Automatic Emotion Recognition using Prosodic Parameters

Iker Luengo, Eva Navas, Inmaculada Hernáez, Jon Sánchez

Department of Electronics and Telecommunication
University of the Basque Country, Spain
{ikerl, eva, inma, ion}@bips.bi.ehu.es

Abstract

This paper presents the experiments made to automatically identify emotion in an emotional speech database for Basque. Three different classifiers have been built: one using spectral features and GMM, other with prosodic features and SVM and the last one with prosodic features and GMM. 86 prosodic features were calculated and then an algorithm to select the most relevant ones was applied. The first classifier gives the best result with a 98.4% accuracy when using 512 mixtures, but the classifier built with the best 6 prosodic features achieves an accuracy of 92.3% in spite of its simplicity, showing that prosodic information is very useful to identify emotions.

1. Introduction

With the progress of new technologies and the introduction of interactive systems, there has been a sudden increase in the demand of user friendly human-machine interfaces. As speech is the natural way of communication for humans, in order to achieve a comfortable interface, it is necessary to provide these interfaces with speech generation and recognition mechanisms. Various systems have been already implemented in this way, from avatars and modern interactive entertainment toys to automatic customer service systems, and many researches are being held in this field [1][2][3].

Nowadays, one of the major goals in this kind of interfaces is the naturalness in the communication between human and machine: High quality text to speech engines and more flexible recognition grammars are needed. But, as humans tend to express their emotional state through the voice, capability to generate emotional speech and to recognize the mood of the speaker are also needed in this kind of natural interfaces.

The purpose of this work is to make a first approach towards emotion recognition in speech for the Basque language. More precisely, the objective is to compare the performance of the recognizers when using different feature sets (traditional Cepstral features and prosodic features) and classifier experts (Gaussian Mixture Models, GMM and Support Vector Machines, SVM).

The paper is organized as follows: First a description of the database used during these experiments is given. Then, the process of extracting prosodic features is summarized, followed by a description of the different experiments that were made, together with the results obtained in these experiments. Finally, these results are discussed.

2. Database Description

For these experiments, the emotional speech database for Basque recorded by the University of the Basque Country was used [4]. This database includes the six emotions considered as the *basic* ones [5], namely anger, fear, surprise, disgust, joy and sadness.

A professional dubbing actress was hired for the recordings, i.e. acted speech was used. In order to verify whether the emotions were correctly expressed, a subjective evaluation of the database was carried out. Results in this evaluation show that the emotions are recognized over 70% of the cases, except for the case of disgust, that has also been difficult to identify in other languages [6][7].

The text corpus recorded in the database is divided into two different parts: The first one is phonetically balanced, and includes emotion independent texts, which are common for all emotions. In this part neutral style was also recorded, in order to use it as a reference. The other part contains texts semantically related to each emotion, thus, the texts are different among emotions. Neutral style was not considered in this part.

The common part of the database allows an easier way to compare the acoustical characteristics of the emotions, as the texts are the same in all of them. On the other hand, texts semantically related to the emotion may help the dubbing actress to express that emotion in a more natural way.

One of the goals in these experiments is to find out whether prosodic features can be used to distinguish between neutral and emotional speech, and furthermore, to identify the expressed emotion. Therefore, only the first part of the database was used, as it is the only part in which neutral style was recorded.

The corpus in the used part of the database contains numbers, isolated words and sentences of different length. Table 1 shows a summary of this content. Overall, 97 recordings are available for each emotion, with a total of approximately seven minutes of speech per emotion.

Recordings were made in a professional studio, using 32 kHz sample rate and 16 bits per sample. A laryngograph was used in order to get the glottal closure signal synchronized with the speech. In this way, highly accurate intonation curves can be extracted.

Table 1: Number and type of items for each emotion

Type of item	Number
Isolated digits	21
Isolated words	21
Short affirmative sentences	10
Short interrogative sentences	5
Medium affirmative sentences	22
Medium interrogative sentences	8
Long sentences	10
Total items per emotion	97

3. Extraction of Prosodic Features

For this very first experience, only pitch and energy related features were used. Figure 1 shows how these features were extracted. Intonation and energy curves are extracted from the recordings, both in linear and logarithmical scale. First and second derivative curves are calculated, as the pitch and energy change rate may provide new useful information for the recognition. Finally, different statistical features are computed over these curves, with the help of the voiced/unvoiced and voice activity information.

