sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
E. M. Albornoz, M. Sánchez-Gutiérrez, F. M. Martínez, H. L. Rufiner & J. Goddard; "Spoken emotion recognition using deep learning"
19th Iberoamerican Congress on Pattern Recognition (CIARP 2014), nov, 2014.

Spoken emotion recognition using deep learning

E. M. Albornoz[1], M. Sánchez-Gutiérrez[2], F. Martinez-Licona[2], H.L. Rufiner[1]
and J. Goddard[2]

[1] Centro de Investigación SINC(i), Universidad Nacional del Litoral - CONICET
(Argentina)

[2] Departamento de Ingeniería Eléctrica, Universidad Autónoma Metropolitana
(México)

Abstract. Spoken emotion recognition is a multidisciplinary research area that has received increasing attention over the last few years. In this paper, restricted Boltzmann machines and deep belief networks are used to classify emotions in speech. The motivation lies in the recent success reported using these alternative techniques in speech processing and speech recognition. This classifier is compared with a multilayer perceptron classifier, using spectral and prosodic characteristics. A well-known German emotional database is used in the experiments and two methodologies of cross-validation are proposed. Our experimental results show that the deep method achieves an improvement of 8.67% over the baseline in a speaker independent scheme.

1 Introduction

Emotion recognition has received much attention in recent years, mainly because its result could be useful in various applications [7,27]. Emotions represent a very important part in human communications and they can be perceived in speech signals, in facial expressions, in biosignal as electrocardiograph (ECG), among others. In spite of good results for different signals, the use of speech signals is the most feasible option because the methods to record and use other signals are invasive, complex and impossible in certain real applications. Most of the previous works on emotion recognition have been based on the analysis of speech prosodic features and spectral information [8,4,1,2]. With regard to classification, several standard techniques have been explored for emotion recognition, among which we can mention hidden Markov models, multilayer perceptron (MLP), support vector machines, k-nearest neighbour, bayesian classifiers [8,1,17].

In this paper, restricted Boltzmann machines (RBM) and deep belief networks (DBN) are used in spoken emotion recognition because they are novel to this task. 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 [18,12]. In [24], a *generalized discriminant analysis* based on DBN showed significant improvement over support vector machines using nine databases. However, Brueckner [5] found that the RBM helped in the task but the 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. A regression-based DBN to learn features directly from magnitude spectra is employed in [23]. In that work, the DBN was able to learn representative features from a sparse representation of the spectrum for music emotion recognition. Also the DBN are used for classification of emotions based on lip shape [19]. These are used to initialize a feed-forward neural network during an unsupervised training phase. Kim et al. [16] use deep learning techniques to explicitly capture complex non-linear feature interactions in multimodal data. Their promising results in emotion classification suggest that DBN can learn the high-order non-linear relationships from diverse sources.

In the present work, multilayer perceptron based classifiers and deep classifiers are implemented to classify emotional signals. The behaviour of the proposed classifiers is assessed in speaker-independent and text-independent schemes.

The remainder of the paper is organised as follows: in Section 2 the baseline and the deep classifier are presented; Section 3 gives an introduction of the emotional database used; in Section 4 the experiments are presented, where we explain the features extraction process and the validation schemes; then, the Section 4 exposes the detailed configurations used in the classifiers; Section 5 shows the results; and finally, Section 6 presents conclusions and discusses possibilities for future work.

2 Classifiers

In this section we first present our baseline method, and then, we introduce the deep learning methods.

