EMOTIONAL SPEECH SYNTHESIS: FROM SPEECH DATABASE TO TTS

J.M. Montero, J. Gutiérrez-Arriola*, S. Palazuelos**, E. Enríquez***, S. Aguilera**, J.M. Pardo**

*Grupo de Tecnología del Habla-Departamento de Ingeniería Electrónica-E.T.S.I. Telecomunicación-Universidad Politécnica de Madrid-Ciudad Universitaria s/n, 28040 Madrid, Spain

**Laboratorio de Tecnologías de Rehabilitación-Departamento de Ingeniería Electrónica-E.T.S.I. Telecomunicación-Universidad 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

Modern Speech synthesisers have achieved a high degree of intelligibility, but can not be regarded as natural-sounding devices. In order to decrease the monotony of synthetic speech, the implementation of emotional effects is now being progressively considered.

This paper presents a through study of emotional speech in Spanish, and its application to TTS, presenting a prototype system that simulates emotional speech using a commercial synthesiser.

The design and recording of a Spanish database will be described and also the analysis of the emotional prosody (by fitting the data to a formal model). Using this collected data, a rule-based simulation of three primary emotions was implemented in the Text-to-Speech system. Finally, the assessment of the synthetic voice through perception experiments will classify the system as capable of producing quality voice with recognisable emotional effects.

1. INTRODUCTION

The quality of synthetic speech has been greatly improved by the continuous research of the speech scientists. Nevertheless, most of these improvements were aimed at simulating natural speech as that uttered by a professional announcer reading a neutral text in a neutral speaking style. Because of mimicking this style, the synthetic voice results to be rather monotonous, suitable for some man-machine applications, but not for a vocal prosthesis device such as the communicators used by disabled people [7].

In order to increase the acceptability and improve the performance of synthesised speech in this area of application, the synthesis of attitudes and emotions have to be considered as critical: to departure from a characterless reference point (a well-controlled neutral state) will contribute to add variability and, therefore, naturalness to the voice of the prosthesis.

Past research in this topic area has been mainly focused on speech databases for several languages, speech analysis (either in a macroscopic way or in phoneme by phoneme basis) and prosody modelling [2], [8], [9]. A complete review of the literature on qualitative analysis of emotional speech can be found in [10].

The most complete studies and implementations, include a commercial TTS system in their developments and evaluations [6], but a detailed control of prosodic and, specially, of voice source parameters is not available. The low level control of some fundamental parameters within the synthesiser's internal structure is not allowed by these systems, and this control parameters are necessary for further improvement on the quality of emotional synthetic speech .

The VAESS project TIDE TP 1174 (Voices Attitudes and Emotions in Synthetic Speech) was aimed at developing a portable communication device for disabled persons containing a multilingual INFOVOX 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.

To get this goal, the new GLOVE voice source was used [1], that allows to control Fant's model parameters. This improved source model can correctly characterise each emotion and is able to mimic the vocal “brilliance” of a truly human voice.

2. SPANISH EMOTIONAL SPEECH DATABASE (SES)

As no study or database was available for Spanish,. our first stage was, then, to design, record and label the SES database.

2.1. Description

SES contains two emotional speech recording sessions played by a professional actor in an acoustically treated studio using a table-top high quality microphone, an OROS audio card, a SONY DAT recorder and EUROPEC software, at a 16 kHz sampling rate. Three basic or primary emotions (sadness, happiness and anger) in addition to a neutral speaking style were recorded (a secondary emotion, surprise, was also recorded, but it has not been used yet, as it was not part of the VAESS communicator requirements). The 38 year old male actor that was recorded for the corpus has a Standard Castilian accent and has been a professional actor for more than ten years.

During the recordings, the actor was free to choose the way that he considered the most appropriate for generating the required emotion.

A session comprises 3 passages (4 or 5 long sentences per passage), 15 short sentences and 30 isolated words. All these texts were emotionally neutral, not conveying any emotional charge through lexical, syntactical or semantical means.

An example of the recorded data is passage 1:

"The participants in the conference went to El Escorial afterwards. They arrived there in a bus, where a guide was explaining the more outstanding monuments of the trip. The visit to the monastery was commented by the same guide, that should know a lot about El Greco, whose picture "El martirio de San Mauricio" he extensively commented; his knowledge about the rest of the picture in the museum should not be the same, as he passed through them in a rapid way, giving place to smiles".

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.

This labelling was further revised in an automatic resynthesis-and-listening process using a diphone concatenation system developed in our laboratory [3]. New corrections were applied in order to get a final refinement of the labelling.

2.2. Evaluation of the natural voice

The assessment of the natural voice is aimed at judging the appropriateness of the recordings as a model for readily recognisable emotional synthesised speech.

The voice quality was tested by non-handicapped listeners, since their perception would not be any different. Fifteen normal listeners, both men and women of different ages (e.g. between 20 and 50) were selected from several social environments; none of them was used to synthetic speech [5].

