The I3MEDIA speech database: a trilingual annotated corpus for the analysis and synthesis of emotional speech

Juan María Garrido [2, 1], Yesika Laplaza [1], Montserrat Marquina [1], Andrea Pearman [1]
José Gregorio Escalada [3], Miguel Ángel Rodríguez [3], Ana Armenta [3]

[1] Speech and Language Group, Barcelona Media Centre d'Innovació, Barcelona, Spain
[2] Department of Translation and Language Sciences, Universitat Pompeu Fabra, Barcelona, Spain
[3] Speech Technology Group, Telefónica Investigación y Desarrollo, Spain

E-mail: juanmaria.garrido@upf.edu, yesika.laplaza@gmail.com, montse.marquina@barcelonamedia.org, apearman@gmail.com, jges@tid.es, miguel@tid.es, aalv@tid.es

Abstract

In this article the I3Media corpus is presented, a trilingual (Catalan, English, Spanish) speech database of neutral and emotional material collected for analysis and synthesis purposes. The corpus is actually made up of six different subsets of material: a neutral subcorpus, containing emotionless utterances; a 'dialog' subcorpus, containing typical call center utterances; an 'emotional' corpus, a set of sentences representative of pure emotional states; a 'football' subcorpus, including utterances imitating a football broadcasting situation; a 'SMS' subcorpus, including readings of SMS texts; and a 'paralinguistic elements' corpus, including recordings of interjections and paralinguistic sounds uttered in isolation. The corpus was read by professional speakers (male, in the case of Spanish and Catalan; female, in the case of the English corpus), carefully selected to meet criteria of language competence, voice quality and acting conditions. It is the result of a collaboration between the Speech Technology Group at *Telefónica Investigación y Desarrollo* (TID) and the Speech and Language Group at *Barcelona Media Centre d'Innovació* (BM), as part of the I3Media project.

Keywords: speech database, emotion, multilingual, text-to-speech

1. Introduction

The analysis of emotional speech is currently a hot research topic for several disciplines, including Psychology, Linguistics, Acoustic Phonetics and Speech Technologies. The description of the phonetic parameters involved in the expression of emotions in speech has been goal of active research in recent years (see, for example, in the case of Spanish, Iriondo *et al.*, 2000, Montero, 2003, Adell *et al.*, 2005 or Francisco *et al.*, 2005). Apart from its intrinsic theoretical interest, the recent active research on the generation of expressive and natural speech in the field of Text-to-Speech (TTS) synthesis has augmented the need of such descriptions.

In this paper we present the I3media speech database, a trilingual (Catalan/English/Spanish) set of speech corpora collected to investigate new procedures and tools for the synthetic generation of emotional speech in those languages. It hat has been designed and recorded within I3media, a four-year multidisciplinary project funded by the Spanish government, whose aim was the development of innovative multimedia contents and tools. It is the result of the collaboration between two of the partners of the project, the Speech Technology Group at *Telefónica Investigación y Desarrollo* (TID) and the Speech and Language Group at *Barcelona Media Centre d'Innovació* (BM).

The ultimate goal of the collaboration between TID and BM within the I3media project was the development of synthetic speakers for the TID TTS system (Rodríguez *et al.*, 1998; Armenta *et al.*, 2003), able to be used for the generation of both neutral and emotional speech in a multilingual environment. To achieve this goal, several research topics were established: first, the analysis of the prosodic realisation of emotional speech; second, the use of well-known synthesis techniques, such as corpus synthesis, for the generation of emotional speech; and finally, the use of alternative synthesis techniques in the generation process, such as parametric modification of prosodic parameters. The I3media corpus was designed and collected having in mind all these research goals.

Most past research on emotional speech has been focused on the so-called 'basic' emotions (Ekman *et al.*, 1982): anger, disgust, fear, joy, sadness, surprise. However, these six emotional labels are not enough to describe the emotional variety needed for some current and future TTS applications. The I3media corpus has been conceived to represent a larger variety of emotions than other existing corpora. To do this, one of the previous tasks has been the definition of the emotions inventory that the corpus should cover. The considered inventory of target emotional and physiological states, inspired in the HUMAINE proposal (Douglas-Cowie, Cox *et al.,* HUMAINE deliverable D5f), includes 42

different labels. Three levels of intensity for each emotion have also been considered, a variable not usually included in other existing corpora.