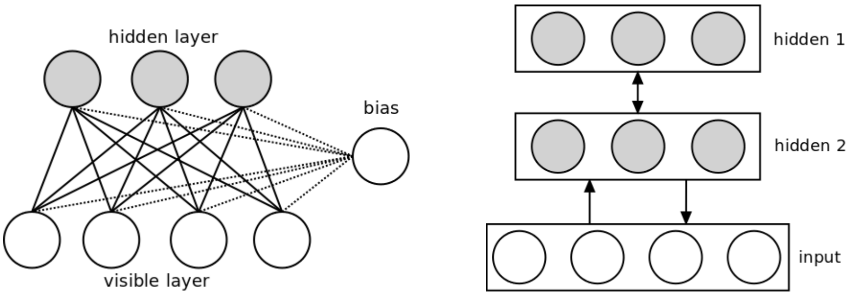
2.1 Multilayer perceptron

Classifiers based on multilayer perceptron were widely used in emotion recognition [8,17] and they are useful as baselines. MLP is a class of artificial neural network and it consists of a set of process units (simple perceptrons) arranged in layers. In the MLP, the nodes are fully connected between layers without connections between units in the same layer. The input vector (feature vector) feeds into each of the first layer perceptrons, the outputs of this layer feed into each of the second layer perceptrons, and so on [11]. The output of the neuron is the weighted sum of the inputs plus the bias term, and its activation is a function (linear or nonlinear) as

$$y = \mathcal{F}\left(\sum_{i=1}^{n} \omega_i x_i + \theta\right)$$

where x_i are the inputs, w_i are the weights and θ is the bias. The network is trained with the back-propagation algorithm, using the error between the current output and the desired output [11].

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
E. M. Albornoz, M. Sánchez-Gutiérrez, F. M. Martínez, H. L. Rufiner & J. Goddard; "Spoken emotion recognition using deep learning"
19th Iberoamerican Congress on Pattern Recognition (CIARP 2014), nov, 2014.



(a) Restricted Boltzmann machine.

(b) Deep belief network.

Fig. 2: Deep classifiers

2.2 Deep learning

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 2(a). The hidden units are usually binary stochastic and the visible units are typically binary or Gaussian stochastic. An RBM represents the joint distribution between a visible vector and a hidden random variable. An RBM only has connections between the units of the two layers, and with the bias unit. One reason for this is that efficient training algorithms have been developed for this *restricted* version(c.f. Hinton's contrastive divergence algorithm [14]) 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, the energy function $E(v,h)$ is defined by:

$$E(v,h) = -a'v - b'h - h'Wv$$

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:

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

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

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

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

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

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
E. M. Albornoz, M. Sánchez-Gutiérrez, F. M. Martínez, H. L. Rufiner & J. Goddard; "Spoken emotion recognition using deep learning"
19th Iberoamerican Congress on Pattern Recognition (CIARP 2014), nov, 2014.

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

$$p(v_j = 1|h) = \sigma(a_i + \sum_j h_j w_{i,j}) \quad \text{and} \quad p(h_j = 1|v) = \sigma(b_j + \sum_i v_i w_{i,j})$$

where

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

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 [13]. 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. [5]) as:

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

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

$$p(h_j = 1|v) = \sigma(\sum_i \frac{v_i}{\sigma_i^2} w_{ij} + b_j)$$

$$p(v_i = v|h) = \mathcal{N}(v| \sum_j h_j w_{ij} + a_i, \sigma_i^2)$$

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

Deep belief networks As Bengio [3] 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. [18,12]). A DBN consists in a number of stacked RBM, as shown in Fig. 2(b). Hinton et al. [15] 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.

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
E. M. Albornoz, M. Sánchez-Gutiérrez, F. M. Martínez, H. L. Rufiner & J. Goddard; "Spoken emotion recognition using deep learning"
19th Iberoamerican Congress on Pattern Recognition (CIARP 2014), nov, 2014.

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
E. M. Albornoz, M. Sánchez-Gutiérrez, F. M. Martínez, H. L. Rufiner & J. Goddard; "Spoken emotion recognition using deep learning"
19th Iberoamerican Congress on Pattern Recognition (CIARP 2014), nov, 2014.

