SPANISH EMOTIONAL SPEECH: TOWARDS CONCATENATIVE SYNTHESIS

J.M. Montero, J. Gutiérrez-Arriola*, R. Córdoba*, E. Enríquez** and J.M. Pardo**

*Grupo De Tecnología del Habla, ETSI Telecomunicación, Universidad Politécnica de Madrid
Ciudad Universitaria s/n, 28040 Madrid, Spain
Phone: +34 91 5495700 ext. 363; Fax: +34 91 3367323
Email: juancho@die.upm.es

**Grupo De Tecnología del Habla, Universidad Nacional de Educación a Distancia
Ciudad Universitaria s/n, 28040 Madrid, Spain

ABSTRACT

Currently, a key point in recognition and synthesis tasks is the addressing of the variability of human speech. One of the main sources of this diversity is the emotional state of the speaker. Speech under emotional conditions can be modelled as a deviation from neutral voice.

Most of the recent work in emotional synthesis has been focused on the prosodic aspects of this kind of speech. In a paper at ICSLP'98 [1], we present a thorough study of emotional speech in Spanish, and its application to TTS, including a prototype system that simulates emotional speech using a commercial synthesiser.

In this paper we describe the evolution of our work towards concatenative synthesis, beginning with copy-synthesis experiments (and their comparison to natural voice evaluation results) and automatic emotional prosody generation.

1. INTRODUCTION

The continuous increase in synthetic speech intelligibility has focused the attention of the research in the area of naturalness. Mimicking the diversity of natural voice is the aim of many current speech investigations. Emotional voice (sometimes under stress conditions) is analysed in many papers in the last few years [9][7][5][4].

The VAESS project TIDE TP 1174 (Voices Attitudes and Emotions in Synthetic Speech) developed a portable communication device for disabled persons using a multilingual synthesiser, specially designed to be capable not only of communicating the intended words, but also of portraying, by vocal means, the emotional state of the device user [8].

The GLOVE voice source that was used [2] allowed controlling Fant's model parameters. Although this improved source model can correctly characterise several voices and emotions (and the improvements are clear when synthesising a happy 'brilliant' voice), the 'menacing' cold angry voice had such a unique quality that we were unable to simulate it in the rule-based VAESS synthesiser (this fact led us to synthesise a hot angry voice, different from the database examples).

Accounting for this, a reasonable following step is to try to implement emotional speech through the use of a concatenative synthesiser [10], taking advantage of the capability of this kind of synthesis to copy the quality of a voice from a database (without an explicit mathematical model)

2. DATABASE

The Spanish Emotional Speech database (SES) contains two emotional speech recording sessions played by a professional male actor in an acoustically treated studio. We recorded 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 (in the VAESS project the secondary emotion was not used).

The recorded database was then phonetically labelled in a semiautomatic way. An automatic pitch epoch extraction software was used, but the outcome was manually revised using a graphical audio-editor programme, the same one that was used for the location and labelling of the phonemes.

3. COPY-SYNTHESIS EXPERIMENTS

Three copy-synthesis sentences were listened to by 21 people in a random-order forced-choice test (including a "non-identifiable" option) [6]. In copy-synthesis experiments, we used a concatenative

synthesiser with diphones and prosody from natural speech. The confusion matrix is shown in Fig. 1.

The copy-synthesis results, although significantly above random-selection level using a Student's test ($p>0.95$), are significantly below natural recording rates except for cold anger [1]. This decrease in the recognition score can be due to the evaluation of a new emotion in the copy-synthesis test, to the use of an automatic process for copying the prosody, and to the distortion introduced by prosody modification algorithms. It is remarkable that cold anger resynthesised sentences were evaluated significantly above natural recordings (the concatenation distortion made the voice even more menacing).

Table 1 shows the evaluation results of an experiment with mixed-emotion copy-synthesis (diphones and prosody are copied from different emotional recordings).

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 has 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 cold anger as a segmental emotion, surprise as a prosodic one, while sadness and happiness have important prosodic and segmental components (sadness has a predominant prosodic one; happiness is more easy to recognise by means of segmental characteristics).

Using the prosodic analysis described in [1], we created an automatic emotional prosodic module to verify the segmental vs. supra-segmental hypothesis. Combining this synthetic prosody (taken from paragraphs recordings) with optimal-coupling diphones (taken from the short sentences recordings), we carried out a new re-synthesis test. The results are shown in Table 2.

