Hierarchical Clustering and Classification of Emotions in Human Speech Using Confusion Matrices

Manuel Reyes-Vargas[1], Máximo Sánchez-Gutiérrez[1], Leonardo Rufiner[2],
Marcelo Albornoz[2], Leandro Vignolo[2],
Fabiola Martínez-Licona[1], and John Goddard-Close[1]

[1] Universidad Autonoma Metropolitana, Electrical Engineering Depto., Mexico City, Mexico
[2] Centro de I+D SINC(i), Universidad Nacional del Litoral / CONICET, Argentina
{manuel.reyesvargas,edmax86,lrufiner,albornoz.marcelo,
leandro.vignolo}@gmail.com, {fmml,jgc}@xanum.uam.mx

Abstract. Although most of the natural emotions expressed in speech can be clearly identified by humans, automatic classification systems still display significant limitations on this task. Recently, hierarchical strategies have been proposed using different heuristics for choosing the appropriate levels in the hierarchy. In this paper, we propose a method for choosing these levels by hierarchically clustering a confusion matrix. To this end, a Mexican Spanish emotional speech database was created and employed to classify the 'big six' emotions (anger, disgust, fear, joy, sadness, surprise) together with a neutral state. A set of 14 features was extracted from the speech signal of each utterance and a hierarchical classifier was defined from the dendrogram obtained by applying Wards clustering method to a certain confusion matrix. The classification rate of this hierarchical classifier showed a slight improvement compared to those of various classifiers trained directly with all 7 classes.

Keywords: Emotional speech, confusion matrix, hierarchical clustering.

1 Introduction

Speech communication provides the most significant information for humans through the emission of thoughts, ideas and even emotions. Due to the dynamic nature of the speech signal it has high levels of variability: speech production depends on the location and movement of the elements of the vocal tract and the face, and variations of parameters such as local accents, social status or personal style [1]. In this process, the emotional state which is expressed in the spoken words, enhances the message's content. It has been claimed that words account for less than 10% of the meaning of the message for the listener [2], so the analysis of features like prosody, rhythm or voice quality has become important.

Emotions are hard to define in theory, although features like time durations and intensity are most useful in the identification of emotions like joy, anger, and sadness, even though differences in perception between different people may arise. In trying to differentiate emotions, efforts have also focused on describing them in terms of activation and valence indexes, where the amount of energy needed to express the emotion

M. Železný et al. (Eds.): SPECOM 2013, LNAI 8113, pp. 162–169, 2013.

and its correlates to the nervous system are studied [3,4]. This leads to a representation of the archetypal emotions as a kind of elemental emotions palette, similar to the case of the primary colors [5]; these elemental emotions are anger, disgust, fear, joy, sadness and surprise, and the different combinations of them give rise to a wider variety of emotional states. Although most of the natural emotions expressed in speech can be clearly identified by a human, automatic classification systems still display significant limitations on this task. In this paper, automatic classification of Mexican Spanish emotional speech is undertaken, using a hierarchical classifier which is defined from a confusion matrix using Wards method. This gives a method for defining the hierarchies found in the classifier.

In the last few years, several emotional speech databases have been created for purposes such as emotion recognition and expressive speech synthesis, human emotion perception, or to produce virtual teachers [6]. Most of them are recorded in English, but there are also a few in German [7] and Spanish [8], where the recordings may contain natural speech, acted speech, or both. Limitations of these databases have often been reported when they are perceptually tested by humans because of factors such as the poor emotional simulation, the variable quality of the recordings, and the lack of phonetic transcriptions of the phrases. In the particular case of Spanish, the reported databases represent the Spanish language as spoken in Spain, and the creation of databases with different variants of the Spanish language is important in order to obtain a better understanding of the emotional content of speech.

The characteristics which are most commonly extracted from the speech signal for analysis purposes are derived from fundamental frequency, the time durations of phonemes, syllables, or words, energy and formants [9]. By using features which are based on these characteristics it has been possible to identify in general most of the big six emotions in English and Spanish [10]. Some other parameters, as well as variations of the previously mentioned ones, have been applied to the classification of emotions. In [11], pitch based features were used to recognize emotions in German with reasonably good results for six emotions with a Bayesian classifier. Time level recognition has been utilized for classification with support vector machines [12], and the influence of speaking rate in speech emotion recognition has been studied in [13], as well as Gaussian mixture models to classify natural, acted and mixed emotional speech [14]. There are approaches that focus on the search for the optimal set of features including acoustic and linguistic ones [15], or by creating a hybrid system that includes neural networks, fuzzy systems and genetic algorithms [16]. Hierarchical emotion recognition schemes are a novel set of methods for the analysis of the speech signal. Some works use a binary decision tree approach with acoustic features [17], prosodic, spectral and glottal flow features [18], or multiple feature methods [19]. The choice of hierarchical levels tends to be heuristically motivated. Clustering techniques for a hierarchical conversion has been used for speech synthesis [20].

