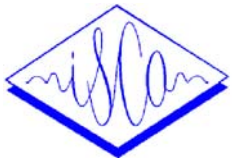

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

DEVELOPMENT OF AN EMOTIONAL SPEECH SYNTHESISER IN SPANISH

J.M. Montero, J. Gutiérrez-Arriola*, J. Colás*, J. Macías*, E. Enríquez**, J.M. Pardo*.*
*Grupo de Tecnología del Habla. Departamento de Ingeniería Electrónica. Universidad Politécnica de Madrid
E.T.S.I. Telecomunicación. 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
juancho@die.upm.es
http://www-gth.die.upm.es

ABSTRACT

Currently, an essential point in speech synthesis is the addressing of the variability of human speech. One of the main sources of this diversity is the emotional state of the speaker. Most of the recent work in this area has been focused on the prosodic aspects of speech and on rule-based formant-synthesis experiments. Even when adopting an improved voice source, we cannot achieve a smiling happy voice or the menacing quality of cold anger. For this reason, we have performed two experiments aimed at developing a concatenative emotional synthesiser, a synthesiser that can copy the quality of an emotional voice without an explicit mathematical model.

Keywords: emotions, emotional speech, concatenative-synthesis, copy-synthesis

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 [2][9][10].

1.1 Formant synthesis

The first systems and experiments in emotional speech synthesis were based on formant synthesisers [5] [7].

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 [1].

The GLOVE voice source that was used [4] 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 lead us to synthesise a hot angry voice, different from the database examples) [1].

1.2 Concatenative synthesis

Only a few papers in the literature describe an emotional synthesiser based on concatenative techniques [6] [11]. Even in these cases, the speech units were not extracted from an emotional but a neutral database (they are just using automatic prosody for modelling emotional speech). As it has been shown in [2], prosodic elements such as pitch, tempo and stress are not the only acoustic correlates of emotion in human speech and, for some emotions, they are not the most relevant ones.

Accounting for all of this, the following step is to try to implement emotional speech through the use of a concatenative synthesiser, 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) [12].

2. THE SES DATABASE: SPANISH EMOTIONAL SPEECH

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 revision and phoneme labelling were performed manually, using a graphical audio-editor programme.

3. THE CONCATENATIVE-SYNTHESIS EXPERIMENTS

3.1 The copy-synthesis experiment

In our first experiment, three copy-synthesis sentences were listened by 21 people in a random-order forced-choice test (including a "non-identifiable" option). We used a concatenative synthesiser [3] with diphones and stylised prosody taken from the database sentences. The results are shown in Table 1.

Although the figures in the diagonal of the confusion matrix are below natural voice tests [1], they are not comparable due to prosody stylisation, prosody modification algorithm and the addition of a new emotion: surprise. Nevertheless, the results are satisfactory:

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

Table 1 Recognition rates for the copy-synthesis experiment

3.2 The automatic-synthesis experiment

Using the prosodic analysis described in [2], we created an automatic emotional prosodic module to verify the segmental vs. supra-segmental hypothesis. In the second experiment, by combining this synthetic prosody (taken from paragraphs) with optimal-coupling diphones (taken from the short sentences), we carried out a new 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 ($p>0.95$).

Emotion	*Recognition rate*
Neutral	72.9%
Happy	65.7%
Surprised	52.9%
Sad	84.3%
Angry	95.7%

Table 2 Recognition rates for the automatic-prosody experiment

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). This new emotion needs an assessment test for the paragraph recordings (the prosody is less emphatic: the pre-pause lengthening and F0 raise are shorter than in isolated sentences), and it also needs further improvements in the prosodic modelling (we have to add new parameters regarding the pre-pause lengthening effect, significantly different for this emotion).

4. UNIT-SELECTION SYNTHESIS

The recognition rates when mixing neutral diphones and emotional prosody are not satisfactory [1]. The F0, duration and energy ranges vary greatly from one emotion to another. In addition to this, voice quality is clearly recognisable for some emotions.

All these facts led us to develop a unit-selection synthesiser. The first parameters to take into account are the prosodic ones: best units are selected using the distance between the target synthetic unit and source database candidates in terms of pitch and duration:

- longer source units are preferable (because of the reverberation noise when period-replication lengthening technique is applied). If the system needs to increase source unit duration in more than a 25 per cent, the target duration is cut to this limit (inserting small changes in the micro-prosody component).

- For the intonation curve, units with the same rising or decreasing tendency and similar F0 mean are selected. F0 distortions of more than a 40 per cent are penalised.

In order to suppress spectral discontinuities (that can be perceived as plosive or pop sounds), we implemented a Viterbi-based search algorithm, accounting not only for the prosodic distance but also for the concatenation cost (measured as Mahalanobis cepstral distance between consecutive units).

Our first informal tests suggest that:

- sad and neutral sentences quality is good: they are the easiest sentences to recognise due to their shorter F0 and duration range,
- happy units exhibit important differences in voice quality: a mixture of units from smiling speech and units from non-smiling speech decreases the happy quality of the synthesised voice,
- surprise: its prosody is the less homogeneous one because, in pre-pause positions, there is a great increase in both F0 and duration (only the units coming from this positions are good for synthesising pre-pause units).

5. CONCLUSIONS

A fully automatic emotional diphone-concatenation system is currently being developed and the preliminary test (using only automatic emotional prosody) exhibits encouraging results.

