Speech Emotion Recognition using Gaussian Mixture Vector Autoregressive Models

Conference Paper *in* Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on · May 2007

DOI: 10.1109/ICASSP.2007.367230 · Source: IEEE Xplore

3 authors:

Moataz M.H. El Ayadi
Cairo University
17 PUBLICATIONS **929** CITATIONS
SEE PROFILE

Mohamed S. Kamel
University of Waterloo
491 PUBLICATIONS **10,533** CITATIONS
SEE PROFILE

Fakhri Karray
University of Waterloo
456 PUBLICATIONS **6,674** CITATIONS
SEE PROFILE

SPEECH EMOTION RECOGNITION USING GAUSSIAN MIXTURE VECTOR AUTOREGRESSIVE MODELS

Moataz M. H. El Ayadi, Mohamed S. Kamel, and Fakhri Karray

Pattern Analysis and Machine Intelligence Lab, Electrical and Computer Engineering, University of Waterloo
{moataz,mkamel,karray}@pami.uwaterloo.ca

ABSTRACT

It is believed that modeling temporal structure of the speech data may be useful for the problem of speech emotion recognition [1]. In this paper, Gaussian mixture vector autoregressive model is proposed as a statistical classifier for this task. The main motivation behind using such a model is its ability to model the dependency among extracted speech feature vectors as well as the multi-modality in their distribution. When applied to the Berlin emotional speech database, the proposed technique provides a classification accuracy of 76% versus 71% for the hidden Markov model, 67% for the k-nearest neighbors, 55% for feed-forward neural networks. The model gives also better discrimination between high-arousal, low arousal, and neutral emotions than the HMM.

Index Terms— expectation maximization algorithm, Gaussian mixture models, maximum likelihood estimation, speech emotion recognition, vector autoregressive models.

1. INTRODUCTION

Speech emotion recognition refers to the process of determining the emotional state of a speaker. Recently, there has been an increasing research interest in speech emotion recognition for it has found a variety of applications such as web interactive movies, information retrieval, medical analysis, in-car board systems and text-to-speech synthesis [2].

Many classification techniques have been applied for speech emotion recognition such as neural network [3], hidden markov models (HMM) [1] and support vector machines [2]. However, an important remark in the majority of these techniques is that they do not model the temporal structure of the training data. The only exception may be the HMM in which the temporal structure of the data is modelled through its states. However, all the Baum-Welch re-estimation formulae are based on the assumption that all the feature vectors are statistically independent. Though this assumption is not valid in practice, the HMM has shown to be a powerful classifier in a variety of applications.

In many speech applications, vector autoregressive (VAR) models have been extensively employed to characterize the correlation between successive speech feature vectors. In fact, autoregressive Markov modelling of speech was originally proposed by Poritz [4]. His model consists of a sequence of states of each which is modelled by a VAR model rather than a Gaussian mixture model (GMM). Various modifications have been proposed to this model such as non-stationary autoregressive hidden Markov model (NAR-HMM) [5] and autoregressive hidden Markov model with duration [6]. In speech emotion classification, only short utterances are often available. Hence, it is preferable to use simpler models in order to avoid over-fitting.

In this paper, we employed a special form of vector autoregressive model as a classifier for the problem of speech emotion recognition. In particular, we assumed the distribution of the innovation sequence to be a mixture of Gaussian densities. We shall refer to this model as Gaussian mixture vector autoregressive (GMVAR) model. Our basic motivation behind using GMVAR for classification is its ability to model the dependency between successive feature vectors as well as the possible multi-modality in the distribution of the data. In order to use GMVAR for classification, we generalized the expectation maximization algorithm to adapt the case of having one or more multivariate time series realizations. The proposed model not only provides a higher classification accuracy than other techniques but also is more consistent with our intuition of dependency between successive feature vectors extracted from speech.

This paper is organized as follows. The GMVAR model is briefly reviewed in section 2. The proposed classification technique is explained in section 3. Experimental evaluation of the proposed method is given in section 4. Finally, important conclusions and a possible extension to this work are addressed in section 5.

