# Spanish Expressive Voices: corpus for emotion research in Spanish

**R. Barra-Chicote[1], J.M. Montero[1], J. Macias-Guarasa[2], S.L. Lufti[1], J.M. Lucas[1], F. Fernandez-Martinez[1], L.F. Dharo[1], R. San-Segundo, J. Ferreiros[1], R. Cordoba[1], M. Pardo[1]**

Universidad Politecnica de Madrid[1] Universidad Alcalá de Henares[2]

{barra, juancho, syaheerah, juanmak, efhes, lfdharo, lapiz, jfl, cordoba, pardo}@die.upm.es[1], macias@depeca.uah.es[2]

**Abstract**

A new emotional multimedia database has been recorded and aligned. The database comprises speech and video recordings of one actor and one actress simulating a neutral state and the Big Six emotions: happiness, sadness, anger, surprise, fear and disgust. Due to a careful design and its size (more than 100 minutes per emotion), the recorded database allows comprehensive studies on emotional speech synthesis, prosodic modelling, speech conversion, far-field speech recognition and speech and video-based emotion identification. The database has been automatically labelled for prosodic purposes (5% was manually revised). The whole database has been validated thorough objective and perceptual tests, achieving a validation score as high as 89%.

## 1. Introduction

Several multimedia corpora have recently been developed involving speech studies. However, some of them are limited to either the analysis of emotion expression, or detection.

Most of these corpora focus on meeting settings (Chen et al,2005) (Mana et al, 2007) and even then these studies are mostly focusing on multispeaker solutions, lip-reading correlating to human understanding, personality traits and social behaviours while discussing and interacting - and not particularly concentrating on the synthesis or detection on emotions per se. In addition, one of the main problems in speech recognition tasks is how to adapt classifiers to affective speech.

Several studies (Mana et al, 2007)(Castellano et al, 2007)(Sebe et al, 2005) has highlighted that an ideal system for automatic recognition of human affective information should be multimedia. This integration is also exhibited in psychology studies such as (Schrerer et al, 2007). However, the studies mentioned above introduce multimedia approaches that are limited to automatic affect sensing and not for the use of affective expressions such that those in text-to-speech systems.

The paper is organised as follows: previous considerations and description of target information is explained. Next, whole acquisition equipment and set-up is described. Finally, the evaluation of the corpus is presented and its labelling method explained.

## 2. Corpora Design

### 2.1. Previous Considerations

What makes SEV unique is that it is a combination of four Spanish previous corpora. Each of them designed to meet a particular goal and equipped with various underlying objectives of speech studies which are not just limited to detection of emotion, but also to acceptable synthesize of emotions (for expressions).

The acquisition of enough far-field emotional speech would give the possibility of the adaptation of beam-forming techniques to emotional speech that could perform the speech applications growing in human-machine interfaces where no close-talk speech signal is available.

The addition of video information could provide features that help to provide more specific information such as detection of level of intensity of a particular emotion or even make up for information lost due to corrupting influences in the audio content.

Several considerations like what kind of emotions would be recorded, what emotions should be recorded or who will be the speakers should be taken into account.

First discussion is the pros and cons of acted versus real emotions. As pointed in (Burkhardt et al., 2005), so-called full-blown emotions very rarely appear in real world and ethic problems appear when recording people experiences of real emotions. These things make almost impossible to work with real data and do it in a clean and high quality acquisition set-up.

The multimedia character of SEV makes more difficult to approach to real situations without losing quality research aspects. This fact makes a difficult task the expression of emotions simultaneously through speech and mimic (mainly facial features). This difficulty and one of the main purposes of the corpus, high quality emotional speech synthesis, are the main reasons why two professional actors (one male and one female) were selected.

It seems logical to use distinct terms when acted emotions are investigated. Due to this our SEV corpus focuses on discrete emotion instead of emotional states projected on "emotional dimensions" (PAD model (Schoder, 2004)). Looking for comparison between new

studies and previous works in our Group (Montero et al., 2002)(Barra-Chicote et al., 2006)(Barra-Chicote et al., 2007) we selected original emotions in SES corpus (Montero et al., 1998) (happiness, cold anger, surprise, sadness and neutral reference) and three new ones (fear, disgust and hot anger) in order to complete the group of basic emotions.

It has been pursued that SEV corpus would be well suited for use to analyse data in developing or testing automatic recognition systems or systems involving emotion synthesis and hopefully there is performance increasing in these two tasks.

### 2.2. Speech Content

The main purpose of 'near' talk speech recordings is emotional speech synthesis and recognition and tasks related to emotion identification. Three channels were recorded: a close talk headset-microphone, a lapel microphone and a desktop microphone.

Several length features of the corpora, as average length of words (< w >) or allophones (< A >), are presented in Table 1.