In order to cope with the increase in prosody-modifications range when synthesising emotional speech (both in pitch, duration and energy) and to make use of the whole database, unit-selection techniques, applied to emotional synthesis, are now being tested.

As the database is not rich enough in terms of variety of units (it was designed for the analysis of emotional prosody), the corpus of available units should be increased with new emotional recordings.

6. ACKNOWLEDGEMENTS

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.

7. REFERENCES

[1] Montero, J.M. Gutiérrez-Arriola, J. Palazuelos, S. Enríquez, E. Aguilera, S. and Pardo, J.M. (1998), "Emotional Speech Synthesis: from Speech Database to TTS", in *Proceedings of the International Conference in Speech and Language Processing ICSLP'98*, vol. III pp. 923-926.

[2] Montero, J.M. Gutiérrez-Arriola, J. Colás, J. Enríquez, S. and Pardo, J.M. (1999), "Analysis and Modelling of Emotional Speech in Spanish", in *Proceedings of the XIII International Conference of Phonetic Sciences* (to be published).

[3] Pardo J.M. et al (1995), "Spanish text to speech: from prosody to acoustic", in *Proceedings of the International Conference on Acoustics*, vol. III pp. 133-136.

[4] Karlsson, I. (1994) “Controlling voice quality of synthetic speech”, in *Proceedings of ISCLP’94*, pp. 1439-1442.

[5] Murray I.R. and Arnott, J.L. (1995) “Implementation and testing of a system for producing emotion-by-rule in synthetic speech”, in *Speech Communication 16*, pp. 359-368.

[6] Heuft, B. Portele, T. and Rauth, M. (1996) “Emotions in time domain synthesis”, in *Proceedings of ISCLP’96*, pp. 1974-1977.

[7] Rutledge (1995), “Synthesising styled speech using the klatt synthesiser”, in *Proceedings of the International Conference in Speech and Signal Processing ICASSP'95*, pp. 648-649.

[8] Higuchi, N., Hirai, T. and Sagisaka (1997), Y. "effect of Speaking Style on Parameters of fundamental Frequency Contour", in “Progress in speech synthesis”, pp.417-427.

[9] Scherer K.R. (1996) "Adding the affective dimension: a new look in speech analysis and synthesis" in *Proceedings of ICSLP'96*.

[10] Amir N. and Ron S. (1998) "Towards an automatic classification of emotions in speech" in *Proceedings of ICSLP'98*.

[11] Rank E. and Pirker H. (1998) "Generating emotional speech with a concatenative synthesiser" in *Proceedings of ICSLP'98*, vol. III pp. 671-674.

[12] Iida A. Campbell N. Iga S. Higuchi F. and Yasumura M. (1998) "Acoustic nature and perceptual testing of corpora of emotional speech" in *Proceedings of ICSLP'98*, vol. IV pp. 1559-1562.

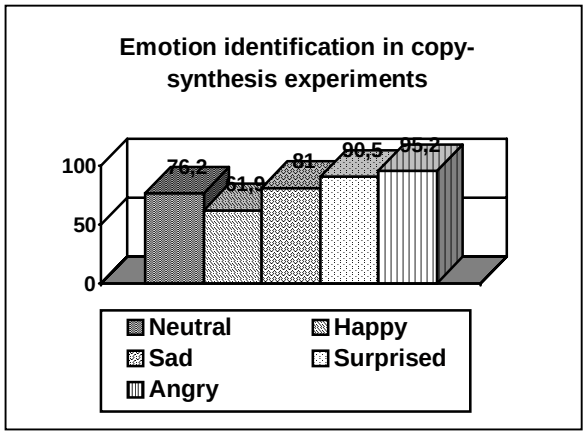

Figure 1 First experiment recognition rate

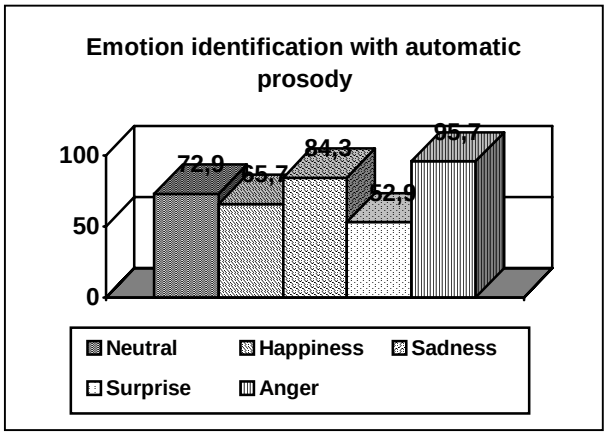

Figure 2 Second experiment recognition rate

Identified Vs. Synthesized.	Neutral	Happy	Sad	Surprised	Angry	Unidentified
Neutral	76,2 %	3,2%	7,9 %	1,6 %	6,3 %	4,8 %
Happy	3,2 %	61,9 %	9,5 %	11,1 %	7,9 %	6,3 %
Sad	3,2 %	0	81 %	4,8 %	0	11,1 %
Surprised	0	7,9 %	1,6 %	90,5 %	0	0
Angry	0	0	0	0	95,2 %	4,8 %

Table 3 Copy-synthesis confussion matrix .

Identified Vs. Synthesized	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 4 Automatic prosody confussion matrix .