2. GAUSSIAN MIXTURE VECTOR AUTOREGRESSIVE MODEL

A vector time series $\{\mathbf{x}[n]\}_{n=1}^{N}$, $\mathbf{x}[n] \in \mathbb{R}^d$ can be modelled by a VAR model of order P of the the following form.

$$\mathbf{x}[n] = \sum_{p=1}^{P} \mathbf{A}_p \mathbf{x}[n-p] + \mathbf{e}[n] = \tilde{\mathbf{A}}\mathbf{y}[n] + \mathbf{e}[n] \quad (1)$$

where

$$\tilde{\mathbf{A}} \equiv \begin{bmatrix} \mathbf{A}_1 & \mathbf{A}_2 & \dots & \mathbf{A}_P \end{bmatrix},$$

$$\mathbf{y}[n] \equiv \left[\mathbf{x}^{\mathsf{T}}[n-1] \quad \mathbf{x}^{\mathsf{T}}[n-2] \quad ... \quad \mathbf{x}^{\mathsf{T}}[n-P]\right]^{\mathsf{T}},$$

and $\mathbf{e}[n]$ is a sequence of iid random vectors and is called the *innovation sequence*[1]. Intuitively, the matrix $\tilde{\mathbf{A}}$ represents the degree of dependency of each feature vector on its past.

In our GMVAR model it is assumed that the distribution of the innovation sequence is a mixture of M Gaussian densities, i.e.,

$$f_{\mathbf{e}[n]}\left(\mathbf{e}[n]\right) = \sum_{m=1}^{M} w_m \mathbb{N}(\mathbf{e}[n]; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m), \tag{2}$$

where $\mathbb{N}(\mathbf{e}[n]; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$ is the multivariate Gaussian distribution with mean vector $\boldsymbol{\mu}_m$ and covariance matrix $\boldsymbol{\Sigma}_m$. The mixture weights (priors) $w_m, m = 1, 2, ..., M$ are positive and sum to one. For convenience, we shall denote all the model parameters by λ. Conditional on the first P pre-sample vectors being constant, the likelihood function of a certain time series realization, $X = \{\mathbf{x}[1], ..., \mathbf{x}[N]\}$, is given by:

$$\mathrm{p}\left(X|\lambda\right) = \prod_{n=1}^{N} \sum_{m=1}^{M} w_m \mathbb{N}\left(\mathbf{x}[n] - \tilde{\mathbf{A}}\mathbf{y}[n]; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\right) \tag{3}$$

2.1. The proposed technique

The above model can be adapted to our speech emotion recognition problem as follows. Each emotion is modelled by a GMVAR. Generally, there are more than one training utterance for each emotion. The feature vectors extracted from each utterance are considered as one time realization for its emotional class. It is also reasonable to assume that these realizations are statistically independent of each others. Hence, the likelihood of all the time series realizations of a certain emotion is given by

$$\mathrm{p}\left(X_1, ..., X_K|\lambda_c\right) = \prod_{k=1}^{K} \mathrm{p}\left(X_k|\lambda_c\right), \tag{4}$$

where $X_k = \{\mathbf{x}_k[1], ..., \mathbf{x}_k[N]\}$ is k^{th} time series realization. Thus, our proposed model can handle more than one multivariate time series realizations. For model training, the likelihood function in the above equation should be maximized with respect to the model parameters. Like many other statistical classifiers, this is done using the expectation maximization (EM) algorithm. Because of the limited length of the paper, we just give our final derived expressions for the update equations. Let the superscript (s) denotes the parameter values at iteration s; the parameter update equations are

$$w_m^{(s+1)} = \frac{\sum_{k,n} P_{knm}(\lambda^{(s)})}{\sum_{k=1}^{K} N_k}, \tag{5}$$

$$\boldsymbol{\mu}_m^{(s+1)} = \frac{\sum_{k,n} P_{knm}(\lambda^{(s)}) \mathbf{e}_k[n]}{\sum_{k,n} P_{knm}(\lambda^{(s)})} \tag{6}$$

[1] When the distribution of the innovation sequence is a mixture of Gaussian, there is no need to add an intercept term.

