INTERSPEECH 2009 BRIGHTON

Cepstral and Long-Term Features for Emotion Recognition

Pierre Dumouchel[1,3], *Najim Dehak*[1,3], *Yazid Attabi*[1,3], *Reda Dehak*[2], *Narjès Boufaden*[3]

[1]École de Technologie Supérieure (ETS), Montréal, Canada
[2]Laboratoire de Recherche et de Développement de l'EPITA (LRDE), Paris, France
[3]Centre de recherche informatique de Montréal (CRIM), Montréal, Canada
{pierre.dumouchel,najim.dehak,Yazid.Attabi,Narjes.Boufaden}@crim.ca
reda.dehak@lrde.epita.fr

10.21437/Interspeech.2009-111

Abstract

In this paper, we describe systems that were developed for the Open Performance Sub-Challenge of the INTERSPEECH 2009 Emotion Challenge. We participate in both two-class and five-class emotion detection. For the two-class problem, the best performance is obtained by logistic regression fusion of three systems. These systems use short- and long-term speech features. Fusion allowed to an absolute improvement of 2.6% on the unweighted recall value compared with [1]. For the five-class problem, we submitted two individual systems: cepstral GMM vs. long-term GMM-UBM. The best result comes from a cepstral GMM and produces an absolute improvement of 3.5% compared to [6].

Index Terms: Gaussian Mixture Models, Support Vector Machines, Logistic Regression, Pseudo-syllable, Legendre Polynomial.

1. Introduction

Emotion recognition is viewed here as a speech processing problem which consists to recognize the emotion state of a speaker from speech signal. Prosodic features based on pitch and energy contour statistics are the most widely used parameters in this field [2] [3]. Recently, modeling the variation of the prosodic contours seems to be helpful to separate between emotions. Furthermore, vocal tract features were successfully applied in order to recognize the speaker emotion state [4].

In this paper, we present our emotion recognition systems that were developed in the context of the INTERSPEECH 2009 Emotion Challenge [1]. We participate in the Open Performance Sub-Challenge which is divided in two tasks: the two-class emotion problem and the five-class emotion problem. The emotion categories for the two-class problem are idle (IDL) and negative (NEG) emotions. Emotion categories for the five-class problem are Angry, Emphatic, Neutral, Positive and Rest. We propose to fuse several systems which operate on short-term cepstral features and long-term prosodic and vocal tract features. These features are modeled by Gaussian mixture models and support vector machines.

The outline of the paper is as follows. Section 2 describes two short-term cepstral feature systems. In Section 3, we present a system based on long-term prosodic and vocal tract features. Section 4 introduces score fusion techniques used to combine systems developed during this competition. The experiments and results are presented in section 5. Concluding remarks are presented in Section 6.

2. Short-term cepstral systems

Two systems are carried out in order to operate in the short-term features. The first system (S1) is based on the classical Gaussian mixture models and the second one (S2) consists on a fusion between Gaussian mixture models and support vector machines.

2.1. Feature extraction

Twelve (12) Mel Frequency Cepstral Coefficients (MFCC) together with log energy are extracted using a $25ms$ Hamming window with a frame advance of $10ms$. Delta and double delta coefficients were then computed using a 5 frame window to produce a 39-dimensional feature vectors. Note that silences are removed from the wave files before extracting the MFCC.

2.2. Gaussian mixture models

The Gaussian Mixture Model (GMM) is a generative model widely used in field of speech processing [5] [6] [7] [4]. It is a semi-parametric probabilistic method that offers the advantage of adequately representing speech signal variability. Given a GMM, the probability of observing a feature vector x is computed by the following equation:

$$P\left(x|\lambda\right) = \sum_{i=1}^{C} w_i \mathcal{N}\left(x;\, \mu_i, \Sigma_i\right) \tag{1}$$

where C, w_i, μ_i and Σ_i correspond respectively to the number of Gaussians, the weight, the mean vector and the covariance matrix of the i^{th} Gaussian. We use diagonal covariance matrices. Full covariance matrices are not really necessary even if the features are not statistically independent which is the case for the MFCC parameters. The effect of using a set of full-covariance GMM can be equally represented by a GMM built on larger set of diagonal covariances. The GMM parameters are estimated using Maximum Likelihood (ML) approach which is the Expectation Maximization (EM) algorithm [8].

