PROSODIC AND SEGMENTAL RUBRICS IN EMOTION IDENTIFICATION

R. Barra, J.M. Montero, J. Macías-Guarasa, L.F. D'Haro, R. San-Segundo, R. Córdoba

Speech Technology Group, Dept. of Electronics Engineering, Universidad Politécnica de Madrid
E.T.S.I. Telecomunicación, Ciudad Universitaria s/n, 28040-Madrid, Spain
{barra, juancho, macias, lfdharo, lapiz, cordoba}@die.upm.es

ABSTRACT

It is well known that the emotional state of a speaker usually alters the way she/he speaks. Although all the components of the voice can be affected by emotion in some statistically-significant way, not all these deviations from a *neutral* voice are identified by human listeners as conveying emotional information.

In this paper we have carried out several perceptual and objective experiments that show the relevance of prosody and segmental spectrum in the characterization and identification of four emotions in Spanish.

A Bayes classifier has been used in the objective emotion identification task. Emotion models were generated as the contribution of every emotion to the build-up of a *Universal Background Emotion Codebook*.

According to our experiments, *surprise* is primarily identified by humans through its prosodic rubric (in spite of some automatically-identifiable segmental characteristics); while for *anger* the situation is just the opposite. *Sadness* and *happiness* need a combination of prosodic and segmental rubrics to be reliably identified.

1. INTRODUCTION

As speech recognition rates and speech synthesis naturalness approach the performance of humans, some issues that were known to be involved in speech production and perception but were regarded as secondary, now deserve a new close-up in speech research. Expressive or emotional speech is one of these new relevant issues.

Emotions are mental states that, consciously or not, can catalyze physiological reactions for appraisal or adaptation purposes [1][2]. Some of these prototypical patterns of change can be externally perceived because one of the main functions of human emotions is communicative. Among these physiological alterations that can be used for conveying emotional information to another person, voice changes are some of the most important ones.

Most experiments carried out in emotion identification try to determine the features that best predict the emotional content of one voice recording using the standard training-testing paradigm [3][4]. However, it will be shown that people are able to identify the emotion using some emotion-specific patterns. These prototypical voice rubrics are significantly different from one another, allowing the emotional state to be communicated, but they are not necessarily the only features with relevant variations from one emotion to another one, or from an emotional voice to a *neutral* one [5].

The paper is organized as follows: first a description of the database and the perceptual experiments is provided. Then, the automatic emotion identification process is described, from feature extraction to automatic emotion modeling and discussion of the results. Finally, some conclusions are drawn.

2. DATABASE DESCRIPTION AND HUMAN EVALUATION EXPERIMENTS

In this work, the *Spanish Emotional Speech database (SES)* [9] is being used. It contains two emotional speech recording sessions played by a professional male actor in an acoustically treated studio. Each recorded session includes thirty words, fifteen short sentences and three paragraphs, simulating three basic or primary emotions (*sadness*, *happiness* and *anger*), one secondary emotion (*surprise*) and a *neutral* speaking style. The text uttered by the actor did not convey any intrinsic emotional content. Finally, the recorded database was phonetically labeled in a semiautomatic way.

The assessment of the emotional voice was aimed at judging the appropriateness of the recordings as a model for recognizable emotional speech [6]. Table 1 includes the confusion matrix of the identification task using real speech from the actor, and the results show that the subjects had no difficulty in identifying the emotion that was simulated by the professional actor, and the diagonal numbers are clearly above the chance level. A Chi-square test refutes the null hypothesis, with a confidence level above 95%, that it could not have been obtained from a random selection experiment.

The sum of each line in the table is not 100 %, because the unidentified option was also offered to the listeners, and sometimes it was chosen.

The last row of Table 1 shows the *precision* of the identification task, defined as the number of times the listeners correctly identified that emotion divided by the

Authorized licensed use limited to: Univ de Alcala. Downloaded on May 8, 2009 at 09:34 from IEEE Xplore. Restrictions apply.

number of times they identified it (correctly or incorrectly). Although intended *happiness* and *neutral* were the most difficult emotions to identify, the precision figures show that *sadness* was the less precisely identified emotion because of its confusion with *happiness* and *neutral*. In spite of its low recall figure (76.2%), *neutral* voice has the highest precision (the listeners had a certain tendency to over-identify emotions).

Table 1: Emotion identification rates by human listeners in the SES database

	IDENTIFIED EMOTION				
INTENDED EMOTION	***Happiness***	***Anger***	***Surprise***	***Sadness***	***Neutral***
Happiness	61.9%	7.9%	11.1%	9.5%	3.2%
Anger		95.2%			
Surprise			90.6%	1.6%	3.2%
Sadness	7.9%		4.8%	81.0%	
Neutral	3.2%	6.3%	1.6%	7.9%	76.2%
PRECISION	84.8%	87.0%	83.8%	81.0%	92.3%

This first experiment shows the ability of human listeners in the emotion identification task in spite of not being familiar with the voice or the patterns of the actor. However, we would need to know what clues in his voice were responsible for this easy identification.

