Anchor Model Fusion for Emotion Recognition in Speech

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

ATVS - Biometric Recognition Group, Universidad Autonoma de Madrid, Spain
carlos.ortego@estudiante.uam.es, ignacio.lopez@uam.es

Abstract. In this work, a novel method for system fusion in emotion recognition for speech is presented. The proposed approach, namely Anchor Model Fusion (AMF), exploits the characteristic behaviour of the scores of a speech utterance among different emotion models, by a mapping to a back-end *anchor-model* feature space followed by a SVM classifier. Experiments are presented in three different databases: *Ahumada III*, with speech obtained from real forensic cases; and *SUSAS Actual* and *SUSAS Simulated*. Results comparing AMF with a simple sum-fusion scheme after normalization show a significant performance improvement of the proposed technique for two of the three experimental set-ups, without degrading performance in the third one.

Key words: emotion recognition, anchor models, prosodic features, GMM supervectors, SVM.

1 Introduction

There is an increasing interest in the automatic recognition of emotional states in a speech signal, mainly due to its applications to human-machine interaction applications [1] [2]. As a result, a wide range of different algorithms for emotion recognition have emerged. This fact motivates the use of fusion schemes in order to improve the performance of system by the combination of different approaches.

It is common for this task to be stated as a multiclass classification problem. However, emotion recognition can also be stated as a verification or detection problem. In such case, given an utterance x and a target emotion e, for which an emotion model m_e from a set M is trained. The objective is to determine whether the dominant emotion that affect the speaker in the utterance is e (*target* class) or any other (*non-target* class). In such a scheme, for any model $m_e \in M$ and utterance x, a similarity score denoted as s_{x,m_e} can be computed. Detection is performed by comparing s_{x,m_e} to a threshold, which is generally set according to the minimization of some cost function.

It is important to remark that the behaviour of the scores of a given utterance from a given emotion is different and characteristic for each model in M. Therefore, it is expected that the detection of the emotion in x will benefit not only of the target scores from their comparisons with m_e, but also from the

scores of x compared to the rest of models in M. This motivates a two-level architecture, where models $m_j \in M$, $j \in \{1, .., N_{fe}\}$ are denoted as front-end models in opposition to back-end models which are trained in advance using scores, such as s_{x,m_j}, as feature vectors. This nomenclature has been adopted form language recognition [3], which is a similar problem.

This work proposes a novel back-end approach for the fusion of the information obtained by N_{sys} different emotion detectors. It is based on *anchor models fusion* (AMF) [4], which uses the information of the relationship among all the models in M for improving detection performance. Results presented validate the proposed approach based on a experimental set-up in substantially different databases: *Ahumada III* (speech from real forensic cases) [5], *SUSAS Simulated* and *SUSAS Actual* [6]. AMF have been used to combine scores from two prosodic emotion recognition systems denoted as GMM-SVM and statistics-SVM. Performance results will be measured in terms of equal error rate (EER) and its average among emotions.

This work is organized as follows. The anchor models feature space is described in Section 2. In Section 3, the proposed AMF method is described in detail. Section 4 describes front-end systems implemented as well as the prosodic feature extraction. The experimental work which shows the adequacy of the approach is shown in 5. Finally, conclusions are drawn in Section 6.

2 Anchor models feature space

Given a speech utterance x from an unknown spoken emotion, and a front-end emotion recognition system with N_{fe} target emotion models $m_j \in M$, $j \in \{1, .., N_{fe}\}$, a similarity score s_{x,m_j}, can be obtained as a result of comparing x against each emotion model m_j. Thus, for every utterance x we obtain a N_{fe} dimensional vector $\bar{S}_{x,M} = [s_{x,m_1} \cdots s_{\mathbf{x},m_N}]$ that stacks all possible scores for x. This scheme defines a derived similarity feature space known as *anchor model* space [4] where every utterance x can be mapped. In this new feature space any classifier can be trained in order to discriminate any given emotion in utterance x with respect to the rest, by learning the relative behaviour of the scores of speech utterance x with respect to the models in M. An example of this relative behaviour is shown in figure 1 where utterances from four emotions (*angry*, *question*, *neutral*, *stressed*,) are compared with two different cohorts M of anchor models.

3 Anchor Model Fusion (AMF) back-end

AMF is a data-driven approach that have shown a satisfactory performance in language recognition [7]. In AMF, the cohort of models M is built by including all the available models from the N_{sys} emotion recognition systems in the front-end. The resulting vector of scores for utterance x, denoted as $\bar{S}_{AM}$, stacks the N_{sys} values of $\bar{S}^j_{x,M}$ over all emotion recognition system j in the front-end.

