Deep Learning for Emotional Speech Recognition

Máximo E. Sánchez-Gutiérrez[1], E. Marcelo Albornoz[2], Fabiola Martinez-Licona[1], H. Leonardo Rufiner[2], and John Goddard[1]

[1] Departamento de Ingeniería Eléctrica, Universidad Autónoma Metropolitana (México)
[2] Centro de Investigación SINC(i), Universidad Nacional del Litoral - CONICET (Argentina)
edmax86@gmail.com

Abstract. Emotional speech recognition is a multidisciplinary research area that has received increasing attention over the last few years. The present paper considers the application of restricted Boltzmann machines (RBM) and deep belief networks (DBN) to the difficult task of automatic Spanish emotional speech recognition. The principal motivation lies in the success reported in a growing body of work employing these techniques as alternatives to traditional methods in speech processing and speech recognition. Here a well-known Spanish emotional speech database is used in order to extensively experiment with, and compare, different combinations of parameters and classifiers. It is found that with a suitable choice of parameters, RBM and DBN can achieve comparable results to other classifiers.

Keywords: Emotional speech recognition, restricted Boltzmann machines, deep belief networks.

1 Introduction

The automatic recognition of emotions in human speech is part of the multidisciplinary research area of Human-Machine Communication, and has received increasing attention over the last few years. One reason for this interest is the growing number of applications which have benefitted from the research conducted in the field, like call centers, video games, and lie detection.

However, automatic emotional speech recognition involves many issues which need to be carefully studied, such as: which emotions can we really identify, what are the best features to use for the identification, and which classifiers give the best performance. To illustrate these issues we can mention that although it is common to consider the 'big six' emotions of joy, sadness, fear, disgust, anger, and surprise, along with neutral, Douglas-Cowie, Cox et al. [1], proposed a far greater list of 48 emotion categories; in the INTERSPEECH Challenges from 2009 to 2012, the number of proposed features has increased from 384 to

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
M. Sánchez-Gutiérrez, E. M. Albornoz, F. M. Martínez, H. L. Rufiner & J. Goddard; "Deep Learning for Emotional Speech Recognition"
6th Mexican Conference on Pattern Recognition (MCPR2014), LNCS - Springer, pp. 311-320, jun, 2014.

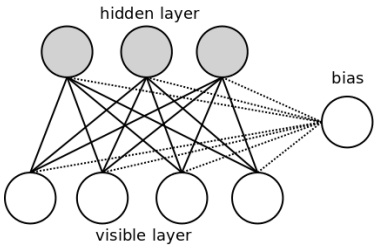

Fig. 1. Restricted Boltzmann Machine.

6125 [2,3] without a final set being decided upon; finally, as Scherer [4] reports, a wide range of classifiers, such as linear discriminant classifiers, k-nearest neighbor (KNN), Gaussian mixture model, support vector machines, decision tree algorithms (DT) and hidden Markov models have all been examined, and no definitive classifier has been chosen.

In this paper, we shall consider the application of RBM and DBN to the problem of classification of emotional speech recognition. The principal motivation lies in the success reported in a growing body of work employing these techniques as alternatives to traditional methods in speech processing and speech recognition [5,6].

Not much work has been conducted using RBM and DBN for the task of automatic emotional speech recognition. In [7], a Generalized Discriminant Analysis based on DBN showed significant improvement over support vector machines on nine databases. However, in [8], on the Likability Sub-Challenge classification task at INTERSPEECH 2012, it was found that the use of RBM helped in the task but that DBN did not. It seems that the parameters involved in training these algorithms are highly sensitive to small modifications, and that there is still work to be done in deciding how to use them for a particular task.

With this in mind, in the present paper we shall conduct an extensive experimentation with RBM and DBN in the context of a Spanish emotional speech database. The organization of the paper is as follows: in section 2 we briefly review the principal ideas of RBM and DBN, continuing in section 3 with information about the emotional speech database we use. In section 4 we describe the experiments we conducted and finally, in sections 5 and 6 discuss the results and end with some conclusions.

2 Deep Learning

2.1 Restricted Boltzmann Machines

An RBM is an artificial neural network with two layers, one layer formed with visible units, to receive the data, and the other with hidden units. There is also a bias unit. This architecture is shown in Figure 1. The hidden units are usually binary stochastic and the visible units are typically binary or stochastic gaussian.

An RBM represents the joint distribution between a visible vector and a hidden random variable.

An RBM only has connections between the units in the two layers, and with the bias unit, but not between the units in the same layer. One reason for this is that efficient training algorithms have been developed (c.f. Hinton's Contrastive Divergence algorithm [9]) which allow the connection weights to be learned.