In this paper a Mexican Spanish emotional speech corpus was created, influenced by [21,22], and used to classify the big six emotions. Considering the variety of possible feature selection and classification systems, our approach uses 14 features together with a hierarchical binary classifier. The hierarchical levels are found by applying Wards method to a certain confusion matrix.

2 Methods

A Mexican Spanish emotional speech database was created and a set of features were defined, in order to classify the six emotions of anger (a), disgust (d), fear (f), joy (j), sadness (sa), surprise (s) and a neutral state (n).

2.1 Data

A Mexican male professional speaker recorded three sets of speech data. Each set had 40 words selected from the Swadesh list for Spanish [23], and 40 sentences that included each word. Swadesh, originally devised by the linguist Morris, contains words that are present in almost all languages and form the basis for communication between humans. The sets of selected words included nouns (numbers, colors, animals, body parts, etc), pronouns and verbs.

The sentences were based on each word of the list and contained the complete structure: subject, verb and predicate. Although there are some emotional speech databases that aim to be phonetically balanced (using nonsense phrases) [8], in our case the objective of using these sentences was to allow the speaker to express himself better in terms of the emotions considered; this was previously agreed on with him. The texts that were recorded belonged to segments from Benito Prez Galdoses novel "Fortunata y Jacinta" (Fortunata and Jacinta), Miguel de Cervantes Saavedra's short novel "La espaola inglesa" (The Spanish English) and Octavio Paz's poem "Primer da" (First day). The texts had around 450 words on average and did not contain any dialog; the poem has 94 words. The recordings were carried out on a desktop PC using the Speech Filing System Version 4.8 [24] with a sampling frequency of 16 KHz; the amplitude and noise levels were controlled by the speaker.

2.2 Features and Classification

14 features were extracted from the speech signal of each utterance using the averages of the first 12 MFCC, fundamental frequency F0 and log energy coefficients. A total of 1562 examples were obtained. Our objective was to form and test a hierarchical binary classifier on the seven emotion classes contained in our data. In order to do this we had to decide how to form the binary hierarchy, and then which binary classifiers to employ at each juncture. The way we automatically found the hierarchy was to first train a classifier using the seven classes, and then apply Ward's hierarchical clustering method to the resulting confusion matrix. The corresponding dendrogram provides the hierarchy. This idea was proposed in [25], for the automatic generation of topic hierarchies.

The Ward's method is one of the hierarchical clustering methods most used in the literature [26,27]. It is a greedy, agglomerative hierarchical method, that determines a diagram, called a dendrogram, that shows the sequence of mergings of clusters into larger clusters. It seeks to form the partitions in a manner that minimizes the loss of information associated with each merging. The information loss is quantified in terms of an error sum of squares criterion, so Wards method is often referred to as the minimum variance method. We used the usual euclidean distance measure for determining the inter-class similarity. Finally, support vector machines (SVM) were taken as the binary classifiers at each juncture.

3 Results

Firstly, we obtained the cross validation error rate, using ten partitions, for several classifiers and all seven emotion classes. The results are shown in Table 1. Here we can observe that the best classifier, SVM, achieves a classification rate of 34.7%. We used Weka [28] to obtain these results, where J48 is a version of Quinlan's C4.5 algorithm, k-NN refers to k nearest neighbors, MLP to a multilayer perceptron. The parameters employed for SVM were a radial basis function kernel with gamma = 0.1, and C=100. In the case of the MLP, the parameters were the standard ones used in Weka.

Table 1. Classification rates with all of the 7 emotions

Classifier	% Error Rate
J48	51.34
1-NN	50.83
3-NN	48.98
5-NN	45.71
7-NN	43.79
SVM	34.7
MLP	37.45

We took the confusion matrix of the worst classifier, J48, as shown below in Table 2, and applied Ward's hierarchical method to it. The idea of using the worst classifier was to emphasize the differences that occur in the confusion matrix. We also tried, however, confusion matrices from the other classifiers and got similar results in terms of the dendrograms.

Table 2. Classification rates with all of the 7 emotions

	Neutral	Joy	Sadness	Anger	Fear	Disgust	Surprise
Neutral	120	2	47	4	34	35	1
Joy	6	136	0	29	5	12	55
Sadness	42	1	110	0	63	26	1
Anger	1	41	2	139	1	3	16
Fear	34	4	58	0	70	33	4
Disgust	29	7	43	2	35	84	24
Surprise	2	58	2	21	3	16	101

The dendrogram given by Ward's method is shown in figure 1. We can see that the clusters which are obtained are successively: {a,j,s} and {d,n,sa,f}, {a} and {j,s}, {d} and {n,sa,f}, {n} and {sa,f}.

