The 2016 Conference on Computational Linguistics and Speech Processing
ROCLING 2016, pp. 142-152

Support Super-Vector Machines in Automatic Speech Emotion Recognition

陳嘉穎 Chia-Ying Chen
國立中山大學資訊工程學系
Department of Computer Science and Information Engineering
National Sun Yat-sen University
m033040029@student.nsysu.edu.tw

陳嘉平 Chia-Ping Chen
國立中山大學資訊工程學系
Department of Computer Science and Information Engineering
National Sun Yat-sen University
cpchen@mail.cse.nsysu.edu.tw

Abstract

In this paper, we use super-vectors in support vector machines for automatic speech emotion recognition. In our implementation, an utterance is converted to a super-vector formed by the mean vectors of a Gaussian mixture model adapted from a universal background model. The proposed method is evaluated on FAU-Aibo database which is well-known to be used in INTERSPEECH 2009 Emotion Challenge. In the case of HMM-based dynamic modeling classifier, we achieve an unweighted average (UA) recall rate of 40.0%, over a baseline of 35.5%, by using the delta features and increasing the number of mixture components. In the case of SVM-based static modeling classifier, we achieve an unweighted average (UA) recall rate of 38.9%, over a baseline of 38.2%, by using the proposed super-vectors.

Keywords：Speech Emotion Recognition, GMM, Super-vector, SVM

1 Introduction

Speech emotion recognition (SER) becomes very popular in recent years [1]. The INTERSPEECH 2009 Emotion Challenge [2] (henceforth referred to as the Challenge) is a large-scale evaluation plan of SER techniques on FAU-Aibo corpus. In the Challenge, the training set and the test set are defined so fair comparison can be carried out. There are 2 classification models, namely the dynamic modeling of hidden Markov model (HMM) on low-level descriptors (LLDs) and the static modeling of support vector machine (SVM) on supra-segmental feature vectors, which are functional values of sequences of LLDs.

In this paper, we focus on the 5-class problem in which a decision among 5 emotional categories has to be made for each test utterance. As published by the organizer of the Challenge, the unweighted average (UA) recall rates of the baseline systems, which use openSMILE toolset for LLD extraction and HTK/Weka toolset for classifiers, is 35.5% for dynamic modeling HMM and 28.9% for static modeling SVM. Furthermore, as part of the evaluation protocol, when the Synthetic Minority Oversampling TEchnique (SMOTE) [3] is applied to deal with the issue of skewed data, the performance of SVM can be improved to 38.2%. These results will be referred to as the baseline performances.

Further progress on FAU-Aibo 5-class problem has been reported over the years after the Challenge. For dynamic modeling, a GMM (equivalent to a one-state HMM) using 13 mel-frequency cepstral coefficients (MFCC) with the first and second derivatives achieves 41.4% UA [4]. A hybrid DBN-HMM system combining deep belief network and hidden Markov model achieves 45.6% UA, which stands as the performance to beat on FAU-Aibo [5]. For static modeling, the anchor model method commonly used in speaker recognition [6] has been transferred to emotion recognition, achieving 43.98% UA with SVM [7].

In this paper, we study the application of Gaussian mixture models (GMM) in the FAU-Aibo 5-class problem. In the dynamic modeling, the LLDs are scored by GMMs, which are equivalent to 1-state HMMs. In the static modeling, GMM is used in the procedure of forming super-vectors for SVM classifier. Super-vectors based on GMM have been widely used for speaker verification tasks [8, 9]. GMM-based super-vectors in combination with SVM have been applied in SER, which outperformed standard GMM system [10].

2 Proposed Methods

2.1 Gaussian Mixture Models

The central idea connecting the static and dynamic classifier frameworks is the Gaussian mixture models (GMM). A GMM is defined by the probability density function (PDF) of

$$p(x) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k) \tag{1}$$

where the weights satisfy

$$\pi_k \geq 0, \quad \sum_{k=1}^{K} \pi_k = 1. \tag{2}$$

Eq. (1) is said to have K components, where the kth component $N(x|\mu_k, \Sigma_k)$ is a Gaussian PDF with μ_k and Σ_k as the component mean vector and covariance matrix.

GMM is very commonly used to model continuous random variables. In theory, GMM is a model general enough to approximate any PDF by increasing the number of components. In practice, parameters in a GMM can be efficiently learned from data by EM algorithm [11, 12].