A given RBM defines an energy function for every configuration of visible and hidden state vectors, denoted v and h respectively. For binary state units, we define the energy function, $E(v, h)$ by:

$$E(v, h) = -a'v - b'h - h'Wv \tag{1}$$

where W is the symmetric matrix of the weights connecting the visible and hidden units, and a, b are bias vectors on the connections of bias unit to the visible and hidden layer, respectively.

The joint probability, p(v, h), for the RBM mentioned above, assigns a probability to every configuration (v, h) of visible and hidden vectors using the energy function:

$$p(v, h) = \frac{\exp^{-E(v,h)}}{Z} \tag{2}$$

where Z, known as the partition function, is defined by:

$$Z = \sum_{v,h} \exp^{-E(v,h)} \tag{3}$$

The probability assigned by the network to a visible vector v is:

$$p(v) = \frac{1}{Z} \sum_{h} \exp^{-E(v,h)} \tag{4}$$

It turns out that the lack of connections in the same layer of an RBM contributes to the property that it's visible variables are conditionally independent, given the hidden variables, and vice versa. This means that we can write these conditional probabilities as:

$$\begin{aligned} p(v_j = 1|h) &= \sigma(a_i + \sum_j h_j w_{i,j}) \\ p(h_j = 1|v) &= \sigma(b_j + \sum_i v_i w_{i,j}) \end{aligned} \tag{5}$$

where:

$$\sigma(x) = \frac{1}{1 + \exp^{-x}} \tag{6}$$

is the sigmoid function.

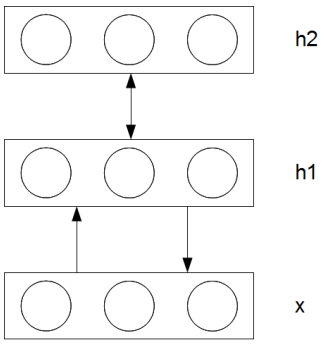

Fig. 2. Deep Belief Architecture. With x as the presentation layer and $h1$, $h2$ as hidden RBM layers

The Contrastive Divergence (CD) algorithm is applied to find the parameters W, a, and b. The algorithm performs Gibbs sampling and is used inside a gradient descent procedure to compute weight update. A guide to training an RBM is given in [10].

When real-valued input data is used, the RBM is modified to have Gaussian visible units, and the energy function is altered to reflect this modification (c.f. [8]) as:

$$E(v,h) = \sum_i \frac{(v_i - a_i)^2}{2\sigma_i^2} - \sum_i \sum_j \frac{v_i}{\sigma_i^2} h_j w_{ij} - b'h \tag{7}$$

With this modified energy function, the conditional probabilities are now given by:

$$\begin{aligned} p(h_j = 1|v) &= \sigma(\sum_i \frac{v_i}{\sigma_i^2} w_{ij} + b_j) \\ p(v_i = v|h) &= \mathcal{N}(v|\sum_j h_j w_{ij} + a_i, \sigma_i^2) \end{aligned} \tag{8}$$

where: $\mathcal{N}(\cdot|\mu, \sigma^2)$ denotes the Gaussian probability density function with mean μ and variance σ^2.

2.2 Deep Belief Networks

As Bengio [11] states: *"there is theoretical evidence which suggests that in order to learn complicated functions that can represent high-level abstractions (e.g. in vision, language, and other AI-level tasks), one needs deep architectures."*

One type of deep architecture is the DBN. Their use has already given excellent results in certain speech representation and recognition problems (c.f. [5,6]).

A DBN consists in a number of stacked RBM, as shown in Figure 2. Hinton, Osindero and Teh [12] proposed an unsupervised greedy layer-wise training, in which each layer is trained, from the bottom upwards, as an RBM using the activations from the lower layer. This stacking method makes it possible to train many layers of hidden units efficiently, although with a large data set training may take a long time, and coding with GPU's has been a recent development.

When a DBN is used for classification purposes, there are essentially two modes we can use once it has been trained: either place a classifier above the top level and train the classifier in a supervised manner with the output from the RBM/DBN (we refer to this as 'mixed'), or, add another layer of outputs and apply back-propagation to the whole neural net.

3 The Emotional Speech Database

Most of the developed emotional speech databases are not available for public use. Thus, there are very few benchmark databases. Table 1 summarizes characteristics of some databases commonly used in speech emotion recognition.