The classification of a test frame sequence $X = \{x_1, x_2, ..., x_T\}$ is based on the Bayes decision. Using an equal prior of all classes, this classification is obtained by computing directly the log-likelihood of the test recording given each GMM associated to each emotion class. This test recording is classified in the emotion class $\widetilde{E}$ that produces the best log-likelihood value as described in the following equation:

$$\widetilde{E} = \arg\max_{i=1,...m} \log P\left(X|\lambda_i\right) \tag{2}$$

where m is the emotion class and $\log P(X|\lambda_i)$ is the log-likelihood of the test sequence X given a GMM λ_i.

2.3. Support vector machines and Gaussian mixture models

The Support Vector Machine (SVM) is a binary classifier used to find a separator between two classes. The main idea of this classifier is to project the input vectors in high dimensional space called feature space in order to find linear separation. This projection is carried out using a mapping function. In practice, SVMs use kernel functions to perform the scalar product computation in the feature space. These functions allow us to compute directly the scalar product in the feature space without defining the mapping function. An SVM discriminant function is given by

$$f(x) = \sum_{i=1}^{M} \alpha_i \, y_i \, k(x, x_i) + b \tag{3}$$

where $k(x, x_i)$ is the kernel function; the x_i's are the support vectors and the y_i's are the corresponding class label ± 1; M is the number of support vectors. The α_i's and b are obtained during training process.

We have experimented a system based on the combination of SVMs with GMMs based on the Universal Background Model (UBM) and Maximum A Posteriori (MAP) adaptation [7]. This system is designed for the two-class emotion problem of the Open Performance Sub-Challenge [1]. The UBM is a large GMM trained by pooling all the MFCC frames of all classes together in order to train only one GMM. This model plays the role of the prior in the MAP adaptation.

The approach that we propose for combining SVM and the GMM uses the GMM supervector as input for the SVM. The GMM supervector is a high dimensional vector comprised by the concatenation of all Gaussian means. The GMMs are obtained by adapting only the UBM component means to each recording file [6]; we keep the weights and covariance matrices unchanged. At the end, we have many GMMs for each emotion class. The SVM is than applied in the GMM space in order to separate between this two emotion classes. This SVM is based on the linear Kullback-Leibler kernel defined in [9]. This kernel was introduced by Campbell *et al.* [9]. It is derived from the Kullback-Leibler distance between two GMMs [6] [9]. This distance corresponds to the Euclidean distance between scaled GMM supervectors e^a and e^b.

$$D_e^2\left(e^a, e^b\right) = \sum_{i=1}^{C} w_i \left(\mu_i^a - \mu_i^b\right)^t \Sigma_i^{-1} \left(\mu_i^a - \mu_i^b\right) \tag{4}$$

where w_i and Σ_i are the i^{th} UBM mixture weights and diagonal covariance matrix, μ_i^a corresponds to the mean of Gaussian i of the GMM recording a. The derived linear kernel is defined as the corresponding inner product of the preceding distance

$$k_{lin}\left(e^a, e^b\right) = \sum_{i=1}^{C} \left(\sqrt{w_i}\Sigma_i^{-1/2}\mu_i^a\right)^t \left(\sqrt{w_i}\Sigma_i^{-1/2}\mu_i^b\right) \tag{5}$$

3. Long-term prosodic and vocal tract system

3.1. Feature extraction

We extracted the pitch, logarithm of energy and the first two formant values computed at $10ms$ intervals with Praat package [10]. Pitch is extracted with the autocorrelation method proposed in [10] and is undefined in unvoiced regions. We used only the voiced part of the speech signal in our modeling. Log energy is normalized on an utterance basis by subtracting the maximum value for the whole utterance.