Another typical feature of most existing emotional corpora is the use of very controlled material, which makes easy direct comparisons but offers less naturalness. This kind of material is not very suitable for real-word TTS applications. For these reasons, the I3media corpus was designed following a different approach: the selection of a set of speaking situations in which emotional speech is very likely to occur, that could be current or future scenarios of use of synthetic speech. Three different types of expressive speech were selected to be present in the corpus: call center messages, football commentaries and SMS expressive readings. The generation of call center messages is one the most typical current application scenarios of TTS systems. Football commentaries and SMS reading were considered to be possible future scenario applications of those systems in a multimedia environment (for example, in the generation of virtual speakers, one of the general research goals of I3media). A general emotional speech was decided to be included in the corpus to have a representation of other emotions considered in the established inventory not appearing in the selected scenarios.

Finally, the use of interjections ('mmm', 'oh', 'ay') and paralinguistic events (coughing, laughing, singing) is very common in natural expressive speech, and their insertion in synthetic speech has been shown to be an effective way to improve its naturalness (Eide *et al.*, 2004). For this reason, a set of such type of expressions has been included and recorded for each language.

2. Contents of the corpus

The I3media dataset includes, for each of the considered languages (Catalan, English and Spanish), six different subcorpora:

- a **'neutral' subcorpus**, a set of emotionless, informative sentences, necessary as base material for the building of TTS voices;
- a **'dialog' subcorpus**, a set of dialog sentences chosen as representative of a call center potential service;
- an **'emotional' corpus**, a set of sentences representative of pure emotional states;
- a **'football' subcorpus**, sentences representing excerpts from football broadcastings;
- a **'SMS' subcorpus**, containing readings of SMS texts;
- a **'paralinguistic elements' corpus**, a set of recordings of isolated interjections and paralinguistic sounds useful in expressive synthesis, but also interesting for analysis purposes.

These six different subcorpora are explained in more detail in the following subsections.

2.1 Neutral subcorpus

As mentioned before, the goal of the I3Media corpus was the development of synthetic multi-style, multilingual speakers for the TID TTS system. This system uses, as many other commercial ones, corpus-based techniques (Hunt and Black, 1996) for the generation for synthetic speech, which require the use of a large corpus with a sufficient phonetic coverage. The neutral subcorpus was designed to obtain the necessary phonetic material to ensure the generation of high-quality synthetic speech using corpus-based techniques in non-expressive situations. It would be used also as control material for comparison purposes in the analysis of the emotional subcorpora.

For each language, the neutral subcorpus contains a set of utterances collected from several sources, but mainly newspaper texts (1.000 in the case of Spanish and Catalan; 1.677 in the case of English), as representative of a neutral speaking style. They where selected, using a greedy algorithm, considering a set of segmental and suprasegmental phonetic variables. Table 1 presents one example for each of the considered languages.

Language	**Utterance**
Catalan	De fet, no sempre és necessari ni imprescindible dir això tan amorós quan el que passa és que fa vergonya expressar amb paraules un plaer purament sensual.
Spanish	Venezuela y Brasil iniciaron negociaciones para definir los términos de dos acuerdos conjuntos de distribución y exploración petrolera, declaró el Canciller brasileño Luiz Lampreia.
English	French West Indian products were freely imported, re-shipped, and exported, thus avoiding the rule of 1756 (85); as a result, the customs revenue leapt in one year from fourteen to twenty millions.

Table 1: Sample utterances of the neutral subcorpus

In the case of English, the neutral subcorpus included also a small subset of neutral sentences in Spanish and Catalan (250 utterances per language), selected from the Spanish and Catalan neutral subcorpora. They were included for synthesis purposes, to have a small set of phonetically rich Spanish and Catalan material, which would allow the development of a trilingual (Catalan/English/Spanish) corpus-based synthetic voice.