- **Emotional Level Corpus**

Fifteen reference sentences of SESII-A corpus were played by actors 4 times, incrementing gradually the emotional level (neutral, low, medium and high level).

- **Diphone concatenation synthesis corpus**

LOGATOMOS corpus is made of 570 logatomos within the main Spanish Di-phone distribution is covered. They were grouped into 114 utterances in order to provide the performance of the actors. Pauses between words were requested to them in the performance in order to be recorded as in an isolated way.

This corpus allows studying the impact and the viability of communicate affective content through voice by no semantic sense words. New voices for limited domain expressive synthesisers based on concatenative synthesis would be built.

- **Unit Selection synthesis corpus**

QUIJOTE is a corpus made of 100 utterances selected from the 1st part of the book Don Quijote de la Mancha and that respects the allophonic distribution of the book. This wide range of allophonic units allows synthesis by unit selection technique.

- **Prosody Modelling In SESII-B corpus**

Hot anger was additionally considered in order to evaluated different kinds of anger.

The 4 original paragraphs in SES (Montero et al., 1998) has been split into 84 sentences. PROSODIA corpus is made of 376 utterances divided into 5 sets. The main purpose of this corpus is to include rich prosody aspects that makes possible the study of prosody in speeches, interviews, short dialogues or question-answering situations.

### 2.3. Far Field Content

The main purpose of 'far' talk speech recordings is to evaluate the impact of affective speech capture in more realistic conditions (with microphones placed far away from the speakers), also in tasks related to speech recognition and emotion identification.

Two microphone arrays were used for recording: a linear harmonically spaced array composed of 12 microphones placed on the left wall, and a roughly squared microphone array composed of four microphones placed on two tables in front of the speaker.

Although the acoustic environment is controlled and reverberation is low, experiments on this data can lead to interesting results on emotional speech processing.

### 2.4. Video Content

Video information was also recorded for every utterance. The main purpose of this capture is allowing research on emotion detection using visual information, face tracking studies and the possibility of study specific head, body or arms behaviour that could be related to features such as intensity level in the recorded speech signals or give relevant information of each emotion played. Also, audio-visual sensor fusion for emotion identification and even affective speech recognition are devised as potential applications of this corpus.

Figure 1 shows some zoom examples of various emotions and Figure 4 presents the camcorder situation in the chamber.

Figure 1: Example frames of emotions

## 3. Recording Equipment and Setup

This section describes the equipment used to record the SEV database. The recording was controlled by an operator in a room next to the recording acoustic treated chamber. The operator controlled the recording

application and was able to talk to the speaker at any time. Recording equipment was able to synchronously record 20 audio channels and a video signal. Figure 2 shows the general architecture of the recording set-up.

### 3.1. Audio Recording Hardware

Every channel's audio is sampled at 48 kHz and samples are 24 bits long. Twenty channels consist of:

- Channel 1: A close-talk head-mounted Shure Microphone.
- Channel 2: An electroglotograph signal is recorded in order to get high-quality pitch marks. The actors were a necklace with two electrodes as shown in Figure 4.
- Channel 3: A desktop microphone.
- Channel 4: A lapel Shure Microphone.
- Channels 5-8: A quasi-squared array composed of 4 PZM microphones.
- Channels 9-20: A harmonically-spaced linear array composed of 11 microphones, plus an additional microphone (for z-axis position discrimination) located 30 cm above the linear array axis.

All audio sources are connected to 3 RME Octamic-D units, in charge of microphone pre-amplification and A/D conversion.

The channels devoted to the Sennheiser mics are phantom powered (using the provided phantom power from the microphone pre-amplifiers). The built-in high pass filter (LO CUT option) of the microphone pre-amplifiers is activated (80Hz, 18 dB/octave).

The 3 ADAT digital streams from the Octamics are optically linked to a RME HDSP 9652 acquisition PCI card installed into a dedicated Dual-Xeon PC running Windows XP. The recording room is acoustically treated to acoustically isolate the chamber from the outside and to reduce reverberation (although there are reflective surfaces inside).

### 3.2. Room Geometry and Microphones Localisation

In Figure 3 we show a wire-frame simulated version of the room, in which several room devices can be seen: the rectangular table in the middle of the room, the door, the window, the frame holding the linear microphone array, and the positions of the desktop microphones and the four ones composing the quasi-squared microphone array.

The MC-REC recording application shows the text and the specific emotion that should be played by the actor. Once the operator clicks the record button, the application indicates the actor to start by showing a red circle.

The video acquisition was carried out by using a dedicated GNU/Linux workstation, running as a video server. The camera was controlled using the dvgrab free software application, modified in order to be able to work in client-server mode. Files were recorded using *720x576* resolution and 25 frames per second.

The MC-REC application is also in charge of requesting video recording to the video server, with a simple protocol (TCP/IP socket based) specifying start/stop commands and filename options.

