ANALYSIS AND MODELLING OF EMOTIONAL SPEECH IN SPANISH

J.M. Montero*, J. Gutiérrez-Arriola*, J. Colás*, E.Enríquez** and J.M.Pardo*
*Grupo de Tecnología del Habla - Departamento de Ingeniería Electrónica E.T.S.I. Telecomunicación- Univ. Politécnica de Madrid, Ciudad Universitaria s/n, 28040 Madrid, Spain
**Grupo de Tecnología del Habla -Departamento de Lengua Española - Universidad Nacional de Educación a Distancia, Ciudad Universitaria s/n, 28040 Madrid, Spain
E-mail: juancho@die.upm.es

ABSTRACT

The importance of speech prosody for conveying emotional information has been extensively underlined in the literature. Major elements such as pitch, tempo and stress are presented as the main acoustic correlates of emotion in human speech. Nevertheless, as several authors have shown, voice quality is also a relevant feature in emotion recognition. In this paper, we present the prosodic analysis, modelling and evaluation of the Spanish Emotional Speech Database including four emotions: happiness, sadness, cold anger and surprise. Our results show that, for Spanish, the contribution of prosody to the recognisability of the uttered emotion greatly varies from one to another, with sadness and surprise being more supra segmental, and happiness and cold anger being rather segmental.

1. INTRODUCTION

Intelligibility is not the focus of current research in speech synthesis. Nowadays, an essential point in recognition and synthesis tasks is the addressing of the variability of human speech, with special emphasis on modelling different personal speaking styles [3].

One of the main sources of this diversity is the emotional state or attitude of the speaker. Speech under emotional [6][7] or stress [8] conditions can be modelled as a deviation from neutral voice, both in a prosodic and in a voice-quality way.

Qualitative descriptions of the typical prosody of emotional speech are present in many papers. Happiness is characterised by an increase in F0 mean and range; sadness includes a decrease in the same factors (a narrow dynamic range of the pitch contour indicates lack of speaker's excitement), probably with a slower articulation rate and lower RMS energy; and so on [5][9].

In the VAESS project (Voices Attitudes and Emotions in Synthetic Speech) [4], we recorded a database of simulated emotional speech in Spanish, and we developed a portable communication device, capable of transmitting the emotional state of the user, through the use of the GLOVE's voice source that implements Fant's model [2].

In this paper, we will present the prosodic analysis, modelling and evaluation of the Spanish Emotional Speech Database and an experiment on the relevance of voice quality in emotional state recognition scores.

2. THE DATABASE

2.1. Collected data

Two sessions (3 passages and 15 sentences of neutral-content text) were recorded by a professional actor simulating four emotions (happiness, sadness, cold anger and surprise) and a neutral speaking style. The recordings were phonetically and prosodically hand-labelled [1]. Up to 2000 phonemes per emotion are available for analysis.

2.2. Emotional data description

During the segmentation phase, we observed that happy recordings exhibited an important differential factor: for the same sentence or phrase, the actor used several clearly distinct intonation contours, depending on the focus word that he decided to choose. An example of this is shown in Fig.2, where the same sentence under the same emotional conditions is uttered with two different F0 patterns.

As our intonation model did not include this variability, we divided happy sentences into three groups: those sentences that could be characterised by a positive declination line, those ones with a rising intonation and, finally, the examples with a focus in the middle of the sentence. In order to achieve a coherent training of the intonation models, only the first ones were used for training and evaluation. As the same sentence can be uttered with different patterns, this reduction of the variability of the speech data cannot be a source of emotional recognition confusion.

The surprise emotion (not previously processed in the course of VAESS) exhibits the highest F0 mean values among the emotions we are considering (with values that could be attributed to a female voice).

Cold anger files has one unique characteristic that makes them easily identifiable: the presence of a pitch-correlated noise that makes the pitch periods quite difficult to mark. A considerable amount of vibrato or tremor in the speech signal can also be observed. Fig 3 shows a comparison of the same frames of a phoneme in a neutral and a cold angry style; the correlated nature of the voice source noise is caused by the way of forcing the glottis (not the vocal tract) in the cold anger emotion.

Finally, sadness was simulated by the actor according to the descriptions in the literature (flat low-pitched F0 contour and slow rhythm)

2.3. Database Evaluation

We conducted an assessment experiment using the association method. Three copy-synthesis sentences were listened by 21 adult people in a random-order forced-choice test (including a "non-identifiable" option). After each sentence in the evaluation process, they had to indicate what they assumed to be the emotional state of the speaker.