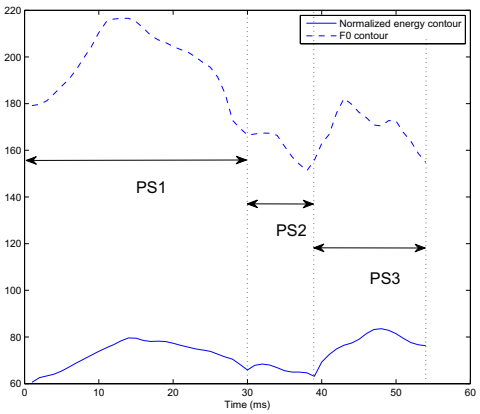

Figure 1: Example of voiced region contour segmentation on pseudo-syllable units.

We will describe now how the pitch and energy contours (containing more than one syllable) are segmented into pseudo-syllables based on unsupervised segmentation according to the energy contour only.

3.1.1. Segmentation:

In order to model prosodic contours based on the pseudo-syllable as a unit, we segment the long prosodic contours into syllable-like regions in the same way as in [5] [6]. This method is based on detecting valley points of the energy of voiced speech contour. In general, these valley points serve as segment boundaries; however we impose a minimum duration constraint of $60ms$. This enables us to calculate six term Legendre polynomial expansions.

We will show in the next paragraph how the pitch and energy contours (based on pseudo-syllable units) are approximated by Legendre polynomials.

3.1.2. Approximation and time normalization:

For each generated segment, we carry out an approximation of the pitch and energy contour by taking the M leading terms in a Legendre polynomial expansion. That is, each contour $f(t)$ is approximated as:

$$f(t) = \sum_{i=0}^{M} a_i \, P_i(t) \tag{6}$$

where t represents time, $P_i(t)$ is the i^{th} Legendre polynomial, and $M = 5$ in our implementation. Each coefficient models a particular aspect of the contour. For example a_0 is interpreted as mean of the segment, a_1 is the slope, a_2 gives information about the curvature of the segment, and a_3, a_4, a_5 model the fine details. However, it is important to carry out time normalization of these coefficients in order to be comparable across segments. All segments are scaled and mapped onto the same $[-1, +1]$ interval. This technique of prosodic contour approximation has

been successfully applied in speaker verification [5], language recognition [6] and recently in emotion recognition [4].

3.2. Long-term feature modeling

In the context of this challenge, we carried out two separate long-term systems based on the GMM approach. The first system (S3) is used in the case of two-class emotion problem [1]. In this modeling, we trained a GMM for each emotion class using an Expectation Maximization algorithm. For the five-class emotion problem [1], the system (S5) starts by training an UBM using all emotion data. Then we adapt the UBM component means in each emotion data separately using MAP adaptation. The reason for using MAP adaptation rather than maximum likelihood (ML) training in this situation is that when we used the pseudo-syllable as unit, we had few vectors in order to train the GMM. The MAP adaptation is more appropriate in this training condition compared to ML.

4. Fusion

We carried out a linear score fusion based on the logistic regression in the case of the two-class emotion problem. The linear score fusion technique consists in combining scores of several subsystems into a single definitive score. In order to carry out this combination, we evaluated the output of each sub-system based on the following score:

$$s = \log \frac{P(\lambda_{IDL}|X)}{P(\lambda_{NEG}|X)} \quad (7)$$

$$= \log \frac{P(X|\lambda_{IDL})}{P(X|\lambda_{NEG})} + \log \frac{P_{IDL}}{P_{NEG}} \quad (8)$$

where λ_{IDL} and λ_{NEG} are respectively the model of IDL and NEG emotions. We used an equal a priori probability for each class which simplifies this score to:

$$s = \log \frac{P(X|\lambda_{IDL})}{P(X|\lambda_{NEG})} \quad (9)$$

This score is compared to the threshold which is equal to zero in order to take the final decision. If the score is greater or equal to zero we classify the test recording as IDL otherwise we classify it as NEG. The linear score fusion is evaluated as follows:

$$s_f(w) = w_0 + \sum_{l=1}^{S} w_l s_l \quad (10)$$

where s_f is the final fused score, s_l is the score for the l^{th} subsystem, S is the number of fused subsystems and $w = (w_0, w_1, ..., w_S)$ is a vector comprised of the score fusion weights. We used the logistic regression in order to estimate the score fusion weights [11]. This approach is based on a supervised training which requires a set of labeled scores for each sub-system in the same corpus. The score labels correspond to idle and negative emotion files. Greater details about the objective function that needs to be optimized are given in [11] [12].