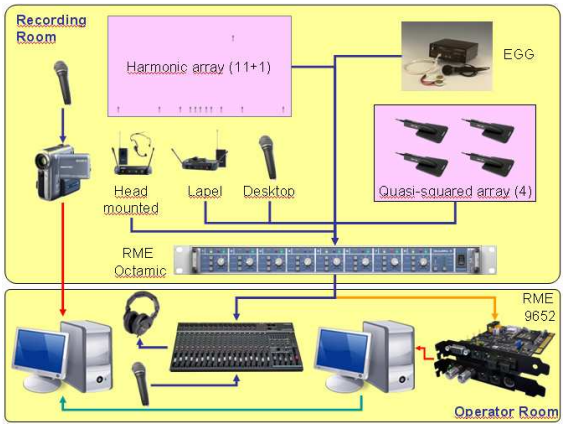


Figure 2: Schematic of recording setup

### 3.3. Cons aspects

The wide range speaking style for the acted emotions, made a difficult task to adjust mic recording levels in order to keep quality constant. Another problem due to the huge amount of recording sessions, was to keep the emotional patterns and emotional intensity constant during all sessions. The acquisition of SEV has taken more than 40 hours of recording sessions, distributed during one month (it was harmful for actors to play more than three hours a day).

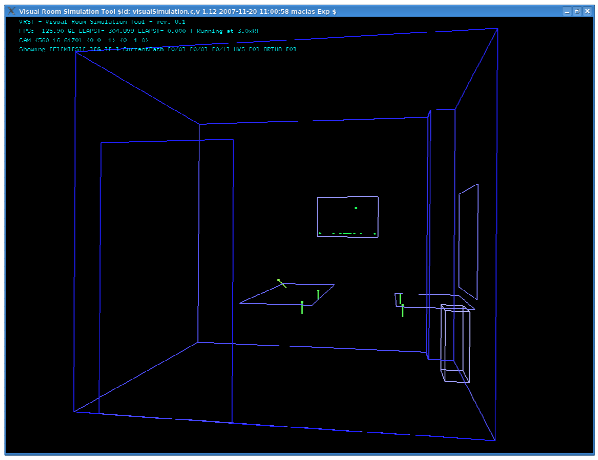

Figure 3: Wireframe simulation of recording room (3D view)

## 4. Evaluation

Close talk speech of SESII-B, QUIJOTE and PROSODIA corpus (3890 utterances) has been evaluated using a web interface. Six evaluators for each voice participated in the evaluation. They could hear each utterance as many times they need. Evaluators were asked for the emotion played on each utterance and its emotional level (choosing

between very low, low, normal, high or very high). 60% of utterances were labelled at least as a high level utterance.

Each utterance was evaluated at least by two people. The Pearson coefficient of identification rates between the six evaluators was 98%. A kappa factor of 100% was used in the validation. 89.6% of actress utterances and 84.3% of actor utterances were validated. Figure 5 plots emotion validation results.

Figure 4: Example of recording session

Whole database has been evaluated by an objective emotion identification experiment that leads 95% identification rate for both speakers. Automatic emotion identification was based on PLP speech features and its dynamic parameters. A 99% Pearson coefficient was obtained between the perceptual and objective evaluation. The mean square error between the confusion matrices of both experiments is less than 5%.

Video material is being carried out using the web interface and equivalent objective experiment to near speech is being performed with far field speech.

## 5. Phonetic and Prosodic Labeling

SEV has been phonetically labelled using HTK software (Gallardo-Antolin et al., 2007) in an automatic way. In addition to this, 5% of each sub-corpus in SEV has been manually labelled, providing reference data for studies on rhythm analysis or on the influence of the emotional state on automatic phonetic segmentation systems. EGG signal has also been automatically pitch-marked and, for intonation analysis, the same 5% of each sub-corpus has been manually revised too.

Video data has been aligned and linked to speech and text data, providing a fully-labelled multimedia database.

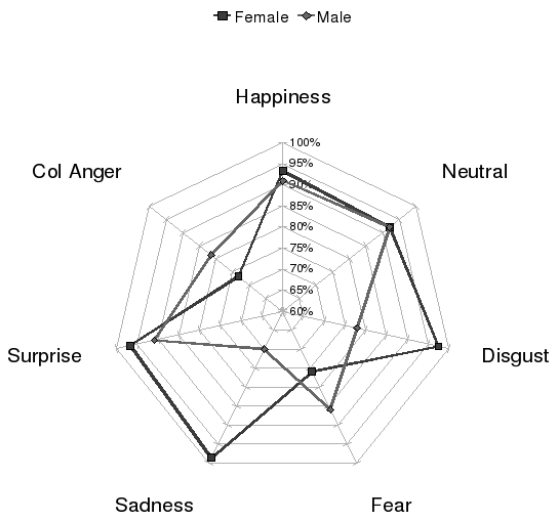


Figure 5: Average of perceptual identification rates for both speakers and emotion

## 6. Conclusions

In this paper we have presented SEV, a multimedia and multi-purpose database for research on emotional speech and video. Although part of the database was specifically designed for high-quality emotional speech synthesis and speech conversion, recorded data can be used for other emotion-related tasks such as emotional visual speech, close-talk or far-field emotion detection and emotional speech recognition or video-based emotion identification. Regarding speech, SEV covers a huge variety of contexts and situations, including recordings for Diphone-based or unit selection synthesis, and special sub-corpora for complex prosodic modelling.