$$\boldsymbol{\Sigma}_m^{(s+1)} = \frac{\sum_{k,n} P_{knm}(\lambda^{(s)})(\mathbf{e}_k[n] - \boldsymbol{\mu}_m^{(s+1)})(\mathbf{e}_k[n] - \boldsymbol{\mu}_m^{(s+1)})^{\mathsf{T}}}{\sum_{k,n} P_{knm}(\lambda^{(s)})} \tag{7}$$

$$\mathsf{vec}\left(\tilde{\mathbf{A}}^{(s+1)}\right) = \left[\sum_{m=1}^{M} \sum_{k,n} P_{knm}(\lambda^{(s)}) \left((\mathbf{y}_k[n]\mathbf{y}_k^{\mathsf{T}}[n]) \otimes (\boldsymbol{\Sigma}_m^{-1})^{(s+1)}\right)\right]^{-1} \times \mathsf{vec}\left(\sum_{m=1}^{M} \sum_{k,n} P_{knm}(\lambda^{(s)})(\boldsymbol{\Sigma}_m^{-1})^{(s+1)}(\mathbf{x}_k[n] - \boldsymbol{\mu}_m^{(s+1)})\mathbf{y}_k^{\mathsf{T}}[n]\right), \tag{8}$$

where $\mathbf{e}_k[n] = \mathbf{x}_k[n] - \tilde{\mathbf{A}}^{(s)}\mathbf{y}_k[n]$,

$$P_{knm}(\lambda) = \frac{w_m \mathbb{N}(\mathbf{e}_k[k]; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)}{\sum_{m'=1}^{M} w_{m'} \mathbb{N}(\mathbf{x}_k[k] - \tilde{\mathbf{A}}\mathbf{y}_k[n]; \boldsymbol{\mu}_{m'}, \boldsymbol{\Sigma}_{m'})},$$

and $\otimes$ denotes the Kronecker product of two matrices and $\sum_{k,n}$ is a shorthand for $\sum_{k=1}^{K} \sum_{n=1}^{N_k}$.

The above update equations are iterated until there is no significant increase in the likelihood function or a maximum number of iterations has been exceeded. Finally, what remains is how to initialize the model parameters. One simple method is apply the estimation procedures several times with different random initializations. This will result in several candidate models. The model with the highest likelihood is picked up. Training is repeated for each emotion and thus we have C trained models $\lambda_1, ...\lambda_C$, where C is the number of emotions.

In the testing phase, it is required to determine from a given set of emotions the one that is most likely produced by a certain unknown testing utterances. If all the emotions are assumed to have equal prior probability, the Bayesian decision rule reduces to the ML decision rule. Denote the speech feature vectors extracted from the test utterance by $\{\mathbf{x}[n]\}_{n=1}^{N}$. The index of the most likely emotion, $\hat{c}$ produced by the unknown testing utterance is given by

$$\begin{aligned} \hat{c} &= \arg \max_{c=1,2,...,C} \mathrm{p}\left(\{\mathbf{x}[n]\}_{n=1}^{N}|\lambda_c\right) \\ &= \arg \max_{c=1,2,...,C} \log \mathrm{p}\left(\{\mathbf{x}[n]\}_{n=1}^{N}|\lambda_c\right) \end{aligned} \tag{9}$$

where the logarithm of the likelihood is taken to avoid possible numerical underflows when multiplying a large number of small values in the calculation of the likelihood value. Substituting (3) in (9) yields

$$\hat{c} = \arg \max_{c=1,2,...,C} \sum_{n=1}^{N} \log \left(\sum_{m=1}^{M} w_{m,c} \mathbb{N}\left(\mathbf{e}_c[n]; \boldsymbol{\mu}_{m,c}, \boldsymbol{\Sigma}_{m,c}\right)\right), \tag{10}$$

where $\mathbf{e}_c[n] = \mathbf{x}[n] - \tilde{\mathbf{A}}_c\mathbf{y}[n]$.

3. EXPERIMENTAL EVALUATION

The above technique was applied to the Berlin emotional speech database [7], which contains 494 utterances with the following adult-directed emotions: *anger*, *boredom*, *fear*, *happiness*, *sadness*, and *neutral*. The sampling rate for all utterances is 16 kHz and the speech quality is high. In order not to favor one of the emotions over the others, the number of training and testing utterances should be the same for all emotions. Since the total number of utterances for each emotion is variable, only fifty utterances are randomly selected without replacement from each emotion. The number of utterances for the *disgust* emotion was fairly low and hence this emotion was discarded from the experiments.

3.1. Speech Processing and Feature Extraction

The speech raw time samples are high-pass filtered using a radiation filter with a pre-emphasis coefficient of 0.97. Hamming windows of duration 25 msec were used to extract features at a rate of one feature every 10 msec. The main speech features extracted from each frame of the speech signal were 12 mel-frequency cepstrum coefficient (MFCC), 12 delta coefficients, 0'th cepstral coefficient, and the speech energy. MFCC is usually used for many speech applications because it models the human perception to speech quite well [8]. In order to increase the distinguishability between emotions, the heteroscedastic linear discriminant analysis (HLDA) [9] was utilized to reduce the dimensionality of feature vectors to 12.