3 Emotional speech database

In spite of the fact that the present goal is to achieve emotion recognition from spontaneous speech, the development of spontaneous-speech datasets is very expensive and they are commonly restricted. Acted emotional expressions may not sound like real expressions, however, these are an interesting approach, especially if the dataset naturalness is judged by expert listeners. Under these assumptions, we employed a well-known emotional acted database developed at the Communication Science Institute of Berlin Technical University [6] and used in several studies [4,1,26][3]. The corpus, consisting of 535 utterances, includes sentences performed under 6 discrete emotions (and neutral emotional state) distributed as: Anger(127), Boredom(81), Disgust(46), Fear(69), Joy(71), Sadness(62) and Neutral(79). The same sentences were recorded in German by 10 actors, 5 females and 5 males. The corpus consists of 10 utterances for each emotion type, from 1 to 7 s. A perception test with 20 individuals was carried out to ensure the emotional quality and naturalness of the utterances, and the most confusing utterances were eliminated [6]. Here, all utterances belonging to the same class are labelled with the name of the class and their transcriptions are ignored. Each one stands for an unique training or validation pattern in a data partition.

4 Experiments

In this section we describe the feature extraction stage and the validation schemes, then we present the configurations used in order to train and test the classifiers.

For every emotional utterance, mel-frequency cepstral coefficients (MFCCs) and prosodic features were extracted. We chose MFCCs because they are the most popular representation used in speech recognition [20] and they are extensively used in emotion recognition [2,8,17]. These are based on a linear model of voice production together with a codification in a psychoacoustic scale [20]. On the other hand, the use of prosodic features in emotion recognition has already been studied and discussed extensively [4]. Here the energy, zero crossing rate and fundamental frequency (F_0) were considered. The first 12 mean MFCCs, the mean F_0, the average of the zero crossing rate and the mean of the energy, plus the means of first derivatives of each one were extracted using the *OpenSMILE* [9]. Hence, each utterance is represented by a 30-dimensional vector in all the experiments.

We propose two validation methodologies for the emotion classifiers, one to ensure speaker independence and other to deal with text independence. Consequently, considering the characteristics of the corpus, ten partitions were obtained for the speaker independent experiments and eight partitions were obtained for the text independent experiments. A leave-one-out scheme was performed for both cases. For both schemes, LOTO (leave one-text-out) and LOSO (leave one-speaker-out), the MLP was used as baseline. The MLPs have one hidden layer with ((# features + # classes)/2) neurons and these were apply using

[3] It is freely accessible at `http://pascal.kgw.tu-berlin.de/emodb/`.

Table 1: Classification results for LOTO and LOSO schemes.

Classifier	LOTO accuracy (avg)	LOSO accuracy (avg)
Multilayer Perceptron	**68.10** [%]	**51.65** [%]
DBN-RBM	**69.14** [%]	**60.32** [%]

Weka Toolkit [10]. A 10% of the training set was left for the generalization test. The MLP training was stopped at 500 epochs or when the network reached the generalization peak with test data [11].

DBN experiments were performed by adding one additional RBM classification layer to a previously trained DBN and using the optimal parameters founded in a previous work with a corpus in Spanish (after a large set of exploratory experiments) [22]. The parameters for RBM/DBN training are *Batch size*=42, *Learning rate*=0.00001, *Hidden units*=112 and *Number of layers*=(1 + RBM). All RBM had Gaussian units and the classification layer had seven output units, one for each class. For deep classifier experiments, we use the toolbox developed by Drausin Wulsin [25]. The deep classifiers were trained up to the generalization peak, with balanced test data, were reached.

5 Results and discussion

In this section, the results of the proposed classifiers on both schemes are presented and discussed. Table 1 shows the performance of classifiers for LOSO and LOTO experiments. In the first column, the classifier is displayed. The second and third columns present the average accuracy of each classifier for LOTO and LOTO tasks. Results indicate that deep classifiers perform better than MLP in both schemes. Furthermore, in LOSO scheme the improvement is really significant (8.67% over the baseline). As can be seen, the emotions are quite dependent on the speaker and the results are better (LOTO) when speaker independence is not taking into account. These results suggest that the DBN could be used in the more difficult schemes and moreover, there is an important correlation between the emotion elicitation and specific speakers.