2.2 GMM and Universal Background Model

A universal background model (UBM) is a model for a data set regardless of the class labels. UBM is often used as the initial point of model adaptation [13]. For example, one way to obtain a set of speaker-dependent models is to first train a UBM using all data, and then adapt the UBM with speaker-dependent data for each speaker. It is common to use GMM for UBM, as GMM is a sound model in theory and in practice. Such a model is called GMM-UBM.

2.3 GMM and Super-Vectors

In this research, we adapt a GMM-UBM for each utterance and obtain utterance-dependent models. The adaptation is base on maximum a posteriori (MAP) criterion [14]. After adaptation, an utterance-dependent super-vector for each utterance is formed by the mean vectors of the corresponding utterance-dependent GMM. The process of creating super-vectors is illustrated in Figure 1. Finally, these utterance-dependent super-vectors are the proposed representation for emotion classification. They are used in the static modeling based on SVM.

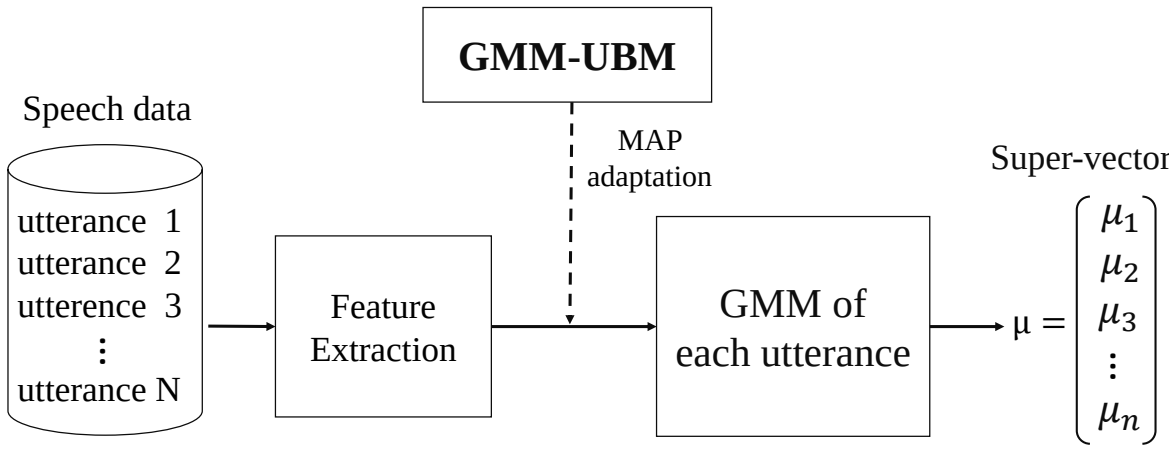

Figure 1: The creation of super-vectors.

2.4 GMM and Dynamic Model

Another way to investigate GMM-UBM is to train a UBM first, and then adapt the UBM with emotion-dependent data. Instead of a set of utterance-dependent models, this approach yields a set of emotion-dependent models. Furthermore, they are different from the models trained directly with emotion-dependent data, as in the case of baseline HMM dynamic modeling.

3 Systems

3.1 Data: FAU-Aibo

FAU-Aibo emotion corpus contains 9.2 hours of spontaneous speech recorded as children are interacting with a Sony pet robot Aibo. The data was collected from 51 German children (31 female and 20 male) at the age of 10 to 13 years from two different schools. There are 11 emotional categories, namely Angry, Touchy, Reprimanding, Helpless, Emphatic, Bored, Other, Neutral, Motherese, Surprised, Joyful. For each utterance, the emotional category of the majority by five persons is the label.

The 5-class problem defined by the Challenge is summarized in Table 1. The 5 emotional categories are A (angry), E (emphatic), N (neutral), P (positive), R (rest). Data from one school (Ohm) was used for training, with 9,959 utterances. Data from the other school (Mont) was used for testing, with 8,257 utterances. This is summarized in Table 2.

Table 1: Five emotional categories defined in INTERSPEECH 2009 Emotion Challenge.