3.2. Results and discussion

The above technique is compared to the k-nearest neighbors (k-NN), the feed-forward artificial neural networks (ANN), and the continuous hidden markov model (HMM) classification techniques. In all simulations, the number of hidden layers in the ANN was fixed to two layers and the back-propagation algorithm is used to train the network. At each instant a single frame is input to the network. While the first two classifiers are considered as representatives for classification techniques that do not model timing dependency altogether, the last two classifiers timing dependency through state transitions. In addition, HMM is very popular in speech applications and has been applied to the problem of speech emotion recognition [1]. In this paper, all simulations of the HMM were done using the hidden Markov toolkit (HTK) [10] thanks to its reliable performance.

For all the classification techniques, it was necessary to apply a model selection technique to determine the following structural parameters: the number of neighbors in the kNN classifiers, the number of nodes in each hidden layer of the ANN classifiers, the number of states and the number of Gaussian components per states in the HMM, and the order and the number of Gaussian components in the GMVAR model. Since the number of available training utterances was limited, it was not reliable to use information-theoretic model selection criteria such as minimum description length (MDL) and Akaike information criterion (AIC). Instead, a model selection technique based on cross-validation was applied to all the above-mentioned classifiers [[11], ch.9]. In particular, for each possible setting of the structural design parameters of the classifier, five-fold cross validation technique was applied to the training data only. The model selection criterion is the minimum average cross validation error. Once the values of structural design parameters are selected, all the training data is used to retrain the selected model and the accuracy with respect to the test set is reported.

In order to demonstrate the importance of modelling the dependency between successive feature vectors, the cross validation accuracies obtained when applying the above model selection technique with GMVAR are averaged with respect to M and plotted versus the number of lags, P. The plot is shown in figure 1. The case of $P = 0$ corresponds to a pure Gaussian mixture model, i.e., there is no modelling of dependency between feature vectors. It is noted that the accuracy increases in general with the increase of number of lags which corresponds to modelling the correlation between a larger number of successive vectors. Thus, taking such a dependency into account results in an increase in the classification accuracy. However, and like many other classifiers, the accuracy decreases when P is too large since the model may be over-fitted to the distribution of the training data.

Table 1 shows the classification accuracy, the classification time and the selected structural parameters of all classification techniques. It is noted from the table that the value of the accuracy corresponding to GMVAR is higher than the peak accuracy in Figure 1. This is expected since the amount of the training data used to obtain the accuracies in Table 1 is larger than that used in Figure 1. It can also be deduced from the table that the techniques that model timing dependency (proposed and HMM) generally perform better than other techniques which ignore dependency completely. Comparing the identification times of different techniques, it is clear that the average time required by the k-NN is from one to two order of magnitudes higher than other methods. This may be undesirable for many practical applications. On the other hand, the average identification times of other techniques are almost comparable. In addition, the classification performance of the ANN is inferior to other techniques. According to literature, it seems that ANNs are not well suited for speech emotion recognition [3]. Based on the table data, it may be deduced that the proposed classification technique achieves the best compromise between the classification accuracy and the classification time.

Tables 2 and 3 show the normalized confusion matrices for both the proposed technique and the HMM technique (the second best classification method). Grouping the emotions into three sets: high-arousal emotions (anger, fear, and happiness) and low-arousal emotions (boredom and sadness) and the neutral emotion, it is noted that the confusion between two emotions in the same set is higher than the confusion between two emotions in different sets. This is consistent with

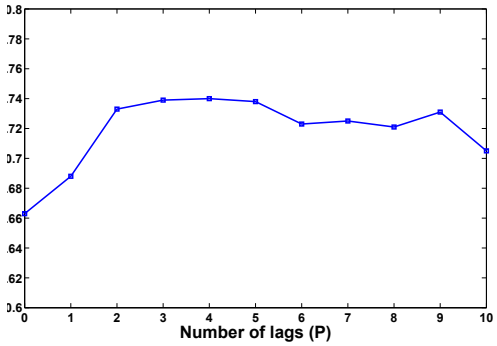

Fig. 1. Average classification accuracy of the proposed GMVAR classification technique when applied to the Berlin emotional speech database.