In copy-synthesis experiments, we used a concatenative synthesiser with diphones and stylised prosody from natural speech. The confusion matrix diagonal obtained was:

Emotion	Recognition rate
Neutral	76.2%
Happy	61.9%
Sad	81.0%
Angry	95.2%
Surprise	90.5%

Table 1. Recognition rate for the copy-synthesis evaluation experiment.

As can be observed from the table, happy recordings are the most difficult to recognise. The copy-synthesis results are significantly above random-selection level using a Student's test ($p>0.95$). The use of an automatic process for copying the prosody, and the distortion introduced by prosody modification algorithms, can reduce the recognition scores. See [1] for an experiment on natural recordings evaluation, although the results are not directly comparable, due to the evaluation of a new emotion in this copy-synthesis test. It is remarkable that cold anger resynthesised sentences were evaluated above natural recordings: the prosody-modification could made the voice to be even more menacing.

3. PROSODIC DATA ANALYSIS AND MODELLING

3.1. Duration analysis and modelling

Phoneme durations were analysed through a multiplicative model including up to 123 parameters, (considering both intrinsic duration and contextual coefficients). Parameters were estimated by error minimisation in the log domain, and optimised in the linear domain using the Levenberg-Marquardt method.

The factors that the model takes into account are: current phone identity, class (articulation manner) of the post vocalic phoneme, the position in the phrase (pre-pausal or not), the type of syllable, the accent status and the length of the word.

Paragraphs were uttered in a slower way (between 11% and 21%) for all emotions, suggesting that there is no absolute value for the intrinsic rhythm of each emotion. Only for sadness, the difference in contextual coefficients is more than than a 10%.

In [4], it is shown that there are no significant differences between the recognition rate of the paragraphs and the isolated sentences in the database (but surprise was not included in that study).

Not only phonemes but also the duration of the pauses influences the global rhythm associated to each emotion. There is a clear division between an emotion with less activity or excitement (sadness) and the other ones (happiness, surprise and cold anger). The small difference between neutral and sadness pauses reflects the fact that the actor decided to accelerate the active emotions, but not to exaggerate the slow tempo of sadness.

Coefficients	Happy/ Neutral	Sad/ Neutral	Surprise/ Neutral	Angry/ Neutral
Consonant	0.92	1.06	1.02	0.98
Semicons	0.96	1.16	1.12	1.05
Vowel	0.99	1.10	1.10	0.91
Prepause Shortening.	0.99	0.86	0.93	1.06
Number of syllables	1.01	1.13	1.19	1.09
Intrinsic duration	1.04	1.26	1.14	1.22
Vowels	1.06	1.02	1.11	1.10
Diphthongs	1.11	1.53	1.25	1.49
Consonants	1.11	1.17	1.12	1.10

Table 2 Duration coefficients for the isolated sentences (ratio between neutral and emotional speech)

Pauses duration	Neutral	Happy	Sad	Surpr	Angr
Before '.'	910	420	1176	547	578
Mean Deviation	167	70	182	75	55
Other pauses	514	316	697	346	376
Mean Deviation	137	82	131	74	87

Table 3 Pause mean duration (in ms) and deviation

3.2. Intonation analysis and modelling

For the intonation analysis, we used a simple model that divides each breadth group into three areas separated by the first and last stressed vowels.

Fifteen parameters were computed by RMS error minimisation. Table 4 summarises the results of intonation analysis and modelling.

Paragraphs intonation is less emphatic. The F0 of the first peak in paragraphs is 16% lower than in isolated sentences for happiness, and 67% for surprise, and a 25% higher for sadness. The same tendencies can be observed for the slope of peaks. Pre-pause intonation figures are similar for both sentences and paragraphs.

4. PROSODY VS SEGMENTAL VOICE QUALITY

In a new copy-synthesis test, we tried to determine the influence of segmental and supra-segmental features in the emotion recognition rate.

Table 5 shows the evaluation results of an experiment with mixed-emotion copy-synthesis (diphones and prosody are copied from different emotional recordings).