5. Experiment

5.1. Database

Our experiments are carried out in the FAU AIBO Emotion corpus. This database is divided into training and test sets. Both sets contain respectively 9959 and 8257 children voice recordings.

5.2. Experimental set-up

In order to carry out a feature selection of the long-term prosodic and vocal tract features and to find the best GMM configuration of the all systems, we carried out nine cross-validations in the training data set of the FAU AIBO Emotion corpus. For each training partition, we have isolated the data of three different children identities. The reason for this cross-validation process is to assure speaker-independency of our systems since the test speaker never occurs in the training partition data.

5.3. Results

The results are presented regarding the Open Performance Sub-Challenge tasks of the 2009 Emotion Challenge. Two metrics are proposed to evaluate the performance of systems: weighted and unweighted average recalls. As in [1], more importance is given to unweighted average recall values.

5.3.1. Two-class emotion problem

In the context of two-class emotion problem, the best performance of the long-term feature system (S3) is obtained using the first three Legendre polynomial coefficients for the pitch and energy and means of the first two formants. These features are then modeled using a GMM comprised with 256 Gaussians. For the cepstral GMM, the optimal GMM configuration uses a mix of 512 Gaussians for each emotion class. We used the same number of Gaussians in the case of the combination of the SVM and the Cepstral GMM.

We carried out four score fusion experiments. In the first fusion experiment (I), we used the logistic regression in order to combine scores of Cepstral GMM and long-term feature GMM. The fusion weights are estimated based on the cross-validation carried out in the training part of the corpus as explain above. The second fusion experiment (II) is based on the fusion weights obtained in experiment I and consists of increasing manually fusion weight associated with Cepstral GMM in order to increase its importance. Note that all these weights are estimated in the training set of FAU AIBO corpus before any test submission. In other words, no system parameters are trained with the test set of FAU AIBO corpus. The third fusion experiment (III) is based on logistic regression of the three systems described above (S1, S2 and, S3). Similar to experiment (II), we increase the weight associated with cepstral GMM system for experiment (IV). The results of all these individual and fused systems with the test data set are given in the Table 1.

Three remarks come from the analysis of results in Table 1. First, cepstral GMM achieves the best overall unweighted average recall. Secondly, the long-term GMM produces the best weighted average recall among our individual system but is outperformed by the 1-state GMM state-of-the-art one [1]. Finally, the best unweighted average recall is obtained by the fusion of our three systems based on the logistic regression training. This fusion achieved an absolute improvement of 2.60% compared with [1].

5.3.2. Five-class emotion problem

For the five-class emotion problem, we submitted two individual systems: cepstral GMM (S4) vs. long-term GMM-UBM (S5). For the cepstral GMM, we used the same feature as system S1 as well as the same number of Gaussian components (512). The long-term feature GMM system S5 uses the same feature as the previous one (S3). Althought it differs in the number

Table 1: Performance for the two-class emotion problem.

System	**Unweighted recall**	Weighted recall
State-of-the-art [1]: 1 state	62.7%	72.6%
3 states	67.6%	68.3%
5 states	67.7%	65.5%
S1: GMM w/ MFCC	**69.72%**	68.03%
S2: SVM-GMM w/ MFCC	60.91%	64.72%
S3: GMM w/ long-term feat.	66.61%	**70.84%**
Experiment I : S1+S3	69.94%	68.67%
Experiment II : S1+S3	70.12%	69.53%
Experiment III : S1+S2+S3	**70.29%**	68.68%
Experiment IV : S1+S2+S3	70.14%	70.39%

Table 2: Performance for the five-class emotion problem.