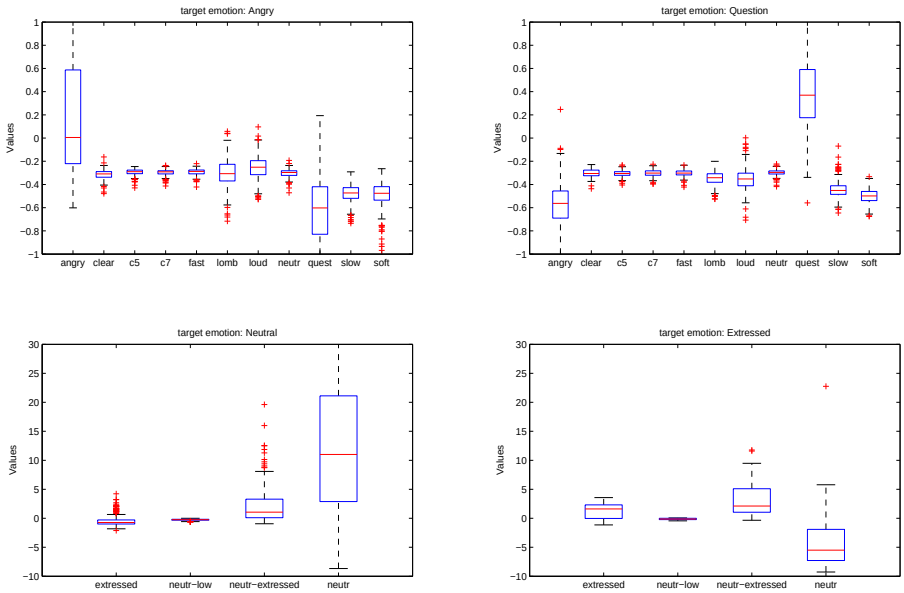

Fig. 1. *At the top, the distribution of* angry *(left) and* question *(right) utterances over the a set M formed by the emotion models in SUSAS Simulated speech. At the bottom, the distribution range of* neutral *(left) and* stressed *(right) utterances over the a set M form by the emotion models in Ahumada III*

.

$$S_{AM}^{-}(x, M) = \left[\bar{S}_{x,M}^{1}, \cdots, \bar{S}_{x,M}^{N_{sys}}\right] \tag{1}$$

Fig. 2 illustrates the process in which $S_{AM}(x, M)$ is obtained by projecting x into the anchor model space defined by M. Hence, the number of dimensions of anchor model space is $d = \sum_{j=1}^{N_{sys}} N_j$, where N_j is the number of models in the front-end system j. At this point, the objective is to boost the probability of finding a characteristic behaviour of the speech pattern in the anchor model space, by increasing d and with the limits of the *curse of dimensionality*. This objective can be achieved by including more anchor models and/or systems to fuse.

It is important to note that any emotion can be trained in the anchor model space, not only those in M. These so-called back-end emotional set M' will be the actual set of target emotions. Thus, once every testing utterance is projected over the anchor model space, any classifier can be used for training any back-end emotions in M'. In this work, SVM were applied due to its robustness while the dimension of the anchor model space increases.

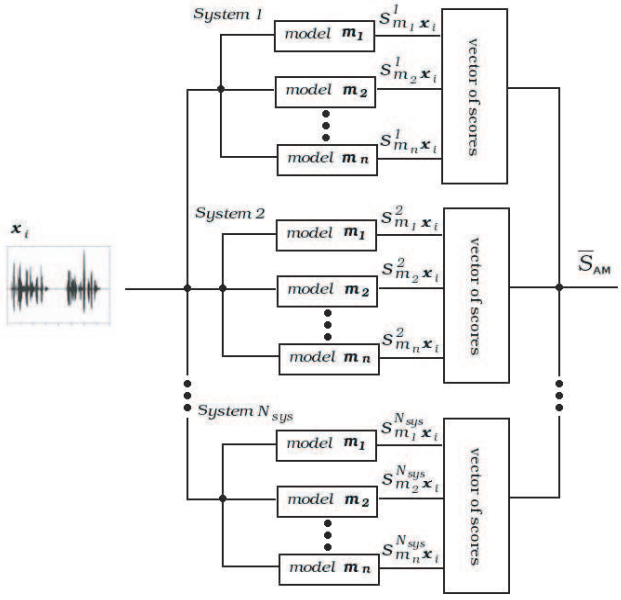

Fig. 2. *Diagram of generation of features in the AMF space.* $\bar{S}_{AM}(x, M)$ *stacks the scores of* $\boldsymbol{x}_i$ *over the set of models* m_j^l*, for all languages* j *and all subsystems* l

.

4 Emotion recognition systems front-end

This section details the prosodic features extracted from the audio signal, and used as input vectors for both front-end systems implemented. Subsections 4.2 and 4.3 describes in more detail their implementation.