2.2 Dialog subcorpus

One of the typical applications of TTS systems is the generation of system messages in an Interactive Voice Response (IVR) system. For this reason, a subset of utterances typically used in automatic call center services (500 in the case of Spanish and Catalan; 195 in the case of English) was also chosen to be recorded. They were collected from a corpus of messages of actual call center services, and selected using also a greedy algorithm. In this case, in addition to the phonetic variables considered for the selection of the neutral corpus, the speech act of the utterance was also included as a selection criterion. To do this, the candidate utterances had been manually annotated with speech act tags. Table 2 includes one sample text per language of this subcorpus.

Language	Utterance
Catalan	\adh=SOLICITUD\ Què desitja, abandonar el sistema o dur a terme alguna altra cosa
Spanish	\adh=NEUTRO_DIALO\ Para confirmar el campo de golf Laukariz, pulsa 1. Si no, pulsa 3.
English	\adh=BIENVENIDA\ Wanda, welcome to the MoviStar Savings Phone Number List Consultation Service.

Table 2: Sample utterances from the dialog subcorpus. Tags indicate the corresponding speech act.

2.3 Emotional subcorpus

As stated before, the I3Media emotional corpus was designed to go beyond the prototypical approaches to emotion analysis, which usually cover only basic emotions. For this reason, an inventory of 42 emotions and physiological states, inspired in the HUMAINE proposal (Douglas-Cowie, Cox *et al.,* HUMAINE deliverable D5f), was defined specifically for the classification of the emotional material in the corpus. Tables 3 and 4 present the inventory of emotions and physiological states, respectively.

In addition to emotion itself, the intensity level of emotions (a variable not usually included in other existing corpora) was also considered for the definition of the emotion labels. Three levels of intensity were considered for each emotion: low (level 1), mid (2) and high (3). Then 126 different labels (42 emotions/physiological states x 3 levels) were finally used for the definition of the emotional states appearing in the corpus. Each label was the combination of an emotion and an intensity label ('JOY_1', ANGER_2', 'PAIN_3').

Positive	Neutral	Negative
Affection	Surprise	Mockery
Joy	Indifference	Irony
Fun		Boredom
Trust		Doubt
Excitement		Distrust
Interest		Dejection
Complicity		Disappointment
Pride		Resignation
Relief		Worry
Compassion		Envy
Admiration		Disapproval
		Nostalgia
		Shame
		Guilt
		Sadness
		Disgust
		Impatience
		Anger
		Impotence
		Fear

Table 3: Inventory of emotions considered in the I3Media corpus.

Positive	Neutral	Negative
Pleasure	Sleepiness	Fatigue
Relaxation	Agitation	Pain
		Sickness
		Cold
		Heat

Table 4: Inventory of physiological states considered in the I3Media corpus.

The emotional subcorpus was designed to have a balanced, small set of pure emotional utterances representing the whole set of defined emotional labels in the corpus. The ultimate goal of this subset was the acoustic analysis of the phonetic parameters expressing emotions, and the exploration of the use of parametric techniques for the prediction of prosody in expressive speech synthesis, taking small corpora as base material. For this reason, the number of items per considered label is rather small: three items per label in the case of Spanish and Catalan, which gave a total amount of 378 items (126 labels x 3 utterances) per subcorpus; four items per label for English, which gives 504 items in total (126 labels x 4 utterances). Table 5 includes one sample utterance per language.

Language	Utterance
Catalan	\emo=EMO_PREOCUPACION_2\ Es curarà, senyor metge? El pot curar, vostè?...
Spanish	\emo=EMO_ASCO_2\ ¿¿Tocarlos?? ¿Tocar las reliquias con las manos?
English	\emo=EMO_ALEGRIA_1\ I'm happy to see you, Charles. Very happy to see you."

Table 5: Examples of sentences of the emotion subcorpus. Tags indicate the corresponding emotion.

In this case, the corpus selection procedure was fully manual: the utterances finally selected were extracted by linguists from literary texts coming from several sources (mainly texts in electronic format, but also paper books). The texts had to express as clearly as possible the intended target emotion and intensity. No length limitation was established. Texts included, in addition to the target emotional utterances, a short context that could help the speaker to interpret the expected emotion in each case.