We have also evaluated the statistical significance of these results by computing the probability that a given experiment is better than our baseline classifier [21]. In order to perform this test we assumed the statistical independence of the classification errors for each utterance and we approached the errors' Binomial distribution by means of a Gaussian distribution. In this way, for the LOSO scheme we have that the confidence of the relationship obtained between error rates of DBN and MLP (reference) is

$$\mathsf{Pr}(err < err\_ref) > 99.85\%$$

On the other hand, the improvement using the LOTO scheme is not significant.

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
E. M. Albornoz, M. Sánchez-Gutiérrez, F. M. Martínez, H. L. Rufiner & J. Goddard; "Spoken emotion recognition using deep learning"
19th Iberoamerican Congress on Pattern Recognition (CIARP 2014), nov, 2014.

6 Conclusions and future work

In this work we evaluated the restricted Boltzmann machines and deep belief networks in spoken emotion recognition. We proposed two validation methodologies in order to ensure speaker independence and text independence. A feature set based on spectral and prosodic characteristics was used. Results show that the deep classifiers are better than MLP classifiers, in both LOSO and LOTO schemes.

In future works the deep classifier will be tested with noisy signals.

7 Acknowledgements

Authors wish to thank: *SEP* and *CONACyT* (Program SEP-CONACyT CB-2012-01, No.182432) and the *Universidad Autónoma de México* 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.

References

1. Albornoz, E.M., Milone, D.H., Rufiner, H.L.: Spoken emotion recognition using hierarchical classifiers. Computer Speech & Language 25(3), 556–570 (2011)
2. Batliner, A., Steidl, S., Schuller, B., Seppi, D., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Aharonson, V., Kessous, L., Amir, N.: Whodunnit - Searching for the most important feature types signalling emotion-related user states in speech. Computer Speech & Language 25(1), 4–28 (2011)
3. Bengio, Y.: Learning Deep Architectures for AI. Foundations and Trends® in Machine Learning 2(1), 1–127 (jan 2009)
4. Borchert, M., Dusterhoft, A.: Emotions in speech - experiments with prosody and quality features in speech for use in categorical and dimensional emotion recognition environments. In: Proc. of IEEE Int. Conference on Natural Language Processing and Knowledge Engineering (NLP-KE). pp. 147–151 (Oct 2005)
5. Brueckner, R., Schuller, B.: Likability classification - a not so deep neural network approach. pp. 1–4. INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association, Portland, USA (2012)
6. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B.: A Database of German Emotional Speech. In: Proc. of 9th European Conference on Speech Communication and Technology (Interspeech). pp. 1517–1520 (Sep 2005)
7. Devillers, L., Vidrascu, L.: Speaker Classification II: Selected Projects, Lecture Notes in Computer Science, vol. 4441/2007, chap. Real-Life Emotion Recognition in Speech, pp. 34–42. Springer-Verlag, Berlin, Heidelberg (2007)
8. El Ayadi, M., Kamel, M., Karray, F.: Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition 44(3), 572–587 (2011)
9. Eyben, F., Wöllmer, M., Schuller, B.: Opensmile: the munich versatile and fast open-source audio feature extractor. In: Proceedings of the international conference on Multimedia. pp. 1459–1462. MM '10, ACM, New York, NY, USA (2010)

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
E. M. Albornoz, M. Sánchez-Gutiérrez, F. M. Martínez, H. L. Rufiner & J. Goddard; "Spoken emotion recognition using deep learning"
19th Iberoamerican Congress on Pattern Recognition (CIARP 2014), nov, 2014.

10. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. SIGKDD Explor. Newsl. 11(1), 10–18 (Nov 2009)
11. Haykin, S.: Neural Networks: A Comprehensive Foundation. Prentice Hall, 2nd edn. (Jul 1998)
12. Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., Kingsbury, B.: Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. Signal Processing Magazine, IEEE 29(6), 82–97 (2012)
13. Hinton, G.E.: A practical guide to training restricted boltzmann machines. In: Montavon, G., Orr, G.B., Müller, K.R. (eds.) Neural Networks: Tricks of the Trade, Lecture Notes in Computer Science, vol. 7700, pp. 599–619. Springer Berlin Heidelberg (2012)
14. Hinton Geoffrey E.: Training Products of Experts by Minimizing Contrastive Divergence. Neural Computation 14(8), 1771–1800 (2002)
15. Hinton Geoffrey E., Osindero Simon, Teh Yee-Whye: A Fast Learning Algorithm for Deep Belief Nets. Neural Computation 18(7), 1527–1554 (2006), doi: 10.1162/neco.2006.18.7.1527
16. Kim, Y., Lee, H., Provost, E.M.: Deep learning for robust feature generation in audiovisual emotion recognition. In: ICASSP. pp. 3687–3691. IEEE (2013)
17. Koolagudi, S., Rao, K.: Emotion recognition from speech using source, system, and prosodic features. International Journal of Speech Technology 15, 265–289 (2012)
18. Mohamed, A., Sainath, T., Dahl, G., Ramabhadran, B., Hinton, G., Picheny, M.: Deep belief networks using discriminative features for phone recognition. In: IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5060–5063 (2011)
19. Popovic, B., Ostrogonac, S., Delic, V., Janev, M., Stankovic, I.: Deep architectures for automatic emotion recognition based on lip shape. In: The 12th Int. Scientific-Professional Symposium (INFOTEH). Bosnia and Herzegovina (mar 2013)
20. Rabiner, L., Juang, B.H.: Fundamentals of Speech Recognition. Prentice-Hall, Inc., Upper Saddle River, NJ, USA (1993)
21. Rufiner, H.L., Torres, M.E., Gamero, L.G., Milone, D.H.: Introducing complexity measures in nonlinear physiological signals: application to robust speech recognition. Physica A: Statistical Mechanics and its Applications 332(1), 496–508 (2004)
22. Sanchez-Gutierrez, M., Albornoz, E.M., Martinez-Licona, F., Rufiner, H.L., Goddard, J.: Deep learning for emotional speech recognition. In: 6th Mexican Conference on Pattern Recognition. Cancún, México (Jun 2014), [Accepted]
23. Schmidt, E.M., Kim, Y.E.: Learning emotion-based acoustic features with deep belief networks. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). pp. 65–68. IEEE, New Paltz, NY, USA (2011)
24. Stuhlsatz, A., Meyer, C., Eyben, F., ZieIke, T., Meier, G., Schuller, B.: Deep neural networks for acoustic emotion recognition: Raising the benchmarks. In: IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP). pp. 5688–5691 (2011)
25. Wulsin, D.: DBN Toolbox v1.0. Department of Bioengineering, University of Pennsylvania (2010), `http://www.seas.upenn.edu/~wulsin/`
26. Yang, B., Lugger, M.: Emotion recognition from speech signals using new harmony features. Signal Processing 90(5), 1415–1423 (2010), Special Section on Statistical Signal & Array Processing
27. Yildirim, S., Narayanan, S., Potamianos, A.: Detecting emotional state of a child in a conversational computer game. Computer Speech & Language 25(1), 29 – 44 (2011)

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
E. M. Albornoz, M. Sánchez-Gutiérrez, F. M. Martínez, H. L. Rufiner & J. Goddard; "Spoken emotion recognition using deep learning"
19th Iberoamerican Congress on Pattern Recognition (CIARP 2014), nov, 2014.