4.1 Prosodic features for emotion recognition

Prosodic features are often considered as input signals for emotion recognition systems due to their relation with the emotional state information [8]. In this work prosodic features consist of a set of $d = 4$ dimensional vectors with the sort-term coefficients of energy; the logarithm of the pitch; and their velocity coefficients, also known as Δ features. These coefficients are extracted only from voiced segments with an energy value higher than the 90% of the dynamic range. Mean normalization have been used for energy and Δ-energy coefficients. Pitch and energy have been computed using Praat [9].

4.2 prosodic GMM-SVM

Previous works have shown the excellent performance of SVM-GMM supervectors in the tasks of language and speaker recognition, while the application of

this technique to the prosodic level of the speech were firstly introduced in [10]. This technique can be seen as a secondary parametrization capable to summarize the distribution of the feature vectors in x, into a single high-dimensionality vector. This high-dimensionality vector is known as a GMM supervector. In order to build the GMM supervector, first the prosodic vectors of x are used to train a M-mixtures GMM model λ_x. This model is obtained from a Universal Background Model (UBM) λ_{UBM} using Maximum-A-Posteriori (MAP) adaptation of means. The GMM supervector of the utterance x is the concatenation of the M vectors of means in λ_x.

GMM supervector are often considered as kernel functions $\mu(x)$ that maps prosodic features from dimension of d into a high-dimensional feature space of size $L' = M * d$. Once every utterance is mapped into this L'-dimensional supervector space, linear SVM models are used to train the front-end emotion models. Therefore, any m_j is a L'-dimensional vector that represent an hyperplane that optimally separates supervectors of utterances form the target emotion j with respect to supervectors from other emotions.

4.3 prosodic statistics-SVM

This scheme is based on a previous work presented in [11]. It consist of a statistical analysis of each prosodic coefficient followed by a SVM. The distribution of the prosodic values is characterized by computing $n = 9$ statistical coefficients per feature (table 1). Once every utterance is mapped into this derived feature space of dimension $L = d * n$, front-end emotions models are obtained as linear one-vs-all SVM models.

Table 1. *Statistical coefficients extracted for every prosodic stream form each utterance in the statistics-SVM approach.*

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

A test-normalization scheme has been used for score normalization prior to AMF. First, the scores distribution for every testing utterance x with respect to M has been estimated assuming Gaussianity. The values of mean and variance of this distribution are then used to normalize the similarity scores of x over any model m.

5 Experiments

5.1 Databases

The proposed emotion recognition system has been tested over Ahumada III and SUSAS (Speech Under Simulated And Actual Stress) databases. Ahumada III consists of real forensic cases recorded by the Spanish police forces (*Guardia Civil*) and authored by the Spanish law under confidence agreements. It includes speech from 69 speakers and 4 emotional states (*neutral*, *neutral-low*, *neutral-stressed*, *stressed*) with 150 seconds training utterances and testing utterances among 10 and 5 seconds long. SUSAS database is divided in two subcorpora from simulated and real spoken emotions. SUSAS Simulated subcorpus 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 subcorpus 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, each one with 2 realization, for every speaker and speaking style.

5.2 Results

The GMM-SVM front-end system requires a set of development data for building the model λ_{UBM}. Therefore every database were split in two different non-overlapping sets. The first one was used for training a M=256 mixtures GMM UBM λ_{UBM}. The second set were used for implementing a double 10 folds cross-validation scheme: first cross-validation stage is for training and testing front-end models, while back-end models are trained and tested during the second one.

AMF cohort M is built with models from all databases and systems. Therefore, for each front-end system we obatined 4 models from Ahumada III corpus, 11 models from SUSAS Simulated corpus and 5 models from SUSAS Actual corpus. A third system is included as the sum fusion of both front-end systems. Thus, this scheme leads to a AMF space of $(4 + 11 + 5) \times 3 = 60$ dimensions.

In order to compare AMF with a *baseline* fusion technique we performed a standard sum fusion between the scores of GMM-SVM and statistics-SVM systems. Note that sum fusion outperforms the results obtained from any of both front-end systems individually.

Tables 2 and 3 summarize the results of the proposed approach. It can be seen that the average EER for all the emotions in *Ahumada III* and *SUSAS Simulated* respectively improves 15.52% and 18.61%. Remarkable good results are obtained for *neutral-low*, *loud* and *fast* emotion models while for models *scream* and *angry* a significant loss of performance is obtained, probably due to non-modeled variability factors such as the speaker identity. The results for *SUSAS Actual* shows neither improvement not degradation in the average EER. This can be due to the enbiromental conditions of SUSAS Actual corpus (amusement park roller-coaster, and helicopter cockpit recordings). Under such conditions,

Table 2. AMF performance improvement vs. sum fusion for the systems *Ahumada III*. Results are presented in EER(%) and its relative improvement (R.I.).