Rearranging the confusion matrix according to the clusters {a,j,s} and {d,n,sa,f}, gives the confusion matrix shown in Table 3. We can observe that a certain order has been brought to this table where lower values are found in the lower left and upper right submatrix blocks, something difficult to perceive in Table 2.

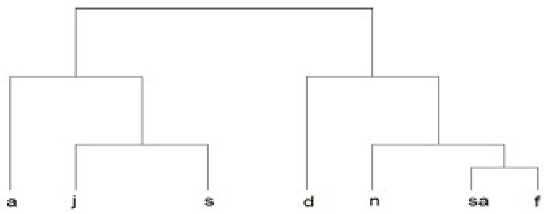

Fig. 1. Dendrogram obtained from a confusion matrix using Ward's method

The initial separation with Ward's method gives the following two clusters: {f,sa,n,d} and {j,s,a}. We also applied the clustering technique of k-means, with k=2, and in this case got the clusters: {n,sa,f}, {j,a,d,s}; as we can observe, the emotion disgust changes from one cluster to the other. We can ask what difference this makes in terms of the classification.

To try to answer this, we treated the data as two, two-class problems, with the classes formed in each case by the two different partitions which we found using Ward's method and k-means. We then found the cross validation rates for the classifiers given in Table 4.

Table 3. Rearranged confusion matrix of J48

	Fear	Sadness	Neutral	Disgust	Joy	Surprise	Anger
Fear	70	58	34	33	4	4	0
Sadness	63	110	42	26	1	1	0
Neutral	34	47	120	35	2	1	4
Disgust	35	43	29	84	7	24	2
Joy	5	0	6	12	136	55	29
Surprise	3	2	2	16	58	101	21
Anger	1	2	1	3	41	16	139

Table 4. Classification rates with two 2-class problems given by the clustering

Classifier	% Error Rate	% Error rate
	{n,sa,f,d}, {j,a,s }	{n,sa,f},{j,a,d,s}
J48	5.7	12.48
1-NN	7.1	13.12
3-NN	6.34	11.14
5-NN	6.53	10.37
7-NN	5.76	10.24
SVM	3.27	9.86
MLP	4.55	10.76

We see that the error rate was worse for each classifier. This suggests that the choice of clustering method employed, and hence also the hierarchical classifier constructed, will make an important difference to the result.

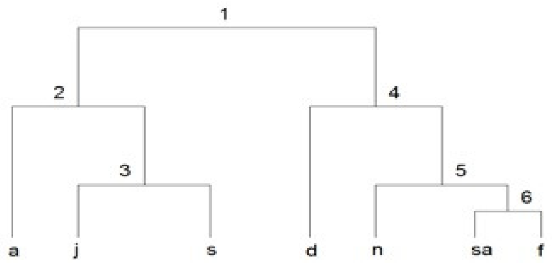

Fig. 2. Hierarchical classifier obtained from the dendrogram in Figure 1

Finally, from the dendrogram, we obtain the following binary hierarchical classifier, shown in figure 2. The numbers signify a different SVM binary classifier in each juncture where: 1 separates the classes formed by {a,j,s} and {d,n,sa,f}, 2 separates the classes formed by {a} {j,s}, etc. In total we have a hierarchical classifier formed using 6 binary classifiers. The error rate obtained using this classifier was 33.59%.

4 Conclusions

In this paper, we have applied Ward's hierarchical clustering method to a confusion matrix which we obtained from a classifier employed on the 7 class problem of a Mexican Spanish emotional speech database. We then used the corresponding dendrogram to define a hierarchical classifier, placing a binary SVM at each juncture. This gave a more principled way of defining the hierarchical structure of the classifier. We have seen that the choice of dendrogram, and so the hierarchy obtained, is important in defining the resulting classifier, as using a different clustering method (the 2-means algorithm) to cluster the confusion matrix produced worse classification results. Finally, the hierarchical classifier was applied to the 7 class problem and produced slightly better results than all of the other classifiers. It should be noted that very little parameter tuning was done to the binary classifiers, only binary SVM classifiers were applied at each juncture, and the same number of data features was used at all levels of the hierarchical classifier. It is to be expected that classification results could be improved by taking some of these factors into consideration and the authors hope to do this, as well as employing other emotional speech databases, in future work.

Acknowledgement. This work was supported by the SEP and CONACyT, through the Program SEP-CONACyT CB-2012-01, No.182432, and by the UAM, in Mexico, as well as CONICET, UNL, and UNER from Argentina.