	Neutral	Happy	Sad	Cold Anger	Surprise
1st syllable	135 (+-10.5)	175 (+-25.2)	113 (+-14.8)	130 (+-19)	217 (+-32.2)
Peaks slope	38.8 (+-11.6)	70.9 (+-28.7)	29.6 (+-10.3)	127 (+-19.1)	-56 (+-32.1)
1st valley	110 (+-9.6)	112 (+-16.2)	90 (+-10.2)	98 (+-11.1)	181 (+-60.5)
Valleys slope	9 (+-29)	24.9 (+-43.1)	11.2 (+-14.3)	-4.4 (+-17.6)	4.6 (+-81.7)
1st syllable	108 (+-5)	113 (+-14.2)	82 (+-21)	97 (+-6.7)	121 (+-9.9)
Last valley	114 (+-8.1)	104 (+-17)	78 (+-6.1)	103 (+-8.1)	167 (+-42)
Last peak	106 (+-7.3)	140 (+-38.6)	84 (+-10.4)	126 (+-21)	266 (+-26.4)
Last phoneme	68 (+-14.8)	73 (+-12.8)	68 (+-18.5)	85 (+-13.9)	121 (+-18.5)
Last valley	106 (+-11.5)	122 (+-15)	89 (+-9.8)	120 (+-19.4)	154 (+-24.8)
Last peak	124 (+-25.8)	192 (+-62)	113 (+-18.1)	166 (+-53.7)	146 (+-60.9)
Last phoneme	158 (+-14)	177 (+-49.5)	102 (+-13)	142 (+-25.8)	246 (+-23.6)

Table 4 Intonation data (in Hz) for statements and interrogatives sentences (last three rows)

Diphones	Prosody	Classification rate	Identified emotion
Neutral	Happy	52.4%	Neutral
Happy	Neutral	52.4%	Happy
Neutral	Sad	66.6%	Sad
Sad	Neutral	45.2%	Sad
Neutral	Angry	23.8%	Surprise
Angry	Neutral	23.8%	Angry
Neutral	Surprise	76.2%	Surprise
Surprise	Neutral	33.3%	Non ident.

Table 5 Summary of the mixed emotion experiment

As we can clearly see, cold anger is not prosodically marked, and happiness, although having a prosody that is significantly different from the neutral one, it presents more recognisable differences from a segmental point of view.

We can conclude that prosodic modelling of emotional speech is not enough to make it recognisable (it does not convey enough emotional information in the supra segmental level). Finally, we can classify anger and happiness as segmental emotions, while sadness and surprise are rather prosodic emotions (sadness has also an important segmental component).

5. CONCLUSIONS AND FUTURE WORK

Although prosodic modelling is not enough to convey emotional information, we can classify anger and happiness as segmental emotions and surprise and sadness as prosodic emotions.

A full diphone-concatenation system is currently being developed, using the modelling data presented in this paper, and unit-selection techniques that can take advantage of the whole database of speech voice segments. This data-driven approach is particularly useful for emotional synthesis, after we have showed that different voice qualities can be identified in several emotions from our database. We want to take advantage of the capability of this kind of synthesis to copy the quality of a voice from a database.

A preliminary test of concatenative synthesis using only automatic emotional prosody confirms the proposed classification hypothesis.

ACKNOWLEDGMENTS

This work has been funded by CICYT project TIC 95-0147. Special thanks go to Mª Ángeles Romero, Gerardo Martínez, Sira Palazuelos, Ascensión Gallardo, Ricardo Córdoba and all people in GTH, especially those who participated in the evaluation tests.

REFERENCES

1. Montero, J.M. Gutiérrez-Arriola, J. Palazuelos, S. Enríquez, E. Aguilera, S. Pardo, J.M., "Emotional Speech Synthesis: from Speech Database to TTS", ICSLP'98.
2. Karlsson, I. "Controlling voice quality of synthetic speech", ISCLP'94, 1439-1442, 1994
3. Rutledge, "Synthesising styled speech using the klatt synthesiser", ICASSP'95, 648-649, 1995
4. "TIDE TP 1174 Final Report", 1997
5. Murray I.R. and Arnott, J.L. "Implementation and testing of a system for producing emotion-by-rule in synthetic speech", Speech Communication 16, pp. 359-368, 1995.
6. Heuft, B. Portele, T. and Rauth, M. "Emotions in time domain synthesis", ICSLP'96, 1974-1977, 1996
7. Dellaert et al, "Recognising emotion in speech", ICSLP'96, 1970-1973, 1996
8. Bou-Ghazade et al, "Synthesis of stressed speech from isolated neutral speech using HMM-based models", ICSLP'96, 1860-1863, 1996
9. Pereira, C and Watson, C., "Some acoustics characteristics of emotion", ICSLP'98, 1998.

