent. According to results we conclude that the proposed approach achieved equal or better results than the baseline. Remarkably enough, the fusion of both approaches in a speaker-dependent framework yields performance improvements by a factor of 17.4% or 19.7% respectively for Actual and Simulated subcorpus. We also conclude that by removing speaker inter-variability the system performance significantly improves. The relative improvement is about 30.64% when it is evaluated in Actual subcorpus and about 43.21% for Simulated subcorpus.

Table 6: *Comparation between speaker independent and speaker dependent experiments*

Subcorpus	Approach	Spk. Ind.	Spk. Dep.	R.I.%
Actual	Baseline	30.13	23.16	+23.13
	Proposed	29.94	19.70	+34.20
	Combined	26.82	18.60	+30.64
Simulated	Baseline	28.89	20.47	+29.14
	Proposed	36.04	17.0	+52.83
	Combined	29.78	16.91	+43.21

The use of new improved configurations for pitch continuous estimation will be addressed in future work as well as the combination of prosodic and acoustic level of features.

6. References

[1] Rosalind W. Picard, *Affective Computing*, The MIT Press, September 1997.

[2] L.C. De Silva, T. Miyasato, and R. Nakatsu, "Facial emotion recognition using multi-modal information", Sep 1997, vol. 1, pp. 397–401 vol.1.

[3] Björn Schuller, Stefan Steidl, and Anton Batliner, "The interspeech 2009 emotion challenge", 2009.

[4] Zhihong Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions", *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 1, pp. 39–58, 2009.

[5] D. Ramos, J. Gonzalez-Rodriguez, J. Gonzalez-Dominguez, and J. J. Lucena-Molina, "Addressing database mismatch in forensic speaker recognition with ahumada iii: a public real-case database in spanish", in *Proceedings of Interspeech 2008*, September 2008, pp. 1493–1496.

[6] J.H.L. Hansen and S.E. Bou-Ghazale, "Getting started with susas: a speech under simulated and actual stress database", in *EUROSPEECH-1997*, 1997, pp. 1743–1746.

[7] J.H.L. Hansen and S. Patil, "Speech under stress: Analysis, modeling and recognition", in *Speaker Classification (1)*. 2007, vol. 4343 of *Lecture Notes in Computer Science*, pp. 108–137, Springer.

[8] Oh-Wook Kwon, Kwokleung Chan, Jiucang Hao, and Te-Won Lee, "Emotion recognition by speech signals", in *EUROSPEECH-2003*, 2003, pp. 125–128.

[9] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 5.1.04) [computer program]", Ap 2009, http://www.praat.org/.

[10] Hao Hu, Ming-Xing Xu, and Wei Wu, "Gmm supervector based svm with spectral features for speech emotion recognition", in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, 2007, vol. 4, pp. IV–413–IV–416.