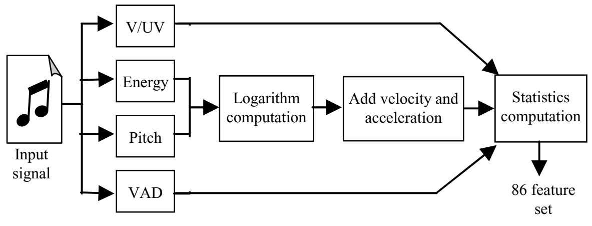

Figure 1: Prosodic feature extraction diagram.

3.1. Intonation Curve Estimation

A highly accurate intonation curve was estimated for each of the utterances, using the signal recorded with the laryngograph. Pitch values were computed as the inverse of the time elapsed between two consecutive glottal closures. The intonation curves were estimated with a 1kHz sample frequency, i.e. one pitch sample per millisecond.

The voiced/unvoiced (V/UV) information was extracted by detecting the segments where there was glottal closure information and in the ones where the glottis was kept open.

3.2. Power Curve Estimation

Recordings were analysed with windows of 25 ms every 10 ms and the mean power was calculated for each windowed frame. This gives a power sample every 10 ms. All curves were normalized to the mean value of the neutral style.

3.3. Voice Activity Estimation

Voice activity estimation is necessary in order to reject those frames in which there is no vocal information. In this way, noise level during speech silences will not corrupt the calculated features. A voice activity detector (VAD) was implemented, based on the computation of the long term spectral deviation (LTSD) between vocal and noisy frames. The implemented system is based in the one presented in [8], in which an adaptive decision threshold is used in order to get the best performance for each noise level.

3.4. Jitter and Shimmer Estimation

Jitter and shimmer are related to the micro-variations of the pitch and power curves. So, they can be estimated as the slope change rate for these curves. In this work, jitter and shimmer were computed as the number of zero crossings of the derivative curves. The result was normalized to the number of frames used for this computation, in order to take into account the length of the utterance.

3.5. Statistical Feature Computation

As there is no *a priori* knowledge about which features are best for emotion recognition, it was considered best to compute a large number of features, and discard those which are redundant later on.

Once power and pitch curves and their derivatives were estimated, different statistical features were computed for each of them:

- Mean value
- Variance
- Maximum value
- Minimum value
- Range
- Skew
- Kurtosis

Computation of pitch related features was accomplished using only frames for which pitch value existed, discarding unvoiced frames. In a similar way, energy related features were computed using only frames for which the VAD estimated vocal activity.

For each utterance 12 curves had been extracted (pitch, power, their logarithmic versions, and the first and second derivatives of all of them), and for each curve 7 statistical features were computed. This gives 84 features per utterance. When jitter and shimmer are added, a total of 86 prosodic features per utterance are obtained.

4. Experiments and results

Three different classification experiments were made: The first one, with a traditional GMM system using spectral features; the second, with SVM and prosodic features; and the last one with a GMM system with prosodic features.

The precision of each system was computed using a Jack-knife test. First of all, the utterances corresponding to each emotion were randomized, and then, divided into five blocks. This randomization ensures that the blocks are balanced, as the database contains different types of utterances (from isolated words to long sentences). Then, five different systems were trained, and five tests were made, with a leave-one-out method. When all five tests were finished, the overall confusion matrix and classification accuracy were calculated. This accuracy was estimated as the number of correctly classified utterances normalized to the total number of utterances in the tests.

4.1. GMM Models with Spectral Features

The recordings from the database were converted into Mel frequency Cepstral coefficients (MFCC), with first and second derivatives, using frames of 25 ms every 10 ms, with Hamming windowing and pre-emphasis factor of 0.97.

HTKv3 software [9] was used for GMM training and testing. Training of the first model sets, with just one and two Gaussian mixtures, was accomplished by three loops of Baum-Welch training. The remaining model sets were trained from the set with two mixtures, in a loop process in which the mixture number was increased in two and two loops of Baum-Welch re-estimation were applied, until the desired number of mixtures was reached.

Figure 2 shows the accuracy obtained in this experiment, for different number of mixtures. As expected, the accuracy of the system increases with the number of mixtures, until saturation. Table 2 shows the confusion matrix for the 512 mixture case.

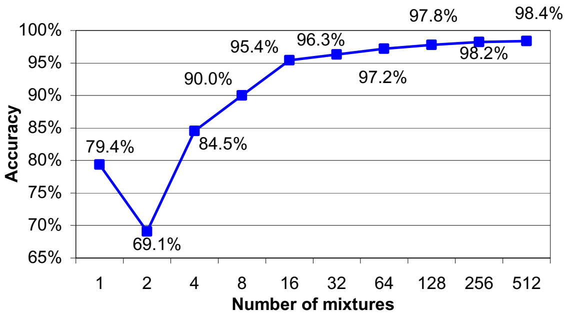

Figure 2: Recognition accuracy with MFCC features and GMM, for different number of mixtures.