The stimuli contained 5 emotionally neutral sentences (not conveying any intrinsic emotional content). 3 sentences came from the short sentences set and the other 2 were part of the passages. As 3 emotions and a neutral voice had to be evaluated, 20 different recordings per listener and session were used (only one session per subject was allowed).

In each session the audio recordings of the stimuli were presented to the listener in a random way. Each piece of text was played up to 3 times.

As can be observed in Table 5, the subjects had no difficulty in identifying the emotion that was simulated by the professional actor and the diagonal figures are clearly above the chance level (20 %). A Chi-square test rejects (with $p<0.05$) the random-selection null hypothesis.

Analysing the results in a sentence by sentence basis, none of them was significantly worse recognised (the identification rate varied from 83.3% to 93.3%)

2.3 Analysis and modelling

Phoneme durations were analysed by means of a multiplicative model with up to 123 parameters including intrinsic durations of vowels and consonants, phonetic context influence or pre-pause lengthening. A summary can be found in Table 1.

Phonemes	Happy/ Neutral	Sad/ Neutral	Angry/ Neutral
Vowels	1.06	1.02	1.10
Consonants	1.11	1.53	1.49
Prepause length.	0.99	0.86	1.07
Contextual coefs.	0.98	1.08	0.94

Table 1: Duration-model parameters.

For the intonation analysis we used a simple model that divides each breadth group in three areas separated by the first and last stressed vowels [4]. For happiness, the recordings contain at least 3 types of different contours; for a coherent training of the model, only the initial-focus ones were selected and analysed. Fifteen parameters for inquiring and non-inquiring sentences were computed by RMS error minimisation. The most outstanding results are:

F0 Features	Happy/ Neutral	Sad/ Neutral	Angry/ Neutral
1st F0 peak	1.29	0.83	0.96
F0 slope	1.82	0.76	-0.05
Last F0 peak	1.32	0.79	1.19
Last phoneme F0.	1.07	1.78	1.25

Table 2: F0-model parameters.

Pauses were only present at the passages recordings. Their mean duration values measured in seconds were:

Pause type	Neutral	Sad	Happy
Before stop	0.91	1.17	0.42
Intra-sentence	0.51	0.69	0.31

Table 3: Duration of pauses.

Significant differences between passages and short sentences prosody were observed (paragraphs prosody was less emphatic)

3. FORMANT SYNTHESIS TTS

We implemented the emotional voices in the rule-based GLOVE synthesiser. Only a small set of new emotional rules had to be written, as all the emotions were implemented through the adequate setting of general parameters in a fine-tuning process carried out by an expert phonetician.

Only one of the F0 contours used by the actor to simulate happiness was implemented (the negative-slope one). as synthesising high F0 levels resulted in a rather childish voice, the phonetician implemented a more emphatic voice.

The menacing cold anger of the recordings was the most difficult emotion to implement (in spite of increasing the additive noise and the spectral tilt). As the results were disappointing, a hot anger voice was fine-tuned.

3.1. Customizable Parameters of emotional synthesis

Speaking rate: ranging from 150 for a sad voice to 179 for an angry one; the default value for a neutral voice is 160.

F0 Range (differences between the peaks and throughs in F0 contour): sad voice had the smallest range and angry voice had the highest one (hot anger).

Mean Pitch level: a happy voice had the highest mean F0; a sad voice had the lowest one.

F0 Slope: angry and happy voices shared a high descendant slope; sad voice was rather flat.

Spectral Tilt: a lower tilt value increases the high frequency contents of the voice source, producing clearer voices; it is a specially useful parameter for happy voices.

Additive Noise: pitch-synchronous noise added to the voice source; used for sadness and anger.

Type of emotion: special rules are applied for happy and neutral voice, changing the default intonation contour characteristics.

3.2. Evaluation of the synthetic voice

In Table 6, the identification percentages are between 42,6% for angry sentences and 82,6% for sad sentences. However, there is an evolution in the results leading to the conclusion that a bigger number of sentences would improve the results.

As we can observe in table 7, the results for the last 10 sentences are much better than the global ones. At the beginning of the test, the subjects are not used to synthetic voice and none of the emotions (but sad) was correctly identified. Nevertheless, he/she adapts quickly to the synthesised emotions, increasing the identification percentages by the end of the test, from a 31-40% (excluding sad) to 60-86%. Subjects commented that, after several sentences, they could hear differences among the different emotions, and the identification was much easier in spite of receiving no feedback from the evaluation supervisor.

However, there was still confusion between emotions (for example, angry sentences were identified as happy more than 25% times due to a similar F0 contour), but happy sentences were nearly perfectly identified by the end of the test (the mistakes were identifying the happy sentences as euphoric (happier) ones).

In some of the sentences, they suggested that the recognised emotion was other one, like stressed, euphoria, affirmation.

Sad sentences were easier to identify, even in the first part of the test (86,6%).

4. PROSODY VS SEGMENTAL VOICE QUALITY

In a new re-synthesis test we tried to determine the influence of segmental and supra-segmental features in the emotion recognition rate [6].