Table 1. Characteristics of common emotional speech databases. Adapted from [13]

Corpus	Access	Language	Size	Emotions
LDC Emotional Prosody Speech and Transcripts	Commercially available	English	1050 utterances	Neutral, panic, anxiety, hot anger, cold anger, despair, sadness, elation, joy, interest, boredom, shame, pride, contempt
Berlin emotional database	Public and free	German	535 utterances	Anger, joy, sadness, fear, disgust, boredom, neutral
Danish emotional database	Public with license fee	Danish	260 utterances	Anger, joy, sadness, surprise, neutral
Natural	Private	Mandarin	388 utterances	Anger, neutral
ESMBS	Private	Mandarin	720 utterances	Anger, joy, sadness, disgust, fear, surprise
INTERFACE	Commercially available	English, Slovenian, Spanish, French	186, 190, 184, 175 utterances respectively	Anger, disgust, fear, joy, surprise, sadness, slow neutral, fast neutral

The variability between sentences, speech styles and even speakers, as well as the number of emotions considered, presents a challenge in emotional speech recognition. Here we shall minimize the number of variables involved by choosing

a well-known Spanish emotional speech database [14], one female speaker, and take the emotions of joy, sadness, anger, fear, disgust and surprise along with neutral.

The database was created by the Center for Language and Speech Technologies and Applications (TALP) of the Polytechnic University of Catalonia (UPC) for the purpose of emotional speech research. The database was part of a larger project, INTERFACE, involving four languages, English, French, Slovene, and Spanish. In the case of Spanish, two professional actors, a man and a woman, were used to create the corpus. The speech corpus consisted of repeating 184 sentences with the big six emotions together with several neutral styles.

The 184 sentences include isolated words, sentences, which can also be in the affirmative and interrogative forms. The distribution is shown in Table 2.

Table 2. Spanish Corpus Contents.

Identifier	Corpus contents
001 - 100	Affirmative
101 - 134	Interrogative
135 - 150	Paragraphs
151 - 160	Digits
161 - 184	Isolated words

In a subjective test of the database with 16 non-professional listeners (UPC engineering students), it was found that over 80% of the sentences were correctly classified initially, and given a second choice, more than 90%. Each expression was correctly classified by at least half of the listeners.

It is interesting to note that errors were generally committed on the isolated words or short phrases, while all sentences and longer texts were classified correctly initially by all listeners. This subjective test is useful because we can compare it to the results obtained automatically by classifiers.

4 RBM & DBN experiments

In this section we describe the details of the experiments performed in the paper. We first discuss the feature extraction stage, then we describe the configurations used in order to train and test the different classifiers.

We use audio files from the female Spanish speaker of the database. For every utterance within these groups (1,100 total), two kinds of characteristics were extracted: mel-frequency cepstral coefficients (MFCCs) and prosodic features. All the experiments were performed using a selected 70% of the patterns for training (770 patterns), 25% for testing (275 patterns) and 5% for validation purposes (55 patterns), all selected in a balanced manner. As usual, the training process continued until the generalization peak with respect to validation set was

reached. In order to avoid a class biased data problem, each subset was sampled in a supervised manner to ensure that it was properly balanced.

The most popular feature representation currently used for speech recognition is MFCC [15]. It is based on a linear model of voice production together with a codification in a psychoacoustic scale.

Prosodic features have been used extensively for emotional speech recognition, such as energy, zero crossing rate and fundamental frequency, F_0, calculated for the speech signals considered [16]. In fact, many parameters can be extracted from the temporal evolution of prosodic features. Usually the minimum, mean, maximum and standard deviations over the whole utterance are used. This set of parameters has already been studied and some works have reported an important improvement in speech emotion discrimination [17].

In this work we have computed the average of the first 12 MFCC over the entire utterance, the average of F_0, the average of the zero crossing rate and the average of the energy, each one with their respective first derivative, all extracted using the OpenSMILE tool [18]. Hence, we represented each utterance with a 30-dimensional feature vector: $(1+\Delta)(12\,\mathsf{MFCCs}+\mathsf{mean\ F_0}+\mathsf{mean\ ZCR}+\mathsf{mean\ Energy})$.

For RBM and DBN experiments, we use the toolbox developed by Drausin Wulsin [19]. A large number of experiments were conducted in order to determine the best configurations and parameters for the RBM. These experiments consisted in different combinations of: varying the size of the batch (number of training vectors used in each pass of each epoch for the Contrastive Divergence algorithm), the learning rate, the number of hidden units, and the number of stacked RBM. All RBM had Gaussian units. DBN experiments were performed by adding one additional RBM layer to a previously trained DBN, and using the parameters shown in Table 3. The classification layer had seven output units, one for each class. The most probable class was considered as the unit with the highest activation level.