In order to answer this question, we designed another experiment: Twenty one listeners, non of which was used to listening synthetic speech, were involved in a re-synthesis test comprising 5 *neutral*-content sentences from the database. As 4 emotions and a *neutral* voice had to be evaluated, 25 different recordings per listener were used. In these experiments, only one listening session per subject was allowed and no feedback was provided along the test to avoid the listeners to learn the patterns in a supervised way. In each session, the audio recordings of the stimuli were presented to the listener in a random way and each text segment of text was played up to 3 times.

The stimuli recordings were generated by re-synthesizing *neutral* recordings with emotional prosody and vice versa, by re-synthesizing emotional recordings with a *neutral* prosody, using a PSOLA-like prosody-modification algorithm. When using *neutral* speech with superimposed emotional prosody, *surprise* was the best identified emotion (76.2%), followed by *sadness* (66.65%), *happiness* (19%) and *anger* (7.1%). When using *neutral* prosody superimposed to emotional speech, *anger* was the best identified emotion (95.2%), followed by *happiness* (52.4%), *sadness* (45.2%) and *surprise* (9.5%).

This experiment gives interesting clues regarding the clues responsible for emotion identification: *surprise* is mostly identified through its prosodic rubric, *anger* is communicated by segmental patterns in our actor's speech, and there is a mixture of both factors in the identification of *sadness* and *happiness*.

In the rest of the paper, we will describe the task of evaluating till what extent these findings are actually supported by objective evidence, by using an automatic emotion identification system. This system will use, in turn, segmental or/and prosodic related features, so that we will be able to determine the importance of both rubrics in the identification of every emotion, and their correlation with the experiments using humans in the identification task.

3. EMOTION IDENTIFICATION SYSTEM

In this work we are not concerned with the optimization of an emotion identification task, but with relating the findings of the perceptual experiments to the results of an automatic emotion identification system. Due to that, instead of using a great amount of features, related parameters and complex pattern recognition classifiers as other research groups [10], we will base the paper on the wide qualitative studies described in [6] and summarized in the previous section.

3.1. Feature extraction

In this work, we have used the following features related to the segmental and prosodic rubrics described above. More precisely:

- Segmental features**:** We selected the MFCC as representative features of the segmental information, which have demonstrated a reasonable performance in similar emotion identification tasks [4].
- Prosodic features**:** Following the results of [7], we decided to focus on F0 related features, leaving energy, duration and speech-rate-related features for future work. Six parameters have been computed from the F0 contour: average F0, F0 standard deviation, F0 average variation, minimum F0, maximum F0 and F0 range.

The segmental and suprasegmental information has been extracted using different strategies:

- The MFCC segmental features have been calculated using Matlab Auditory Toolbox (by Interval Research Corp.), extracting 13 MFCC per frame, using 25 ms window length every 10ms, with a Hamming windowing and a pre-emphasis factor of 0.97.
- The prosodic features have been calculated by processing voiced segments: Every paragraph has been automatically split in voiced and unvoiced segments using PRAAT. The F0 contour was extracted every 8 ms and then divided to voiced segments. This strategy (instead of calculating prosody features at frame rate such as in [8]) provides a more robust suprasegmental information, while reducing the number of available vectors for training.

All the training features were extracted from the paragraphs in the SES corpus and then, the identification process has been carried out using the available sentences. In the MFCC case this is not a problem, but when using the prosodic features, this will affect the identification rates because the prosody contours in the paragraphs are smoother than in the sentences (the reason for this is that the actor had more time to transmit the intended emotion).

Authorized licensed use limited to: Univ de Alcala. Downloaded on May 8, 2009 at 09:34 from IEEE Xplore. Restrictions apply.

3.2. Emotion Modeling

We have modeled every emotion as its contribution to each of the 256 centroids to a *Universal Background Emotion Codebook (UBEC)*, generated with the training vectors of all emotions. In [8][9], an alternative modeling for emotion identification in the SES database is used, where the same emotions where modeled, excluding *surprise*, with a similar strategy. Figure 1 shows the general procedure for model training.