We mixed diphones and prosody from two different emotional sentences (one of them was always a neutral recording). For instance, 3 sentences were resynthesised using the diphones from a neutral recording, but the prosody from a sad one.

Diphones	Prosody	Classified as	Rate (%)
Neutral	Happy	Happy	19
Happy	Neutral	Happy	52.4
Neutral	Sad	Sad	66.6
Sad	Neutral	Sad	45.2
Neutral	Angry	Angry	7.1
Angry	Neural	Angry	95.2

Table 4: Prosody vs. voice quality test

The results in Table 4 show that anger was mainly simulated by the actor through voice quality and glottal source changes.

5. CONCLUSIONS

A whole emotional TTS system has been developed, showing good recognition rates after a short adaptation period.

Prosodic modelling is not enough to convey emotional information, and preliminary results show that a data-driven diphone-concatenation system gets promising results

6. ACKNOWLEDGEMENTS

This work has been funded by EC TIDE project TP-1174 VAESS and by CICYT project TIC 95-0147. Special thanks go to the actor, Eduardo Jover, to Johan Bertenstam and Kjell Gustafsson from KTH, to Mª Angeles Romero, Ascensión Gallardo, Gerardo Martínez Salas and all people in GTH, specially those that participated in the evaluation tests.

7. REFERENCES

1. Carlsson, R., Granstrom, B.,. and Nord, L. "Experiments with emotive speech, acted utterances

and synthesised replicas", ICSLP 92, pp. 671-674, 1992

2. Engberg, I.S., Hansen, A.V., Andersen, O. and Dalsgard, P. "Design, recording and verification of a Danish Emotional Database", Eurospeech 97, pp. 1695-1698, 1997

3. Pardo, J.M. et al, "Spanish text to speech: from prosody to acoustic", International Conference on Acoustics 95, vol. III pp. 133-136, 1995

4. Moreno, P.J. et al "Improving naturalness in a text to speech system with a new fundamental frequency algorithm", Eurospeech 89, pp. 360-363, 1989

5. Palazuelos, S., Aguilera, S., Montero, J.M. and Pardo, J.M. "Report on the evaluation of the emotional voice for Spanish", EC TIDE TP-1174 VAESS project Report, pp 1-19, 1997

6. Murray, I.R. and Arnott, J.L. "Synthesising emotions in speech: is it time to get excited?", ICSLP 96, pp. 1816-1819, 1996

7. Higuchi, N., Hirai, T. and Sagisaka, Y. "effect of Speaking Style on Parameters of fundamental Frequency Contour", en "Recent advances in speech synthesis", pp.-, 1997

8. Noad, J.E.H., Whiteside, S.P. and Green, P.D. "A macroscopic analysis of an emotional corpus", Eurospeech'97, pp. 517-520, 1997

9. Vroomen, J., Collier, R. and Mozziconacci, S. "Duration and intonation in emotional speech", Eurospeech 93, pp. 577-580, 1993

10. Murray, I.R. and Arnott, J.L. "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion", JASA Vol. 93(2), pp. 1097-1108, 1996

Emotion Identified / Simulated	Neutral	Happy	Sad	Angry	No Identif.
Neutral	**67 (89,3%)**	1 (1,33%)	1 (1,33%)	3 (3,99&)	3 (3,99%)
Happy	13 (17,3%)	**56 (74,6%)**	1 (1,33%)	1 (1,33%)	4 (5,33%)
Sad	1 (1,33%)	0	**70 (90,3%)**	1 (1,33%)	3 (3,99%)
Angry	0	1 (1,33%)	2 (2,66%)	**67 (89,3%)**	5 (6,66%)

Table 5: Confusion matrices for natural voice evaluation test

Emotion Identified / Synthesised	Neutral	Happy	Sad	Angry	No Identif.	Other
Neutral	**44 (58,6%)**	0	22 (29,3%)	8 (10,6%)	1 (1,3%)	0
Happy	18 (24%)	**35 (46,6%)**	7 (9,3%)	2 (2,6%)	10 (13,3%)	3 (4%)
Sad	7 (9,3%)	0	**62 (82,6%)**	3 (3,9%)	1 (1,3%)	2 (2,6%)
Angry	16 (21,3%)	16 (21,3%)	1 (1,3%)	**32 (42,6%)**	4 (5,3%)	6 (8%)

Table 6: Confusion matrices for synthetic voice evaluation test

Emotion Identified / Synthesised	Neutral	Happy	Sad	Angry	No Identif.	Other
Neutral	**32 (61,5%)**	0	4 (7,7%)	7 (13,5%)	1 (1,9%)	8 (15,3%)
Happy	0	**13 (86,6%)**	0	0	0	(13,3%)
Sad	7 (16,6%)	0	**33 (78,5%)**	2 (4,7%)	0	0
Angry	0	8 (26,6%)	0	**18 (60%)**	1 (3,3%)	(10%)

Table 7: Confusion matrices for synthetic voice evaluation test (last 10 sentences)