Table 2: Confusion matrix with MFCC and GMM with 512 mixtures

		INPUT						
		Ang	Fea	Sur	Dis	Joy	Sad	Neu
OUTPUT	Ang	97	-	-	-	-	-	1
	Fea	-	96	-	-	-	-	-
	Sur	-	1	97	-	-	-	-
	Dis	-	-	-	93	-	-	-
	Joy	-	-	-	-	93	-	-
	Sad	-	-	-	1	-	97	1
	Neu	-	-	-	3	4	-	95
	Acc(%)	100	99.0	100	95.9	95.9	100	97.9

4.2. SVM Models with Prosodic Features

LibSVM v2.6 function library [10] was used for the training and testing of SVM. A RBF kernel and one-against-one approach were used for multi-class classification.

In a first experiment, all 86 features that had been extracted were used, achieving an overall accuracy of 93.50%. Nevertheless, it is expectable that many of these features are redundant, even more if we take into account that two versions of the same curves of pitch and power are used to calculate these statistical features, one in lineal scale and the other in logarithmical. Therefore, a feature selection process was implemented.

A Forward 3-Backward 1 wrapper method was used for the selection of features. During the process, the feature which maximizes the system's accuracy is selected in each step, where the accuracy is obtained by training a whole new classifier with a Jack-knife test. After three consecutive selections have been made, the least useful feature is taken out, the one which, once eliminated, reduces the system's performance least.

Results of this experiment are shown in Figure 3. It is clear that, even if using fewer features gives a slightly worse performance than using all 86, the computational cost of extracting all these features and training such a complex system may not be worthy. With as few as six features, 92.32% of accuracy is obtained, only 1.18% less than using all 86 features. Table 3 presents the confusion matrix for the case in which only the *best* six features were used. These features were:

- Mean pitch
- Mean energy
- Pitch variance
- Skew of logarithmic pitch
- Range of logarithmic pitch
- Range of logarithmic energy

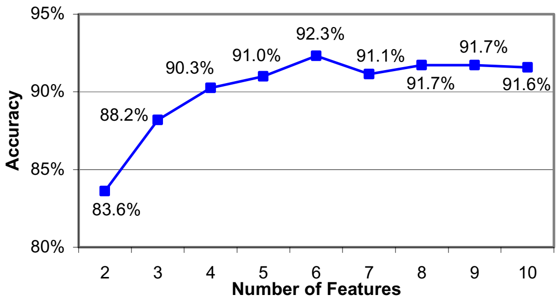

Figure 3: Recognition accuracy with prosodic features and SVM, for different number of features.

Table 3: Confusion matrix with prosodic features and SVM with the best 6 features

		INPUT						
		Ang	Fea	Sur	Dis	Joy	Sad	Neu
OUTPUT	Ang	92	-	-	1	2	-	-
	Fea	-	94	9	-	-	-	-
	Sur	-	3	86	-	-	-	-
	Dis	-	-	-	80	-	4	3
	Joy	2	-	-	-	88	-	1
	Sad	2	-	-	10	-	93	1
	Neu	1	-	-	6	7	-	92
	Acc(%)	94.9	96.9	90.5	82.5	90.7	95.9	94.9

4.3. GMM Models with Prosodic Features

Once again, HTKv3 software was used for GMM training and testing. Because of the use of a Jack-knife test and due to the fact that a single feature vector was obtained from each utterance, only about 80 vectors were available for model training. So, single mixture models were used, due to the lack of training material. These models were created by three loops of Baum-Welch training.

Two experiments were made, one with all 86 features, and the other one with just the 6 features that gave best result in the SVM experiment. While the system with 86 features obtained an accuracy of 84.79%, the one with 6 features improved the performance up to 86.71%. This increase in the performance when using fewer features can be due to overtraining of the models, because of the use of too many features for so little training material. Table 4 and Table 5 show the confusion matrixes for both cases.

Table 4: Confusion matrix with prosodic features and GMM with 86 features

		INPUT						
		Ang	Fea	Sur	Dis	Joy	Sad	Neu
OUTPUT	Ang	88	1	1	3	4	-	-
	Fea	3	89	5	-	-	-	-
	Sur	4	13	78	-	-	-	-
	Dis	1	-	-	76	3	7	10
	Joy	4	-	-	1	68	-	24
	Sad	-	1	-	7	-	89	-
	Neu	-	-	-	2	8	1	86
	Acc(%)	90.7	91.8	82.1	78.4	70.1	91.8	88.7

Table 5: Confusion matrix with prosodic features and GMM with 6 features

		INPUT						
		Ang	Fea	Sur	Dis	Joy	Sad	Neu
OUTPUT	Ang	89	2	4	-	4	-	-
	Fea	-	90	8	-	-	-	-
	Sur	1	5	83	-	-	-	-
	Dis	2	-	-	73	0	14	1
	Joy	4	-	-	-	82	-	8
	Sad	-	-	-	14	-	83	14
	Neu	1	-	-	10	11	-	87
	Acc(%)	91.8	92.8	87.4	75.3	84.5	85.6	89.7

5. Conclusions

Results in emotion recognition experiments are hard to compare, because different database designs are used: Some use acted speech, whereas others collect real emotions, some are multi-speaker and others are not, different basic emotions sets are considered...