Table 3. Configuration parameters details for RBM/DBN training.

Parameter	Values
Batch size	[6, 12, 18, 24, 30, 36, 42, 48, 54, 60]
Learning rate	[0.01, 0.001, 0.0001, 0.00001]
Hidden units	[28, 56, 84, 112, 140, 168]
Number of layers	[1, 2, 3, 4, . . . , 13, 14, 15]

For the experiments performed with Support vector machines (SVM) we used the LIBSVM toolbox [20] and for the rest of the classifiers: K-nearest neighbors (KNN), Decision trees (DT) and Multilayer Perceptron (MLP), the Statistics and Neural Networks Toolboxes from Matlab were used [21]. The SVM was used

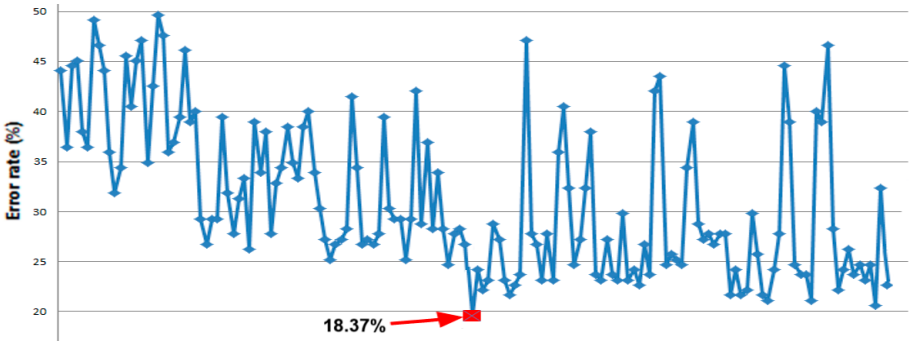

Fig. 3. Extract of some error rates of DBNs experiments for the different combinations of configurations seen in Table 3.

with a radial kernel. For KNN three neighbors and the cosine distance measure were used. The DT was constructed using Gini's diversity index, and then pruned in order to obtain better generalization capabilities. The MLP was trained in a traditional way with one hidden layer and ten hidden units, all the activation functions were sigmoid except in the last layer where "tansig" functions were used. We also performed some "mixed" experiments where SVM, KNN, DT and MLP classifiers were fed the outputs of an RBM.

5 Results and Discussion

In this section we present the results of the previously described experiments. With DBN classifiers, some of the results for the different configurations presented in Table 3 are shown in Figure 3. The combination of parameters that yields the best result was: 112 hidden units, a batch size of 42 and a learning rate of 0.00001. With this configuration the DBN achieved an error rate of 18.37%.

In order to perform a second set of experiments with several stacked RBMs, we used the DBN with the best configuration. The results can be seen in the Figure 4. For the case of mixed classifiers, the outputs of RBMs can generally help the other classifiers to achieve better performance, as seen in Table 4. These results were obtained by feeding the classifier with the output of three trained RBMs as described above, as can be seen, the results obtained are better than those obtained with the other classifiers.

It also can be seen that the best result was achieved with only one, two and three layers of RBMs and then it got worse. A possible explanation for this result is that with the aim of minimizing the number of variables, we are using a small subset of emotions and data; the more stacked RBMs, the more free parameters to train, requiring additional training data in order to properly estimate the parameters. Further work needs to be done so we can prove this affirmation.

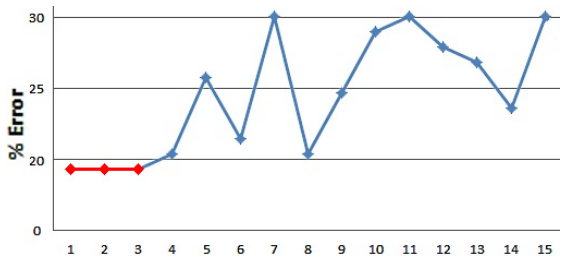

Fig. 4. Error rates against number of stacked RBMs. The best results were obtained with 1, 2 and 3 RBMs

Table 4. Best performance obtained for the different classifiers.

Classifier	Error rate (%)
DBN - RBM	**18.37**
K-nn	31.63
DBN - K-nn	24.49
DT	34.69
DBN - DT	23.47
MLP	40.82
DBN - MLP	20.41
SVM	25.43
DBN - SVM	18.97

6 Conclusions

In this work we considered the application of Restricted Boltzmann machines and Deep Belief Networks to the task of automatic Spanish emotional speech recognition. The results obtained are comparable, and in fact better than the results of other selected classifiers, when the parameters were correctly chosen. Future work includes extending these experiments to other languages in the database, and another using Mexican Spanish, that is being developed by us.