2.4 Football subcorpus

The football subcorpus includes a set of utterances representing a football live broadcasting situation (500 in the case of Spanish and Catalan; 242 in the case of English). This speaking style was selected as prototypical of emotional speech, and also as a possible future application of synthetic speech in expressive situations. Table 6 presents some text examples of this subcorpus in all three languages.

In this case, the selection procedure for each language involved three steps: first, a collection of electronic texts from actual live football chronicles on the internet ('minute-by-minute') was manually collected; then the collected texts were manually annotated using the emotion labels described in the previous section; and finally, the target utterances were selected, using automatic means in the case of Catalan and Spanish, manually in the case of English, in order to define a set which was representative of the distribution of the emotional labels in the whole input corpus.

Language	Utterance
Catalan	\emo=EMO_ALEGRIA_2\ Surten els jugadors dels dos equips al terreny de joc! Això és a punt de començar!
Spanish	\emo=EMO_ENFADO_1\ No me está gustando Drente... El holandés está siendo rebasado por Cazorla en todas las acciones y en la última de ellas, un pase del ex del Recre ha estado a punto de ser aprovechado por Rossi.
English	\emo=EMO_TENSION_2\ Villa win a corner on the right and after Berbatov heads straight up in the air from under his own crossbar Young is waiting on the edge of the box to volley the ball as it comes back to earth. \emo=EMO_TENSION_1\ Unfortunately for Villa his effort ricochets up off the floor and some way over the bar.

Table 6: Samples utterances from the football subcorpus.

2.5 SMS subcorpus

SMS reading-aloud is another possible TTS application in which the expression of emotions plays an important role. For this reason, a SMS subcorpus was included as part of the I3Media emotional corpus. It includes recordings of SMS text readings (500 in the case of Spanish and Catalan; 252 in the case of English). Table 7 presents examples in Catalan, Spanish and English.

In the case of the Spanish subcorpus, SMS texts were automatically selected from a SMS database collected specifically for this project. Volunteers were told to write SMS with emotional content on a web application especially conceived for this purpose. SMS texts were not fully realistic, in the sense that participants were asked to write down the text without the typical abbreviations in this type of messages. However, smileys were allowed. Then, as in the case of the football source corpus, the obtained SMS were manually annotated with emotional labels, and finally, a selection was made automatically among them to obtain a representative sample of the emotional labels found in the source corpus. For Catalan and English, no SMS database was collected: the Spanish SMS source database was manually translated before the selection process (automatic in the case of Spanish and Catalan; manual in the case of English).

Language	Utterance
Catalan	\emo=EMO_INTERES_2\ Hola, com va? Què fas per Saragossa? \emo=EMO_BURLA_1\ Vens a buscar manyes? \epa=EPA_RISA_1\Jaja\epa=EPA_NINGUNO\. \emo=EMO_NEUTRA\ Aquest cap de setmana vaig a Madrid.
Spanish	\emo=EMO_SORPRESA_2\ ¡No me lo puedo creer! \smi=SMI_SORPRESA_2\ :-O \smi=SMI_NINGUNO\ ¿En serio? \emo=EMO_ALEGRIA_2\ ¡Eso es estupendo, me alegro un montón! \emo=EMO_AFECTO_2\ ¡Un besazo! \smi=SMI_BESO_1\ :-* \smi=SMI_NINGUNO\
English	\emo=EMO_ADMIRACION_2\ Wow! You were right! The view is incredible from the Eiffel Tower. I'll send you photos. \emo=EMO_AFECTO_2\ \smi=SMI_BESO_1\ :-* \smi=SMI_NINGUNO\ Love and hugs!

Table 7: Sample utterances from the SMS subcorpus. Tags indicate smileys and emotions.