The differences between this final experiment and the first one are significant (using a chi-square test with 4 degrees of freedom and $p>0.95$) due to the bad recognition figure for surprise. In a one by one basis, and using a Student's test, anger, happiness, neutral and sadness results are not significantly different from the copy-synthesis test ($p<0.05$). An explanation for all these facts is that the prosody in this experiment was trained with the paragraphs prosody and it was never evaluated before for surprise (both paragraphs and sentences were evaluated in the VAESS project for sadness, happiness, anger and neutral style).

There is an important increase in happiness recognition rates when using both happy diphones and happy prosody, but the difference is not significant with a 0.95 threshold and a Student's distribution.

4. FUTURE WORK

The following step will be the development of a fully automatic emotional diphone concatenation synthesiser. As the range of the pitch variations is larger than for neutral-style speech, we shall use several units per diphone to cover this increased range.

As the segmental differences between emotions play an important role in their recognisability, we plan to apply voice-conversion techniques [3] to characterise the transformations that are necessary to produce emotional voice from neutral diphones, so we can obtain a multi-voice synthesiser without the need of new emotional recordings. These transformations can be applied to voice source and to vocal tract.

To analyse the voice source we shall use a polynomial-plus-noise model (the glottal source is modelled as a mixture of a polynomial function and a certain amount of additive noise). Given the characteristic of the emotional speech we plan to change only the glottal source and try to demonstrate that this is enough to convey the information needed to simulate natural emotions. We should compare these results with the complete transformation (glottal source plus vocal-tract) and we expect the results to be very similar. A very preliminary experiment transforming just one sentence from neutral to angry shows that we can be in the right hypothesis.

5. CONCLUSIONS

We have classified emotions as prosodic or segmental. We consider sadness and surprise as emotions that are mainly prosodic, while happiness and cold anger have to be considered as mainly segmental.

We have presented an emotional synthesiser where emotional information is transmitted through variations in the prosodic model and through an increase in the number of concatenation units (in order to be able to cover the variability introduced by emotions).

As we have shown, emotions can not be transmitted using only supra segmental information; so it is interesting to consider that emotional speech synthesis is a transformation of the neutral voice. We shall apply transformation techniques (parametric and non-parametric) in order to produce new emotional voices for a new speaker without having to record a new emotional database.

6. ACKNOWLEDGEMENTS

This work was supported 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, specially those who participated in the evaluation tests.

7. REFERENCES

[1] J.M. Montero, J. Gutiérrez-Arriola, S. Palazuelos, E. Enríquez, S. Aguilera, J.M. Pardo, "Emotional Speech Synthesis: from Speech Database to TTS", *Proc. of International Conference on Spoken Language Processing*, Sydney, Australia, 1998, Vol. 3 pp 923-926.

[2] I. Karlsson, "Controlling voice quality of synthetic speech", *Proc. of International Conference on Spoken Language Processing*, Yokohama, Japan, 1994, pp. 1439-1442.

[3] J. Gutierrez-Arriola et al, "Speech synthesis and prosody modification using segmentation and modelling of the excitation signal", *Proc. of European Conference on Speech Communication and Technology*, Rhodes, Greece, 1997, pp. 1059-1062.

[4] J. Rutledge, "Synthesising styled speech using the Klatt synthesiser", Proc. of International Conference on Acoustics, Speech and Signal Processing, Detroit, USA, 1995, pp. 648-649.

[5] I.R. Murray, J.L. Arnott. "Implementation and testing of a system for producing emotion-by-rule in synthetic speech", Speech Communication 16, pp. 359-368.

[6] B. Heuft, T. Portele, M. Rauth, "Emotions in time domain synthesis", *Proc. of International Conference on Spoken Language Processing*, Philadelphia, USA, 1996, pp. 1974-1977.

[7] F. Dellaert et al, "Recognising emotion in speech", *Proc. of International Conference on Spoken Language Processing*, Philadelphia, USA, 1996, pp. 1970-1973.