References

1. Benzeghiba, M., De Mori, R., Deroo, O., et al.: Automatic speech recognition and speech variability: A review. Speech Communication 49, 763–786 (2007)
2. Mehrabian, A.: Communication without words. Psychology Today 2, 53–56 (1968)
3. Williams, C., Stevens, K.: Vocal correlates of emotional states. In: Speech Evaluation in Psychiatry. Grune and Stratton (1981)
4. Fernandez, R.: A computational model for the automatic recognition of affect in speech. Ph.D. Thesis, Massachussetts Institute of Technology (2004)
5. Cowie, R., Douglas, E., Tsapatsoulis, N., Kollias, S., Fellenz, W., Taylor, J.: Emotion recognition in human-computer interaction. IEEE Signal Process. Mag. 18, 32–80 (2001)
6. Ververdis, D., Kotropoulos, C.: A Review of Emotional Speech Databases. Department of Informatics, Aristotle University, Greece (2003)
7. Burkhardt, F., Paeschke, A., et al.: A database of German emotional speech. In: Proceedings of the Interspeech, Lisbon, pp. 1517–1520 (2005)
8. Barra-Chicote, R., Montero, J.M., Macias-Guarasa, J., Lufti, S., Lucas, J.M., Fernandez, F., D'haro, L.F., San-Segundo, R., Ferreiros, J., Cordoba, R., Pardo, J.M.: Spanish Expressive Voices: Corpus for Emotion Research in Spanish. In: Proc. of 6th International Conference on Language Resources and Evaluation (LREC 2008), Morocco (2008)
9. Ei Ayadi, M.: Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition 44, 572–587 (2011)
10. Muñoz, A., Jiménez, F.: La expresion de la emocón a traés de la conducta vocal. Revista de Psicología General y Aplicada 43, 289–299 (1990)
11. Yang, B., Lugger, M.: Emotion recognition from speech signals using new harmony features. Signal Processing 90, 1415–1423 (2010)
12. Schuller, B., Rigoll, G.: Timing levels in segment-based speech emotion recognition. In: Proceedings of Interspeech, Pittsburg, pp. 1818–1821 (2006)
13. Phlilippou-Hübner, D., Vlasenko, B., Böck, R., Wendemuth, A., von Guericke, O.: The performance of the speaking rate parameter in emotion recognition from speech. In: Proceedings of IEEE International Conference on Multimedia and Expo, Melbourne, pp. 248–253 (2012)
14. Sungrack, Y., Chang, Y.: Loss-scaled large-margin Gaussian mixture models for speech emotion classification. IEEE Transactions on Audio, Speech and Language Processing 20, 585–598 (2012)
15. Batliner, A., Stedi, S., Schuller, B., et al.: Whodunnit - searching for the most important feature types signaling emotion-related user states on speech. Computer and Speech Language 25, 4–28 (2011)
16. Gharavian, D., Scheikhan, M., Nazeriech, A., Garoucy, S.: Speech emotion recognition using FCBF feature selection method and GA-optimized fuzzy ARTMAP neural network. Neural Computer & Applications 21, 2115–2126 (2012)
17. Lee, C., Mower, E., Busso, C., Lee, S., Narayanan, S.: Emotion recognition using hierarchical decision tree approach. Speech Communication 53, 1162–1171 (2011)
18. Giannoulis, P., Potamianos, G.: A hierarchical approach with feature selection for emotion recognition from speech. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation, Istanbul, pp. 1203–1206 (2012)
19. Albornoz, E., Milone, D., Rufiner, H.: Spoken emotions using hierarchical classifiers. Computer Speech and Language 25, 556–570 (2011)
20. Chung-Hsien, W., Chi-Chun, H., Chung-Han, L., Mai-Chun, L.: Hierarchical prosody conversion using regression-based clustering for emotional speech synthesis. IEEE Transactions on Audio, Speech and Language Processing 18, 1394–1405 (2010)

21. Vaughan, B., Cullen, C.: Emotional speech corpus creation, structure, distribution and re-use. In: Young Researchers Workshop in Speech Technology (YRWST 2009), Dublin (2009)
22. Van Eyne, F., Gibbon, D. (eds.): Lexicon Development for Speech and Language Processing. Springer (2000)
23. Swadesh lists for Spanish, `http://en.wiktionary.org/wiki/Appendix:Spanish_Swadesh_list`
24. Speech Filing System, University College London, `http://www.phon.ucl.ac.uk/resource/sfs/`
25. Godbole, S.: Exploiting Confusion Matrices for Automatic Generation of Topic Hierarchies and Scaling Up Multi-Way Classifiers. Technical report, IIT Bombay (2002)
26. Everitt, B.S., Landau, S., Leese, M., Stahl, D.: Cluster Analysis. John Wiley & Sons Inc. (2011)
27. Gan, G., Ma, C., Wu, J.: Data Clustering Theory, Algorithms, and Applications. ASA-SIAM Series on Statistics and Applied Probability. SIAM, Philadelphia (2007)
28. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA Data Mining Software: An Update. SIGKDD Explorations 11 (2009)
29. van der Maaten, L.J.P., Hinton, G.E.: Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research 9, 2579–2605 (2008)