Acknowledgments. The authors wish to thank: *SEP* and *CONACyT* (Program SEP-CONACyT CB-2012-01, No.182432) and the *Universidad Autónoma Metropolitana* from México; *ANPCyT* and *Universidad Nacional de Litoral* (with PAE 37122, PACT 2011 #58, CAI+D 2011 #58-511) and *CONICET* from Argentina, for their support. We also want to thank ELRA for supplying the, Emotional speech synthesis database, catalogue reference: ELRA-S0329.

References

1. Douglas-Cowie, Cox: Humaine d5f deliverable, obtainable from http://emotion-research.net/download/pilot-db/

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
M. Sánchez-Gutiérrez, E. M. Albornoz, F. M. Martínez, H. L. Rufiner & J. Goddard; "Deep Learning for Emotional Speech Recognition"
6th Mexican Conference on Pattern Recognition (MCPR2014), LNCS - Springer, pp. 311-320, jun, 2014.

2. Schuller, B., Steidl, S., Batliner, A.: The interspeech 2009 emotion challenge. INTERSPEECH 2009, 10th Annual Conference of the International, Speech Communication Association (2009) 312–315
3. Schuller, B., Steidl, S., Batliner, A., Noth, E., Vinciarelli, A., Burkhardt, F., van Son, R., Weninger, F., Eyben, F., Bocklet, T., Mohammadi, G., Weiss, B.: The interspeech 2012 speaker trait challenge. Proc. INTERSPEECH (2012)
4. Scherer, K.R.: A blueprint for affective computing: a sourcebook. Oxford: Oxford University Press (2010)
5. Mohamed, A., Sainath, T., Dahl, G.E., Ramabhadran, B., Hinton, G., Picheny, M.: Deep belief networks using discriminative features for phone recognition. ICASSP-2011 2012, ISCA, Portland, OR, USA (2012)
6. Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A.: Deep neural networks for acoustic modeling in speech recognition. IEEE Signal Processing Magazine (2012)
7. Stuhlsatz, A., Meyer, C., Eyben, F., Zielke, T., Meier, G., Schuller, B.: Deep neural networks for acoustic emotion recognition: Raising the benchmarks. Proc. Int. Conf. on Acoustics, Speech, and Signal Processing, ICASSP 2011, Prague, Czech Republic (2011) 5688–5691
8. Bruckner, R., Schuller, B.: Likability classification - a not so deep neural network approach. Proceedings INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association (2012)
9. Hinton, G.E.: Training products of experts by minimizing contrastive divergence. Neural Computation **14** (2002) 1771–1800
10. Hinton, G.: A practical guide to training restricted boltzmann machines. UTML TR 2010-003, University of Toronto (2010)
11. Bengio, Y.: Learning deep architectures for ai. Foundations and Trends in Machine Learning **2** (2009) 1–127
12. Hinton, G.E., Osindero, S., Teh, Y.: A fast learning algorithm for deep belief nets. Neural Computation **18** (2006) 1527–1554
13. El Ayadi, M., Kamel, M.S., Karray, F.: Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition **44**(3) (2011) 572–587
14. catalogue, E.: http://catalog.elra.info, emotional speech synthesis database, catalogue reference: Elra-s0329
15. Rabiner, L., Juang, B.H.: Fundamentals of speech recognition. Prentice Hall PTR (1993)
16. Deller, J., Proakis, J., Hansen, J.: Discrete-time processing of speech signals. Prentice Hall PTR, Upper Saddle River, NJ, USA (1993)
17. Albornoz, E., Milone, D., Rufiner, H.: Spoken emotion recognition using hierarchical classifiers. Computer Speech and Language **25** (2011) 556–570
18. Eyben, F., Wollmer, M., Schuller, B.: opensmile - the munich versatile and fast open-source audio feature extractor. Proc. ACM Multimedia (MM), ACM, Florence, Italy (2010)
19. Wulsin, D.: Dbn toolbox v1.0, http://www.seas.upenn.edu/ wulsin/. Department of Bioengineering, University of Pennsylvania (2010)
20. Chang, C.C., Lin, C.J.: Libsvm: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (2011)
21. Guide, M.U.: Mathworks, http://www.mathworks.com. (2011)

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
M. Sánchez-Gutiérrez, E. M. Albornoz, F. M. Martínez, H. L. Rufiner & J. Goddard; "Deep Learning for Emotional Speech Recognition"
6th Mexican Conference on Pattern Recognition (MCPR2014), LNCS - Springer, pp. 311-320, jun, 2014.