Possibly the work presented in [1] is the closest to this one, as acted speech and just one speaker are used, with the same list of emotions. In this paper, Hozjan and Kacic describe the use of the Interface database to train and test an speaker dependent emotional speech classifier based on a total of 144 prosodic features. Accuracy between 60-90% is achieved depending on the language. Other works in emotion recognition in speech get 63.5% in a multi-speaker framework [2], 80.7% with only 3 emotions and neutral style [3] and 82.5% in a multi-speaker environment using HMM [11].

Regarding to the capability of prosodic features to distinguish among emotions, from the results obtained with these experiments, it is clear that traditional GMM based emotional speech classifiers using spectral features achieve a higher accuracy that those which use only prosodic information. Nevertheless, it is noticeable that a simple SVM classifier with only six prosodic features reaches an accuracy of 92.32%, only 6% less than a GMM-MFCC system with 512 mixtures. This error increase may be compensated by the time reduction obtained in the training and testing processes.

As acted speech was used, emotions may be overacted, and classification results in realistic conditions may be worst. Nevertheless, it is a good starting point to see how prosody may help in emotion recognition tasks.

These results are very encouraging. As long term prosodic features seem to be very little correlated with short term spectral features [12], it is expected that the use of both kind of parameters reduces even more the classification error.

Future work includes considering other types of features, such as phoneme duration in order to estimate the speaking rate. Classifier experts fusion will also be held, so that time varying spectral information and long term prosodic features work together to reduce the system error.

6. Acknowledgments

This work has been partially funded by the Spanish Ministry of Science and Technology (TIC2003-08382-C05-03) and the University of the Basque Country (UPV-0147.345-E-14895/2002).

7. References

[1] Hozjan, V., Kacic, Z., "*Improved Emotion Recognition with Large Set of Statistical Features*", Proc. Eurospeech'03, pp 133-136, 2003.

[2] Petrushin, V.A., "*Emotion Recognition in Speech Signal: Experimental Study, Development and Application*", Proc. ICSLP'00, pp 222-225, 2000.

[3] Seppänen, T., Väyrynen, E, Toivanen, J., "*Prosody Based Classification of Emotions in Spoken Finnish*", Proc. Eurospeech'03, pp 717-720, 2003.

[4] Navas, E., Hernáez, I., Castelruiz, A., Luengo, I., *"Obtaining and Evaluating an Emotional Database for Prosody Modelling in Standard Basque"*, Lecture Notes on Artificial Intelligence 3206: 393-400, 2004.

[5] Cowie, R., Cornelius, R., "*Describing the Emotional States that are Expressed in Speech*", Speech Communication 40(1,2): 2-32, 2003.

[6] Iida, A., Campbell, N., Higuchi, F., Yasumura, M.: A *"Corpus-based Speech Synthesis System with Emotion"*. Speech Communication 40(1,2): 161-187, 2003.

[7] Burkhardt, F., Sendlmeier, W.F.: *"Verification of Acoustical Correlates of Emotional Speech using Formant-Synthesis"*. Proc. ISCA Workshop on Speech and Emotion, pp. 151-156, 2000.

[8] Ramirez, J., Segura, J., Benitez, C., de la Torre, A., Rubio, A., "*Efficient Voice Activity Detection Algorithms Using Long Term Speech Information*", Speech Communication 42: 271-287, 2004.

[9] Young, S., Odell, J., Ollason, D., Valchev, V., Woodlans, P., *The HTK book*, Cambridge University, Cambridge, 2000.

[10] Chang, Ch. and Lin, Ch. *LIBSVM: a Library for Support Vector Machines*, 2005, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[11] Nogueiras, N., Moreno, A., Bonafonte, A., Mariño, J., "*Speech Emotion Recognition Using Hidden Markov Models*", Proc. Eurospeech'01, pp. 2679-2682, 2001.

[12] Campbell, J., Reynolds, D., Dunn, R., "*Fusing High and Low Level Features for Speaker Recognition*", Proc. Eurospeech'03, pp. 2665-2668, 2003.