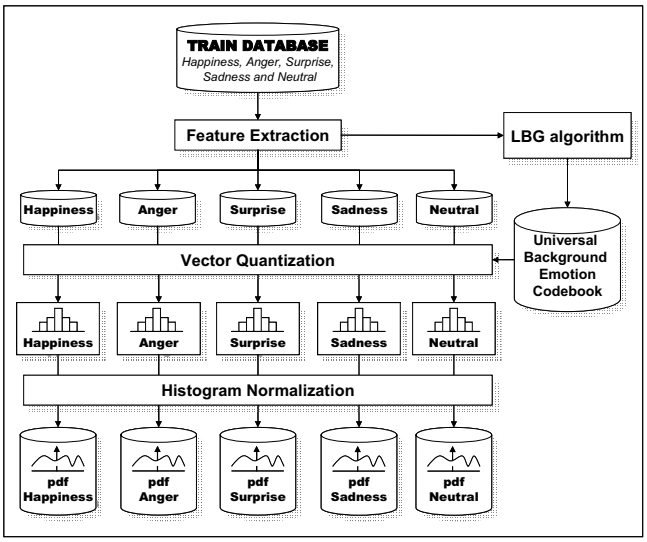

Figure 1: Emotion model generation procedure

In order to evaluate models using segmental and prosodic information, two models were generated for each emotion in the SES database:

- The first model was obtained using around 140,000 MFCC vectors.
- The second model was obtained from 6,100 vectors composed of the 6 F0-related features, extracted from the F0 contour of the voiced segments, as described above.

This significant difference in the amount of training data would lead to a different behavior depending on which model is used.

3.3. The bayesian classifier

We have used a Bayes classifier to implement our emotion recognizer, in which the probability of a certain emotion given the acoustic evidence is calculated by accumulating the probabilities of frames/voiced_segments depending on the classifier being used.

3.4. Identification results and discussion

We have carried out two automatic emotion identification experiments using the classifier described in the previous sections. The first one is based on MFCC features (emotion identification using segmental models related to the segmental rubrics mentioned in section 2) and the second one is based on prosody (using supra-segmental models related to the prosodic rubrics mentioned in section 2).

As mentioned in section 3.1, segmental and prosodic training information were extracted from the paragraphs and the evaluation was carried out using the short sentences of the SES corpus. This mismatch between the training and test sets, in addition to a smaller amount of data in the prosodic model generation, has contributed to achieve worse results with the prosodic classifier.

Table 2 shows the confusion matrix obtained when trying to identify the underlying emotion with the classifier based on segmental information, and Table 3 shows the confusion matrix when using the prosodic one.

Table 2: Confusion matrix when using segmental features for emotion identification

	IDENTIFIED EMOTION				
INTENDED EMOTION	***Happiness***	***Anger***	***Surprise***	***Sadness***	***Neutral***
Happiness	91.1%		6.7%		2.2%
Anger		97.8%			2.2%
Surprise	2.2%	28.9%	68.9%		
Sadness				100.0%	
Neutral		26.7%			73.3%
PRECISION	97.6%	67.7%	91.2%	100.0%	91.7%

Table 3: Confusion matrix when using the prosodic features for emotion identification

	IDENTIFIED EMOTION				
INTENDED EMOTION	***Happiness***	***Anger***	***Surprise***	***Sadness***	***Neutral***
Happiness	44.4%	17.8%	24.4%		13.3%
Anger	31.1%	48.9%	4.4%		15.6%
Surprise	4.4%		95.6%		
Sadness		6.7%		75.6%	17.8%
Neutral	20.0%	13.3%			66.7%
PRECISION	47.6%	59.5%	76.8%	100.0%	48.8%

In the subjective evaluation sessions, *happiness* was the most difficult emotion to identify. However, attending to the segmental classifier, *happiness* identification rate is high (91.1%), with also a high precision (97.6%). In the prosodic experiment, *happiness* has been the worst identified emotion (44.4%); with a high confusion rate (47.6% precision).

Anger is the second best identified emotion in the segmental experiment (97.8% accuracy). However, it is the less precisely identified emotion (67.7%), with a high confusion with *neutral* (26.7%) and *surprise* (28.9%). Prosodically, *anger* is the second worst identified emotion (48.9%) and the precision is also low when using F0-related information (59.5%).

Surprise is poorly identified by segmental means (68.9%), but the identification is very precise (91.2%). On the contrary, using the prosodic recognizer, *surprise* is the best identified emotion (95.6%), with the second highest

Authorized licensed use limited to: Univ de Alcala. Downloaded on May 8, 2009 at 09:34 from IEEE Xplore. Restrictions apply.

precision (76.8%).

The nature of *sadness*, according to the objective experiments, has proven to be unique, with a high identification accuracy and high precision by both segmental and prosodic means. The automatic system was able to outperform humans, due to the speaker-dependent training.

To summarize the results from the two experiments:

- The identification based in segmental information shows high accuracy for *happiness*, *anger* and *sadness*; and lower for *surprise* and *neutral* (which are less segmentally-salient). Prediction precision is higher for *anger*: the trained system exhibits a certain tendency to over-identify *anger* (the emotion with most segmental emotion) and to under-identify *neutral* and *surprise* (the less segmental), and most of the confusion involves exactly this 3 emotions.
- The identification based on prosodic features shows that *surprise* (the most prosodic according to subjective evaluation) and *sadness* are the best identified emotions. Low-pitched *sadness* is never confused, while high-pitched *happiness* is the most easily-confused emotion. The estimation of pitch in a fully automatic way is very difficult for a menacing *anger*.