A	Angry, Touchy, Reprimanding
E	Emphatic
N	Neutral
P	Motherese, Joyful
R	Surprised, Bored, Helpless

Table 2: Summarization of Data Points

Emotion	train data	test data
A	881	611
E	2093	1508
N	5590	5377
P	674	215
R	721	546
sum	9959	8257

3.2 Acoustic Features

We use openSMILE 2.0 to extract the standard features. There are 16 LLDs, including root mean square (RMS) frame energy, zero-crossing-rate (ZCR), 12 mel-frequency cepstral coefficients (MFCCs), harmonics-to-noise ratio (HNR), and pitch frequency (F0). They are enhanced by the delta coefficients. There are 12 functionals which are applied on sequences of LLDs, including mean, standard deviation, kurtosis, skewness, maximum value, minimum value, relative position, range, and two linear regression coefficients with their mean square error (MSE). This is summarized in Table 3. In total, there are $16 \times 2 \times 12 = 384$ standard features per utterance. In this paper, we use 384 standard features for our baseline, and we use 16 LLDs and their deltas for training GMM.

3.3 Classifier

For the static model, support vector machines (SVM) are used with the proposed GMM-based supervectors. SVM [15] is a supervised learning method learning hyperplanes in feature space. Specifically, we use SVM kernel function, sequential minimal optimization learning [16], polynomial kernel, and pairwise multi-class discrimination in the experiments. For the dynamic model, HMMs are used as the backend classifier.

Table 3: Baseline acoustic features [2].

LLDs	Functionals
RMS Energy	mean
ZCR	standard devation
MFCC 1-12	kurtosis, skewness
HNR	extrmes:value, rel.position, range
F0	linear regression:offset, slope, MSE

4 Results

4.1 SVM Static Model with Super-Vectors

The SVMs are implemented as specified by the Challenge. The dimension of the super-vector is related to the number of components in GMM. Since there are 16 LLD, the dimension is

$$16 \times K$$

with LLD alone, and

$$16 \times 2 \times K$$

if the delta LLD are also included in the feature vector. We use notation O (Original) to describe 16 LLDs, and use notation Δ to describe their delta.

The results with varying K are summarized in Table 4.

Table 4: Recall rates in percentage with support super-vector machines, using original data.

feature	no. comp	vector size	UA	WA
O	8	128	26.9	64.9
	32	512	28.4	62.5
	64	1024	28.3	60.9
O + Δ	8	256	**31.0**	64.8
	32	1024	29.8	60.1
	64	2048	30.1	55.5

Methods to balance data in different classes are applied to deal with skewed data issue, as can be seen in Table 2. We use SMOTE [3] to increase the number of data points in the classes of A, E, P, R

to the number of data points of N, resulting in 27,950 data points for the training data. The results with varying K is summarized in Table 5.

Table 5: Recall rates in percentage with support super-vector machines, combining SMOTE for data balance.

feature	no. comp	vector size	UA	WA
O	8	128	**38.6**	37.4
	32	512	35.1	42.2
	64	1024	33.1	42.6
O + Δ	8	256	**38.9**	40.2
	32	1024	34.4	44.7
	64	2048	34.8	43.5

From the results in Table 4 and Table 5, the following observations can be made.

- The proposed super-vectors outperform the baseline feature vectors, with SMOTE for data balance (38.9% over 38.4%) or without SMOTE (31.0% over 28.9%).
- When $K = 8$, the performance of 38.9% UA is better than the performance of 38.2% UA achieved by the baseline feature vectors. Note that this is achieved by a lower dimension of feature space (256 vs. 384).
- We can exclude the delta features to reduce feature dimension to 128, and still get better results than baseline (38.6% vs. 38.2%).

4.2 HMM Dynamic Model for LLD

Following the Challenge [2], we use HMMs for the standard LLDs. The results with baseline settings as follows are shown in Table 6.

- left-to-right HMM
- one model per emotion
- diverse number (1, 3, 5) of states
- 2 Gaussian mixtures
- 6+4 Baum-Welch re-estimation iterations

Table 6: UA recall rates in percentage of baseline HMM-GMM on standard LLDs.

feature	no. states	UA	WA
O	1	**36.1**	37.1
	3	33.8	32.7
	5	33.9	36.1
O + Δ	1	**36.3**	49.3
	3	36.2	35.7
	5	36.2	41.6

Table 7: UA recall rates in percentage of 1-state HMM-GMM on standard LLDs with varying components.

no. comp.	feature	UA	WA
4	O	36.0	33.4
	O + Δ	36.7	38.7
8	O	34.9	25.3
	O + Δ	36.7	40.5
16	O	35.9	34.7
	O + Δ	**40.0**	41.7

Hidden Markov Models (HMM) with Gaussian mixtures Model (GMM) for states. We increase the number of Gaussian components in HMM-GMMs. The results are shown in Table 7. The best performance we achieved by increasing the number of components and including the delta features is 40.0% UA recall rate, which is better than the baseline performance of 35.5% UA recall rate by 4.5% absolute.