System	**Unweighted recall**	Weighted recall
State-of-the-art [1]: 1 state	35.5%	50.8%
3 states	35.2%	34.7%
5 states	35.9%	37.2%
S4: GMM w/ MFCC	**39.40%**	52.08%
S5: GMM-UBM w/ long-term feat.	36.91%	36.87%

of Gaussian (64 instead of 256) and in GMM training by using MAP adaptation based on UBM rather than ML estimation. Performance results are given in Table 2.

Table 2 shows that both S4 and S5 systems outperform the state-of-the-art ones based on unweighted average recall. The best result comes from a Cepstral GMM system and offers an absolute improvement of 3.5% compared to [1].

6. Conclusion

For the Open Performance Sub-Challenge of the INTERSPEECH Emotion Challenge with two-class emotion problem, the best individual system uses GMM with short-term cepstral features and achieves an unweighted average recall of 69.72%. By fusing systems with short- and long-term features, we improve the performance and reach an unweighted average recall of 70.29%. Both individual and fused systems outperform the state-of-the-art performance of 67.7%.

For the five-class emotion problem, two individual systems based on GMM modeling were proposed and differ mainly on feature type: short vs. long term features. The cepstral GMM and the long-term feature GMM obtain unweigted average recalls of 39.4% and 36.9% respectively compared to 35.9% for the state-of-the-art.

7. Acknowledgements

This work was founded in part by the Canadian Heritage Funds for New Media Research Network and Natural Sciences and Engineering Research Council of Canada.

8. References

[1] B. Schuller, S. Steidl, and A. Batliner, "The Interspeech 2009 Emotion Challenge," in *Interspeech*. Brighton, UK: ISCA, 2009.

[2] S. Yacoub, S. Simske, L. Xiaofan, and J. Burns, "Recognition of Emotions in Interactive Voice Response Systems," in *EUROSPEECH*, 2003, pp. 729–732.

[3] J. Ang, R. Dhillon, A. Krupsky, E. Shriberg, and A. Stolcke, "Prosody Based Automatic Detection of Annoyance and Frustration in Human Computer Dialog," in *Conference ICSLP*, 2002.

[4] Y. Attabi, "Reconnaissance Automatique des motions Partir du Signal Acoustique," Master's thesis, cole de technologie suprieure (TS), Montral, 2009.

[5] N. Dehak, P. Dumouchel, and P. Kenny, "Modeling Prosodic Features With Joint Factor Analysis For Speaker Verification," *IEEE Transaction on Speech and Signal Processing*, vol. 15, no. 7, pp. 2095–2103, Sept 2007.

[6] C.-Y. Lin and H-C.Wang, "Language Identification Using Pitch Contour Information," in *ICASSP*, 2005, pp. 601–604.

[7] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker Verification using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[8] A. Dempster, N. Laird, and D. Robin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, vol. B, pp. 1–38, 1997.

[9] W. M. Campbell, D. E. Sturim, D. Reynolds, and A. Solomonoff, "Support Vector Machines Using GMM Supervectors for Speaker Verification," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308–311, May 2006.

[10] P. Boersma and D. Weenink, "Praat: Doing Phonetics by Computer (version 5.0.32)," 2008. [Online]. Available: http://www.praat.org/

[11] N. Brummer, L. Burget, J. Cernocky, O. Glembek, F. Grezl, M. Karaat, D. V. Leeuwen, P. Matejka, P. Schwarz, and A. Strasheim, "Fusion of Heterogeneous Speaker Recognition Systems in the STBU Submission for the NIST Speaker Recognition Evaluation 2006," *IEEE Transaction On Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 2072–2084, september 2007.

[12] D. V. Leeuwen and N. Brummer, "An Introduction to Application-Independent Evaluation of Speaker Recognition Systems," in *Speaker Classification I: Fundamentals, Features, and Methods*. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 330–353.