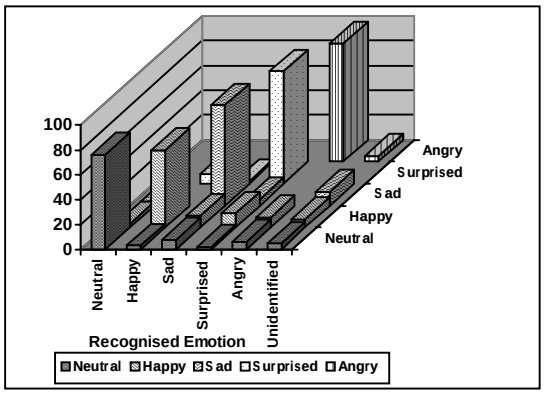
[8] V. Darsinos et al, "Designing a Speaker Adaptable Formant-based TTS System", *Proc. of European Conference on Speech Communication and Technology*, Rhodes, Greece, 1997, pp. 2527-2530.

[9] S. Bou-Ghazade et al, "Synthesis of stressed speech from isolated neutral speech using HMM-based models", *Proc. of International Conference on Spoken Language Processing*, Philadelphia, USA, 1996, pp. 1860-1863

[10] E. Rank, H. Pirker, "Generating Emotional Speech with a Concatenative Synthesiser", *Proc. of International Conference on Spoken Language Processing*, Sydney, Australia, 1998, Vol. 3 pp. 671-674.

Identified Emotion⇒		**Neutral**	**Happy**	**Sad**	**Surprised**	**Angry**	**Un-identified.**
Diphones	**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 1 Prosody vs segmental quality test



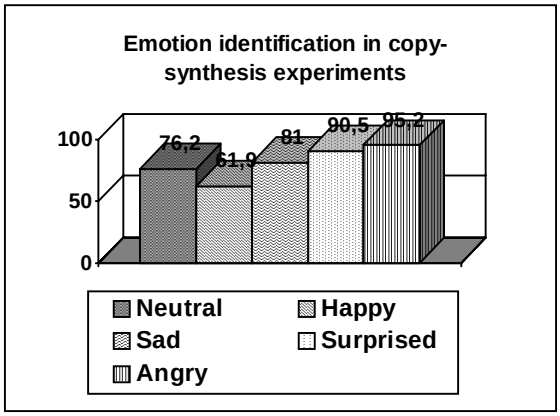
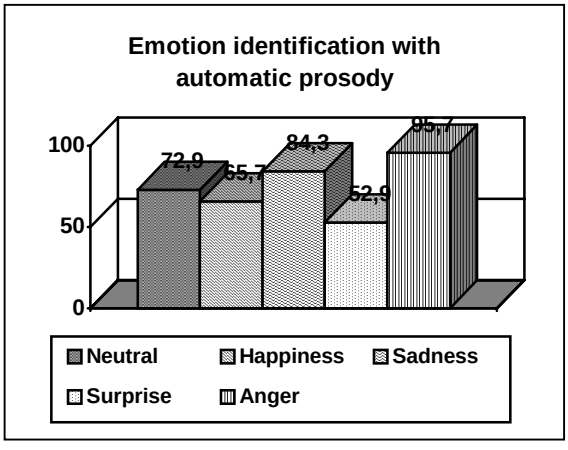


Figure 1 Evaluation of copy-synthesis experiments

Identified Synth.	Neutral	Happy	Sad	Surprised	Angry	Unidentified
Neutral	**72,9 %**	0	15,7 %	0	0	11,4 %
Happy	12,9 %	**65,7 %**	4,3 %	7,1 %	1,4 %	8,6%
Sad	8,6 %	0	**84,3 %**	0	0	17,1 %
Surprised	1,4 %	27,1 %	1,4 %	**52,9 %**	0	17,1 %
Angry	0	0	0	1,4 %	**95,7%**	2,9 %

Table 2 Automatic prosody experiments



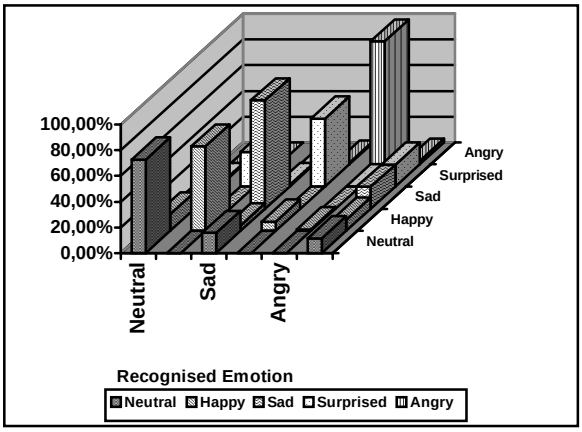


Figure 2 Evaluation of the automatic prosody experiment