2.6 Paralinguistic elements subcorpus

The use of interjections and paralinguistic sounds, such as coughs or kisses, is quite common in expressive speech, and quite often they are carriers of an important part of the expressive contents of utterances. Also, the use of pre-recorded paralinguistic elements has proven to be an effective way to generate realistic synthetic speech (Eide *et al.*, 2004). For this reason, a subcorpus of such elements was included in the I3Media corpus. It contains the recordings of an inventory of linguistic (interjections) and paralinguistic (laughs, coughs, etc.) elements frequently used in expressive speech. The interjection inventory was defined specifically for each language (36 interjections in the case of Spanish, 40 in the case of Catalan, and 49 in the case of English); a common inventory of 15 paralinguistic elements was also defined for all three languages. As in the case of emotions, three different levels of intensity were considered for each element. Also, for some interjections and paralinguistic sounds, different versions were recorded, with varying intonation depending of the emotion being expressed. So, for example, in the case of the English interjection 'ah', up to five variants were recorded, expressing happiness, fear, or surprise, among others. Table 8 presents the inventory of interjections recorded for English, and table 9 shows the complete inventory of paralinguistic sounds.

Ah	Mmm
Aha	Oh
Argh	Ok
Aww	Oops
Ay	Ouch
Bah	Ow
Blech	Phew
Eh	Pshaw
Erm	Psst
Ew	Sheesh
Geez	Shhh
Hey	Ugh
Hmm	Uh
Huh	Uh-huh
Hurrah	Uh-oh
Meh	Whew

Table 8: Inventory of interjections considered in the English paralinguistic elements subcorpus.

Kiss	Yawn
Humming	Click (of the tongue)
Shushing	Sneeze
Grumbling	Crying
Laughter	Snoring
Whistle (admiration)	Whistle (happiness)
Strong exhalation	Coughing
Shigh	

Table 9: Inventory of paralinguistic sounds considered in the paralinguistic elements subcorpus.

3. Speaker selection

The selection of the speakers is also a key task in the development of an emotional corpus, and a special attention was paid to this aspect in the I3media project. The Spanish and Catalan corpora were read by the same speaker, considering that the ultimate goal was the development of a bilingual Spanish/Catalan synthetic voice. A bilingual professional actor, with good language and acting skills, was selected for this task. The selection procedure included a small casting among six candidates, with experience in theatre or dubbing acting, which previously recorded a corpus of basic emotions which served as test material for the final decision. In the case of the English corpus, the selected speaker was a woman, with acceptable skills in Spanish and Catalan and an excellent competence in English. Acting criteria were also considered in the selection procedure, but no casting corpus was recorded in this case. In both cases, technical

criteria (compatibility of the voice with signal processing techniques) were also taken into account.

4. Recording conditions

The corpus was recorded at UPF premises, in a professional sound recording room. The Sony Vegas program running on a PC with a RME Hammerfall HDSP 9652 soundcard, and a Yamaha 02R96 mixer with a ADAT MY16AT card, were used for recording. Speech signal was stored in wav files, at a sampling frequency of 48 KHz. Texts were presented on a TV screen placed in front of the speaker. The speaker was free to move and gesticulate while uttering, using a high quality headset microphone which allowed to keep constant the distance between the speaker's mouth and the microphone during the recordings.

5. Corpus annotation

After recording, the corpora were processed at TID and BM to be phonetically segmented and annotated, both for analysis and synthesis purposes. The resulting annotation includes phonetic segmentation (phones, syllables, intonation groups), F0 labelling, and a series of labels emotion, emotion level, sentence mood, speech act, etc.) relevant for the analysis and synthesis of prosody in emotional situations. Figure 1 shows the appearance of this annotation when converted to a Praat (Boersma and Weenink, 2009) Textgrid file.

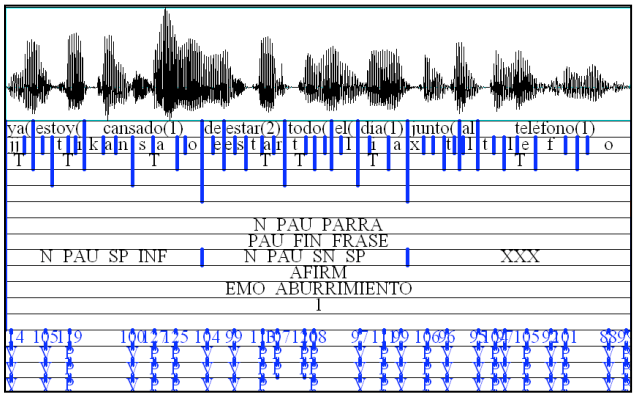

Figure 2: Example of Praat screen showing the speech signal and the corresponding annotation for a sample utterance of the Spanish emotional subcorpus.