4.3 GMM-UBM

Each emoiton is modeled as a single-state HMM and each state distribution is a GMM. In this paper, we call it HMM-GMMs. There are two different approaches to build emotion-dependent GMM models. The first approach is to use emotion-dependent data to train independent models, as is the case with 1-state HMM. The second approach is to use all data to train a UBM, then to adapt the UBM by emotion-dependent data to emotion-dependent models. In GMM-UBM, the second approach is taken. The results are shown in Table 8. The UA recall rate of 39.2% is achieved when the GMMs contain 256

Table 8: UA recall rates in percentage of GMM-UBM on standard LLDs with varying components.

no. comp.	feature	UA	WA
8	O	33.7	21.8
	O + Δ	34.1	20.2
32	O	37.6	29.1
	O + Δ	**39.1**	32.4
64	O	36.2	25.4
	O + Δ	37.9	31.5
256	O	34.2	20.5
	O + Δ	**39.2**	27.6

components.

5 Conclusion

In this paper, we apply super-vectors methods to speech emotion recognition. The construction of super-vectors is based on adaptation of Gaussian mixture models. Evaluated on INTERSPEECH 2009 Emotion Challenge, the proposed system achieves performance gain while reducing the dimension of feature space to 1/3 (128 vectors versus 384 vectors) or 2/3 (256 vectors versus 384 vectors). Furthermore, by increasing the number of components in HMM-GMM and including the delta features, the performance is found to improve significantly. In the future, we will use emo-large (6000x) features in our baseline and compare to super-vectors methods.

References

[1] X. Zhang, Y. Sun, and S. Duan, “Progress in speech emotion recognition,” TENCON 2015 - 2015 IEEE Region 10 Conference,pp. 1 – 6, 2015.

[2] B. Schuller, S. Steidl, and A. Batliner, “The INTERSPEECH 2009 emotion challenge,” in Proceedings of INTERSPEECH, 2009, pp.312–315.

[3] G. Weiss and F. Provost, “The effect of class distribution on classifier learning: An empirical study,” Department of Computer Science,Rutgers University, 2001.

[4] B. Vlasenko, “Processing affected speech within human machine interaction,” in Proc. Interspeech. Brighton, 2009, pp. 2039–2042.

[5] D. Le and E. M. Provost, “Emotion recognition from spontaneous speech using hidden markov models with deep belief networks,”Proceedings of Automatic Speech Recognition and Understanding(ASRU), 2013.

[6] Y. Yang, M. Yang, and Z. Wu, “A rank based metric of anchor models for speaker verification,” Proc. IEEE Intl Conf. Multimedia and Expo (ICME 06), pp. 1097–1100, 2006.

[7] S. Ntalampiras and N. Fakotakis, “Anchor models for emotion recognition from speech,” IEEE Transactions on Affective Computing,vol. 4, pp. 280–290, 2013.

[8] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff, “Svm based speaker verification using a gmm supervector kernel and nap variability compensation,” Proc. of ICASSP 2006, pp. 97–100, 2006.

[9] M. Liu and Z. Huang, “Multi-feature fusion using multi-gmm supervector for svm speaker verification,” International Congress on Image and Signal Processing. Tianjin: IEEE, pp. 1–4, 2009.

[10] H. Hu, Ming-XingXu, and W. Wu, “Gmm supervector based svm with spectral features for speech emotion recognition,” in Proc. Int. Conf.Acoustics, Speech, and Signal Processing, vol. 4, pp.413–416, 2007.

[11] Bishop, Pattern Recognition and Machine Learning. LLC, New York: Springer Science Business Media, 2006.

[12] 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.

[13] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted gaussian mixture models,” Digital Signal Processing, 10(1 - 3), pp. 19–41, 2000.

[14] J. L. Gauvain and C. H. Lee, “Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains,”IEEE Trans. Speech Audio Process, pp. 291–298, 1994.

[15] C. Cortes and V. Vapnik, “Support-vector networks,” Machine learning, vol. 20, no. 3, pp. 273–297, 1995.

[16] J. Platt, “Fast training of support vector machines using sequential minimal optimization,” in Advances in Kernel Methods—Support Vector Learning, B. Scholkopf, C. J. C. Burges, and A. J. Smola,Eds. Cambridge, MA: MIT Press, 1999, pp. 185–208.