Finally, the whole speech database has been evaluated through highly-correlated objective and automatic emotion identification tests. 89% of the recordings have been validated, achieving a general inter-labeller agreement higher than 0.95. 60% of the recordings were evaluated as intense or very intense in an emotional-intensity subjective test.

## 7. Acknowledgements

This work has been partially supported by the Spanish Ministry of Education & Science under contracts EDECAN (TIN2005-08660-C04-04) and ROBONAUTA (DPI2007-66846-c02-02).

## 8. References

R. Barra-Chicote, J.M. Montero, J. Macias-Guarasa, L.F. DHaro, R. San-Segundo, and R. Cordoba (2006). *"Prosodic and segmental rubrics in emotion identification"*. In Proceedingsof ICASSP, pages 1085–1088.

R. Barra-Chicote, J.M. Montero, J. Macias-Garasa, J. Gutirrez-Arriola, J. Ferreiros, and M. Pardo (2007).*"On the limitation of voice conversion techniques in emotion identification"* In Proc. of Interspeech.

F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss (2005). *"A Database of German Emotional Speech*". In Proc. of Interspeech.

G. Castellano, L. Kessous, and G. Caridakis (2007). *"Multimodal emotion recognition from expressive faces, body gestures and speech"*. In Humaine. International Conference on Affective Computing and Intelligent Interaction.

L. Chen et al.(2005). *"VACE Meeting Corpus"*. Lecture Notes in Computer Science. Pages 40-51.

N. Mana et al. (2007). *"Multimodal corpus of multi-party meetings for automatic social behaviour analysis and personality traits detection"*. In ICMI '07 Workshop on Tagging, Mining and Retrieval of Human-Related Activity Information, pages p. 9–14.

A. Gallardo-Antolin, R. Barra-Chicote, M. Schrder, S. Krstulovic, and J.M. Montero (2007)."In *Automatic Phonetic Segmentation f Spanish Emotional Speech"* In Proc. of Interspeech.

J.M. Montero, J. Gutirrez-Arriola, S. Palazuelos, E. Enrquez, and J.M. Pardo (1998). *"Spanish emotional speech from database to tts"*. In Proceedings of ICSLP, pages 923–925, September.

J.M. Montero, J. Gutierrez-Arriola, R. Cordoba, E. Enriquez, and J.M. Pardo (2002). *"The role of pitch and*

*tempo in emotional speech"*. In Improvements in speech synthesis. Ed. Wiley and Sons, pages 246–251

Marc Schoder. (2004). *"Speech and Emotion Research. An Overview of Research Frameworks and a Dimensional Approach to Emotional Speech Synthesis"*. Ph.D. thesis, Institut fur Phonetik. Universitat des Saarlandes, Saarbruken.

K.R. Schrerer and H. Ellgring (2007). *"Multimodal expression of emotion: Affect programs or componential appraisal patterns?"*. In Emotion. 2007 Feb ; vol 7 (1), pages. 158-171

N. Sebe, I. Cohen, and T.S. Huang (2005). *"Mutimodal emotion recognition: Handbook of pattern recognition and computer vision"*. In World Scientific.

| SET | TEXT | *UTT/emo* | Length (*min/emo*) | <W> | <A> |
|---|---|---|---|---|---|
| LOGATOMOS | *Isolated words* | 570 | 18 | - | - |
| SESII-A | *Short sentences* | 45 | 6 | 5 | 21 |
| SESII-B | *Long sentences* | 84 | 17 | 15 | 65 |
| QUIJOTE | *Read speech* | 100 | 22 | 16 | 70 |
| PROSODIA 1 | *A speech* | 25 | 8 | 26 | 125 |
| PROSODIA 2 | *Interview (short anwsers)* | 52 | 8 | 10 | 44 |
| PROSODIA 3 | *Interview (long anwsers)* | 40 | 10 | 20 | 87 |
| PROSODIA 4 | *Question anwsering* | 117 | 10 | 4 | 19 |
| PROSODIA 5 | *Short dialogs* | 142 | 13 | 4 | 22 |
| *TOTAL* | | *1175* | *112* | | |

Table 1: Features related to SEV size