6. Corpus exploitation

The I3Media corpus has been used for the development of two multilingual, corpus-based, synthetic voices, the first one male, bilingual Catalan/Spanish, and the second one female, trilingual English/Catalan/Spanish, compatible with the TID text-to-speech engine. The Catalan/Spanish voice was also integrated, as part of the tasks of the I3Media project, with the Activa Multimedia avatar to develop a prototype of virtual football speaker with expressive voice.

Also, the corpus has been used to carry out several acoustic analysis of the prosodic cues involving the expression of emotions in speech (intonation, speech rate, pausing), both n Catalan and Spanish (see, for example, Garrido *et al.*, 2012). The results of these analyses have been used to explore, develop and integrate in the TID TTS system a new parametric procedure to the generation of prosody in emotional speech.

7. References

Adell, J., Bonafonte, A., Escudero, D. (2005). Analysis of prosodic features: towards modelling of emotional and pragmatic attributes of speech. *Procesamiento del Lenguaje Natural*, 35, pp. 277-283.

Armenta, A., Escalada, J. G., Garrido, J. M., Rodríguez, M. A. (2003). Conversor texto a voz multilingüe de Telefónica I+D. *Procesamiento del Lenguaje Natural*, 31, pp. 331-332.

Boersma, P, Weenink, D. (2009). Praat: doing phonetics by computer (Version 5.1.05) [Computer program]. Retrieved May 1, 2009, from http://www.praat.org/.

Douglas-Cowie, Cox *et al.* HUMAINE D5f deliverable; http://emotion-research.net/download/pilot-db/ [last access: 24.10.2011].

Eide, E., Aaron, A., Bakis, R., Hamza, W., Picheny, M, Pitrelli, J. (2004). A corpus-based approach to expressive speech synthesis, In *SSW5-2004*, pp. 79-84.

Ekman, P., Friesen, W. V., Ellsworth, P. (1982). What emotion categories or dimensions can observers judge from facial behavior? In P. Ekman (Ed.), *Emotion in the human face*. New York: Cambridge University Press, pp. 39-55.

Garrido, J. M., Laplaza, Y., Marquina, M. (2012). On the use of melodic patterns as prosodic correlates of emotion in Spanish. In *Speech Prosody 2012 Proceedings*.

Francisco, V., Gervás, P., Hervás, R. (2005). Análisis y síntesis de expresión emocional en cuentos leídos en voz alta. *Procesamiento del Lenguaje Natural*, 35, pp. 293-300.

Hunt, A. J., Black, A. (1996). Unit selection in a concatenative speech synthesis system using a large speech database**.** In *IEEE International Conference on Acoustics, Speech, and Signal Processing,* 7-10 May 1996. *ICASSP-96. Conference Proceedings*, Volume 1, pp. 373-376.

Iriondo, I., Guaus, R., Rodríguez, A., Lázaro, P., Montoya, N., Blanco, J.M., Bernadas, D., Oliver, J.M., Tena, D., Longhi, L. (2000). Validation of an acoustical modelling of emotional expression in Spanish using speech synthesis techniques. In *Proceedings of the ISCA Workshop on Speech and Emotion, Northern Ireland*, pp. 161–166.

Montero, J.M. (2003). Estrategias para la mejora de la naturalidad y la incorporación de variedad emocional a la conversión texto a voz en castellano. Ph.D Thesis, Universidad Politécnica de Madrid.

Rodríguez, M.A.- Escalada, J. G., Torre, D. (1998). Conversor texto-voz multilingüe para español, catalán, gallego y euskera. *Procesamiento del Lenguaje Natural*, 23, pp. 16-23.