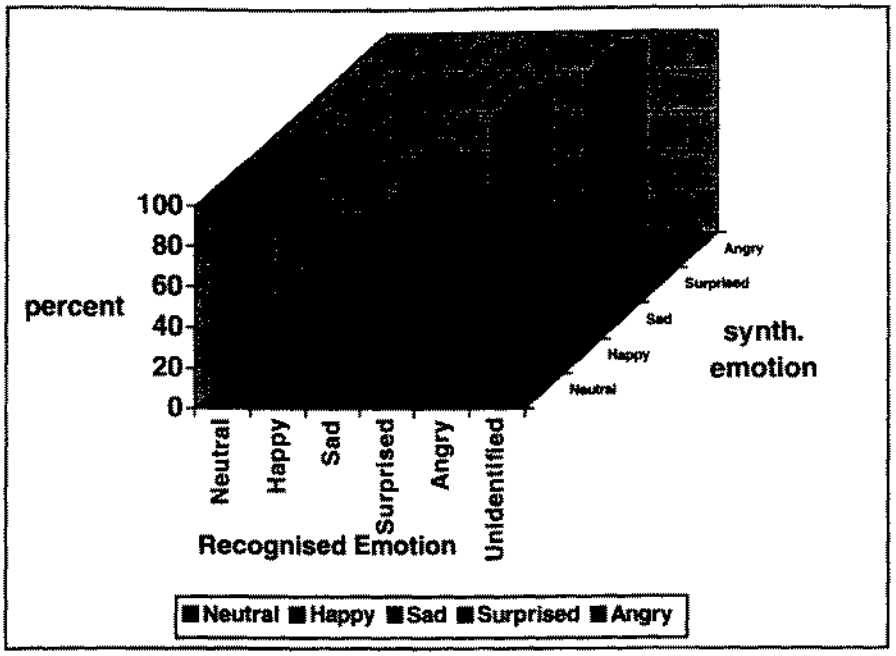

Figure 1 Confusion Matrix of the copy-synthesis experiment

Identified Emotion⇒		Neutral	Happy	Sad	Surprised	Angry	Non identif.
Segments	Prosody						
Neutral	Happy	***52,4 %***	19 %	11,9 %	4,8 %	0	11,9 %
Neutral	Sad	23,8 %	0	***66,6%***	0	2,4 %	7,1 %
Neutral	Surprised	2,4 %	16,7 %	2,4 %	***76,2%***	0	2,4 %
Neutral	Angry	11,9 %	19 %	19 %	***23,8 %***	7,1 %	19 %
Happy	Neutral	4,8 %	***52,4%***	0	9,5 %	26,2 %	7,1 %
Sad	Neutral	26,2 %	2,4 %	***45,2%***	4,8 %	0	21,4 %
Surprised	Neutral	19.0%	11.9%	21.4%	9.5%	4.8%	33.3%
Angry	Neutral	0	0	0	2.4%	95.2%	2,4 %

Table 6 Prosody vs. segmental quality test

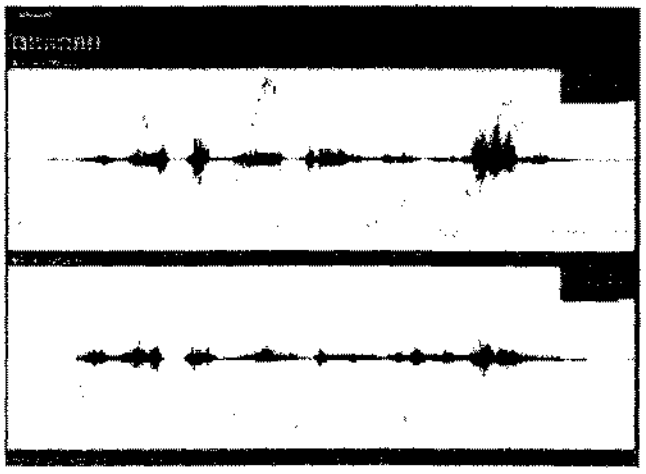

Figure 2 Two happy intonation contours for the same sentence

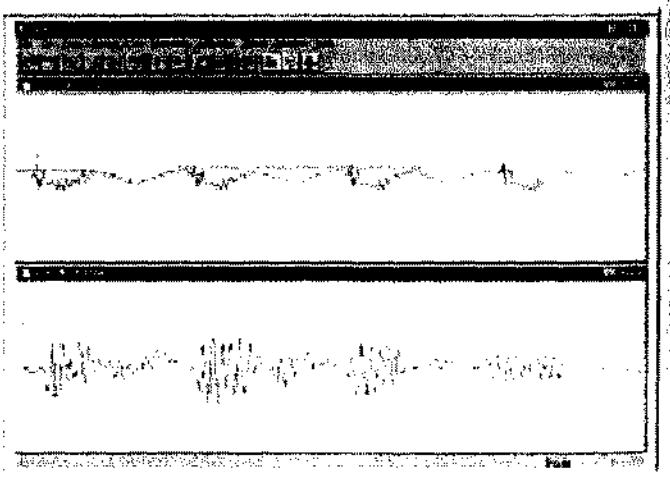

Figure 3 Two segments of neutral vs. angry voice