AhumadaIII			
Emotion	Baseline	AMF	R.I. %
neutral-low	50.21	30.02	-40.21
neutral	33.77	33.92	0.44
neutral-stressed	38.12	33.22	-12.85
stressed	28.69	25.7	-10.42
Avg. EER	37.7	30.72	-18.51

Table 3. AMF performance improvement vs. sum fusion for *SUSAS Simulated* (*a*) and *SUSAS Actua*l (*b*).

SUSAS Simulated			
Emotion	Baseline	AMF	R.I. %
angry	22.93	32.76	42.87
clear	42.91	41.89	-2.38
cond50	41.01	33.57	-18.14
cond70	48.3	30.55	-36.75
fast	30.21	16.81	-44.36
lombard	34.85	38.65	10.9
loud	27.65	13.2	-52.26
neutral	40.53	35.31	-12.88
question	3.86	3.52	-8.81
slow	26.75	20.35	-23.93
soft	22.07	22.54	2.13
Avg. EER	31.01	26.29	-15.22

(a)

SUSAS Actual			
Emotion	Baseline	AMF	R.I. %
neutral	36.54	35.26	-3.5
medst	46.95	50.08	6.67
hist	42.57	39.14	-8.06
freefall	25.86	24.66	-4.64
scream	11.15	14.6	30.94
Avg. EER	32.61	32.75	0.43

(b)

noise patterns can characterististically affect scores in such way that AMF can not improve front-end results.

6 Conclusions

This work introduces Anchor Model Fusion (AMF), a novel approach for fusion of systems in emotion recognition. The approach is based on the anchor model space which maps scores from the so-called *front-end* detectors to a different *back-end* feature space where they can be classified by an SVM. Therefore back-end emotion models M' are supported over the set of front-end models M, which may be trained with different emotions, databases, recording conditions, etc. In this work the proposed AMF approach have been used for fusing two different prosodic emotion recognition systems as well as a third one obtained as the result of the sum fusion of both systems. Thus M have been built with 3 systems, each one with 20 front-end models, leading to a 60-dimensions AMF space. Experiments have been carried out over three corpora: *Ahumada III* (speech

from real forensic cases), *SUSAS Simulated* (speech with acted emotions) and *SUSAS Actual* (speech with actual emotions). Results of the proposed AMF scheme are compared with the sum fusion of both front-end systems, showing a EER relative improvement larger than the 15% for *Ahumada III* and *SUSAS Simulated* corpora.

Future work will focus on the optimal selection of front-end models M, normalization techniques of the anchor-model space vectors and new classification methods for the back-end such as Linear Discriminant Analysis.

7 Ackowledgements

This work has been financed under project TEC2006-13170-C02-01.

References

1. Ververidisa, D., Kotropoulos, C.: Emotional speech recognition: Resources, features, and methods. Speech Communication (9) (September 2006) 1162–1181
2. Picard, R.W.: Affective Computing. The MIT Press (September 1997)
3. Ramabadran, T., Meunier, J., Jasiuk, M., Kushner, B.: Enhancing distributed speech recognition with back-end speech reconstruction. In: Proceedings of Eurospeech 2001. (2001) 1859–1862
4. Collet, M., Mami, Y., Charlet, D., Bimbot, F.: Probabilistic anchor models approach for speaker verification. (2005) 2005–2008
5. Ramos, D., Gonzalez-Rodriguez, J., Gonzalez-Dominguez, J., Lucena-Molina, J.J.: Addressing database mismatch in forensic speaker recognition with ahumada iii: a public real-case database in spanish. In: Proceedings of Interspeech 2008. (September 2008) 1493–1496
6. Hansen, J., Sahar, E.: Getting started with susas: a speech under simulated and actual stress database. In: Proceedings of Eurospeech 1997. 1743–1746
7. Lopez-Moreno, I., Ramos, D., Gonzalez-Rodriguez, J., Toledano, D.T.: Anchor-model fusion for language recognition. In: Proceedings of Interspeech 2008. (September 2008)
8. Hansen, J., Patil, S.: Speech under stress: Analysis, modeling and recognition. In: Speaker Classification (1). Volume 4343 of Lecture Notes in Computer Science., Springer (2007) 108–137
9. Boersma, P., Weenink, D.: Praat: doing phonetics by computer (version 5.1.04) [computer program] (Ap 2009) http://www.praat.org/.
10. Hu, H., Xu, M.X., Wu, W.: Gmm supervector based svm with spectral features for speech emotion recognition. In: Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on. Volume 4. (2007) IV–413–IV–416
11. Kwon, O.W., Chan, K., Hao, J., Lee, T.W.: Emotion recognition by speech signals. In: EUROSPEECH-2003. (2003) 125–128