Table 1. Classification accuracies, average identification times, and selected structural parameters of different classification techniques when applied to the Berlin emotional speech database.

Classification method	Average Accuracy	Classification time (seconds)	Selected structural parameters
GMVAR	76.0%	0.3253	$M = 2$ & $P = 9$
HMM	71.0%	0.3505	$M = 6$ & # states = 5
k-NN	67.3%	16.2132	# neighbors = 6
ANN	55.0%	0.2573	# neurons = 5

what is reported in the literature [1]. From Table 2 and 3, it can be easily deduced the accuracy of classification between high-arousal emotions, low-arousal emotions, and the neutral emotion is 90.33% for the proposed method versus 86.00% for the HMM technique. This is intuitive since the speech rate for low-arousal emotions is significantly less than that of high-arousal ones. Hence, there should be a difference in the temporal profile of features extracted from the two emotion types.

4. CONCLUSIONS

In this paper, classification using GMVAR models has been proposed for speech emotion recognition. GMVAR models have the advantages of modelling the dependency between successive feature vectors and the multi-modality in their distribution. In addition, the proposed classification technique has been found to provide a better classification performance than other techniques such as the HMM and the kNN in terms of the overall accuracy and the discrimination of high-arousal and low-arousal emotions. To further improve the classification performance, we shall study the implementation of a two-stage classifier. In the first stage, emotions are classified into high arousal, low arousal, and neutral emotions using our proposed method. In the second stage, another classifier is used to distinguish between emotions in the same category.

Table 2. Normalized confusion matrix of the proposed classification technique when applied to the Berlin database.

	Recognized emotion					
True emotion	anger	fear	happiness	boredom	sadness	neutral
anger	**0.74**	0.08	0.16	0	0	0.02
fear	0.08	**0.66**	0.12	0	0.04	0.10
happiness	0.18	0.18	**0.62**	0	0	0.02
boredom	0	0.02	0.02	**0.76**	0.04	0.16
sadness	0	0	0	0.02	**0.96**	0.02
neutral	0	0.02	0.04	0.12	0	**0.82**

Table 3. Normalized confusion matrix of the HMM classification technique when applied to the Berlin database.

	Recognized emotion					
True emotion	anger	fear	happiness	boredom	sadness	neutral
anger	**0.78**	0.06	0.16	0	0	0
fear	0.04	**0.7**	0.16	0.04	0.02	0.04
happiness	0.24	0.04	**0.68**	0	0	0.04
boredom	0	0.04	0	**0.42**	0.16	0.38
sadness	0	0	0	0.04	**0.94**	0.02
neutral	0	0.08	0	0.16	0.02	**0.74**

5. REFERENCES

[1] T. Nwe, S. Foo, and L. De Silva, "Speech emotion recognition using hidden markov models," *Speech Communication*, vol. 41, pp. 603–623, 2003.

[2] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," *Proc. ICASSP 2004*, vol. 1, pp. 577–580, 2004.

[3] J. Nicholson, K. Takahashi, and R. Nakatsu, "Emotion recognition in speech using neural networks," *Neural Computing & Applications*, vol. 9, pp. 290–296, 2000.

[4] A. Poritz, "Linear predictive hidden markov models and the speech signals," *ICASSP 1982*, pp. 1291–1294, 1982.

[5] K. Lee and J. Lee, "Recognition of noisy speech by a nonstationary ar hmm with gain adaptation under unknown noise," *IEEE Trans. Speech & Audio Processing*, vol. 9, no. 7, pp. 741–746, 2001.

[6] Y. Ephraim and W. Roberts, "Revisiting autoregressive hidden markov modeling of speech signals," *IEEE Signal Processing letters*, vol. 12, no. 2, pp. 166–169, 2005.

[7] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of german emotional speech," *Proc. Interspeech 2005, Lissabon, Portugal*, 2005.

[8] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, New Jersey, 1993.

[9] K. Martin, G. Frantisek, S. Petr, B. Luks, and C. Jan, "Robust heteroscedastic linear discriminant analysis and lcrc posterior features in large vocabulary continuous speech recognition," *Proc. Fifth Slovenian and First International Language Technologies Conference, Ljubljana, SI*, pp. 1–4, 2006.

[10] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK book (for version 3.1)*, 2002.

[11] C. Bishop, *Neural Networks for Pattern Recognition*, Clarendon Press, Oxford, 1995.