Finally, we performed an experiment of score fusion using a simple weighted linear combination, leading to the results shown in Table 4, in which we can see that the fused classifier outperforms the results obtained by the other two.

Table 4: Confusion matrix when fusing the segmental and prosodic based classifiers

	IDENTIFIED EMOTION				
INTENDED EMOTION	***Happiness***	***Anger***	***Surprise***	***Sadness***	***Neutral***
Happiness	95.6%		2.2%		2.2%
Anger	2.2%	95.6%			2.2%
Surprise		6.7%	93.3%		
Sadness				100.0%	
Neutral		13.3%			86.7%
PRECISION	97.7%	86.0%	97.7%	100.0%	92.9%

A detailed analysis of the results summarized in Table 4 showed that the spectral and prosodic information modeling is complementary, so that the combination of both sources improves the final identification accuracy.

4. CONCLUSIONS

In this paper we have carried out several perceptual and objective experiments that show the relevance of prosody and segmental spectrum in the characterization and identification of four emotions in Spanish.

According to our experiments, *surprise* is primarily identified by humans through its prosodic rubric (in spite of some automatically-identifiable segmental characteristics); while for *anger* the situation is just the opposite: it is better transmitted and detected using segmental patterns. Finally, *sadness* and *happiness* need a combination of prosodic and segmental rubrics to be reliably identified. The results of the proposed classifier are validated as they are clearly in line with the qualitative behavior of human listeners in the emotion identification task.

Objective evaluation shows that the information from both the prosodic and segmental spectrum rubrics is complementary, so that both of them should be used in order to improve emotional speech processing systems (specially taking into account that we provide a perceptual based explanation for such complementary behavior).

5. ACKNOWLEDGMENTS

This work has been partially funded by the Spanish Ministry of Science and Technology under contracts DPI2004-07908-C02-02 (ROBINT) and TIN2005-08660-C04-04 (EDECAN-UPM).

6. REFERENCES

[1] Scherer, K. R., "Personality markers in speech by K. R. Scherer and H. Giles (eds)". Social markers in speech. Cambridge: Cambridge University Press. 1979.

[2] Cowie, R. and Cornelius, R.R. "Describing the emotional states that are expressed in speech, in Speech Comnication",40:5-32. 2003.

[3] Nogueiras, A., Moreno, A., Bonafonte, A. and Mariño, J.B. "Speech Emotion Recognition Using Hidden Markov Models", in Proc. of EUROSPEECH, pp 2679-2682, 2001.

[4] Luengo, I., Navas, E., Hernáez, I., Sánchez, J, "Automatic Emotion Recognition using Prosodic Parameters", in Proc. of INTERSPEECH, pp. 493-496, 2005.

[5] Escudero-Mancebo, D., González-Farreras, C. and Cardeñoso-Payo, V., "Quantitative evaluation of relevant prosodic factors for text-to-speech synthesis", in Proc. Of ICSLP, pp. 1165-1168. 2002.

[6] Montero, J.M., Gutiérrez-Arriola, J., Colás, J., Enríquez, E., Pardo J.M., "Analysis and Modelling of Emotional Speech in Spanish", Proc. of the XIVth International Congress of Phonetic Science, vol. II pp. 957-960, 1999.

[7] Montero, J.M., Gutiérrez-Arriola, J., Córdoba, R., Enríquez, E., Pardo, J.M. "The role of pitch and tempo in emotional speech" in *Improvements in speech synthesis*, pp. 246-251. Ed. Wiley & Sons, 2002.

[8] Amir, N., Kerret O. and Karlinski, D. "Classifying emotions in speech: a comparison of methods", Proc. of EUROSPEECH, pp.127-130, 2001.

[9] Amir, N., Ron, S., Laor, N. "Analysis of an emotional speech corpus in Hebrew based on obective criteria", ISCA workshop on speech and emotion, Belfast, 2000.

[10] Montero, J.M., Gutiérrez-Arriola, J., Colás, J., Macías-Guarasa, J., Enríquez E. and Pardo, J.M. "Development of an Emotional Speech Synthesiser in Spanish". Proc. of EUROSPEECH 1999, vol. 5, pp. 2099-2102, 1999

[11] Schuller, B., Müller, R., Lang, M., Rigoll, G. "Speaker Independent Emotion Recognition by Early Fusion of Acoustic and Linguistic Features within Ensembles", Proc. of INTERSPEECH 2005, pp. 805-809, 2005.

Authorized licensed use limited to: Univ de Alcala. Downloaded on May 8, 2009 at 09:34 from IEEE Xplore. Restrictions apply.