# A Catalan Broadcast Conversational Speech Database

*Henrik Schulz, José A. R. Fonollosa*

Department of Signal Theory and Communications
Technical University of Catalunya (UPC), Barcelona, Spain
hschulz@gps.tsc.upc.edu, adrian@gps.tsc.upc.edu

## Abstract

Data driven methods in speech and linguistic research, and system develoment require appropriate speech databases. A new Catalan speech database has been developed with a particular emphasis on broadcast conversational speech. The article describes origin and nature of the broadcasts and its acoustic environment. Annotation and transcription provide statistics on specific phenomena of exhibited speech, speaker characteristics and acoustic events. It concludes with perspective uses and limitations.

**Index Terms**: Catalan, audio video speech database, broadcast conversation, literal transcription, spontaneous speech

## 1. Introduction

State-of-the-art empirical and statistical data driven methods in speech processing depend to a large extend on sufficient and appropriate sample data, often covering a particular domain, acoustic environment or recording channel. The tremendous variability of phenomena in spoken language and those across speakers and speaking styles often require large amounts of data for robust system development. Generally, an appropriate coverage is often achieved across a large variety of speakers, dialects, gender and age. But, for example, in terms of speech recognition, most system models provide best results for a specific task or domain that limit the acoustic and linguistic variability, leading to task specific collections. Likewise for providing evidence for a particular acoustic or linguistic phenomenon in question, the data not only need to be sufficient, but also need to comply to secondary conditions.

Speech databases for the purpose of speech and linguistic research, and for system development have been collected for Catalan as well as for other languages. Nevertheless, and in particular to study and develop technologies for regional languages, the available data seem still insufficient and demand further efforts.

The article describes a novel Catalan broadcast conversational speech database. Its original video recordings were supplied directly by the Catalan TV3 broadcast station and contain 32 live television broadcasts of the Àgora (Greek: a place of political and juridical assembly) programme, that are debates on selected topics from politics, economy or society.

Each broadcast follows a repeating format: the anchorman is initially presenting the current topic, followed by an introduction of invited participants featuring background music. The main part features the debate between the invited participants, usually public figures. During the debate, public opinions are added, either as e-mails or faxes read by the anchorman, or telephone recordings played back again featuring background music. Between these general elements of the broadcast format short terms of music. Although the original TV broadcasts are usually interrupted once by a commercial, the final recordings do not contain them.

The debate generally comprises spontaneous speech, whereas the introduction of topic and participants features planned speech. Both differ significantly acoustically and linguistically.

Catalan - mainly spoken in Catalonia - exhibits substantial dialectual differences, dividing the language into an eastern and western group. The eastern dialect includes Northern Catalan (French Catalonia), Central Catalan (the eastern part of Catalonia) and Balearic. The western dialect includes North-western Catalan and Valencian (south-west Catalonia)[1].

The origin of the recorded Catalan participants suggests a predominant Central Catalan dialect. Furthermore, the rare selection of alternative pronunciations other than those equal to Central Catalan during the alignment of acoustic data while building ASR acoustic models supports this impression.

Although TV3 is primarily a Catalan television channel, the recorded broadcasts contain a surprisingly high proportion of Spanish speaking participants. The assessment of the database annotation will therefore emphasis on both languages.

## 2. Data and Annotation

The 32 Àgora television broadcasts are originally available as video files, those audio channels were extracted and stored as 16 bit PCM, mono with its original 32 kHz sampling rate. In order to facilitate transcription and further processing the audio data have been downsampled to 16 kHz. Each broadcast was partitioned into two files at the time of the original commercial. The annotation procedure followed a 3-pass approach: a markup of initial segmentation, a full literal transcription and speaker annotation, and finally a verification and refinement of boundaries by another transcriber. The transcription was carried out and formated according to [2] and leads to an XML-like file format.

Several conditions were qualified regarding speaking mode: planned or spontaneous; background : music, noise and speech; and channel: studio, telephone, outside (typically public places).

The original broadcast conversations bear a frequent speaker change, contain segments of music and speaker overlap. Segments of speaking rate acceleration can be observed. Linguistically, the conversations possess frequent repetition and repairs. Furthermore a measurable amount of mispronunciations, incompleteness and filled pauses.

Every speaker turn is decomposed into segments by placing breakpoints at grammatical sentence boundaries. Spontaneous utterances or incomplete phrases are separated at audible natural boundaries, i.e. at pauses, breaths, etc. and at interruptions of the sentence flow. The orthographic literal transcription is case sensitive and contains applicable punctuation marks.

## 3. Acoustic Environment

Primarily, the broadcasts are studio recordings. As the format contains public opinions provided via telephone, several segments have been identified accordingly. Furthermore, these segments feature background music. The speech component of these segments is therefore bandlimited to frequencies from 300 Hz to 3.4 kHz, while the background music employs the full bandwidth. Few segments originate from recordings of public places, again gathering public opinions that feature background music. The tables 1, and 2 display the durations of the given recording and background conditions and follow the distinction of planned and spontaneous speech. The categories 'Music' and 'Speech' denote the respective background condition, whereas 'None' refers to clean speech.

| | Planned [h] | | Spontaneous [h] | | |
|---|---|---|---|---|---|
| | None | Music | None | Speech | Music |
| Studio | 0:30 | 1:50 | 20:37 | 2:51 | 0:59 |
| Telephone | - | - | - | - | 0:52 |
| Outside | 0:03 | 0:02 | 0:30 | 0:06 | 0:28 |

Table 1: Duration breakdown regarding recording environment and background conditions for Catalan segments

Since planned speech originates rather from the anchorman than from invited speakers, there are no segments of planned speech among the Spanish segments in table 2. The share of background speech, i.e. segments possessing events of non-transcribed overlapping speakers is rather large. It denotes solely background speech of speakers that could not be identified for their term. The total number of such events amounts to 2843. A segment featuring an event usually extends to a longer duration than the actual event itself. The phenomenon also indicates a high spontaneity and naturalness of the discussion. There are few segments of pure music and pure silence as well as of a combination of music and background speech of minor duration which have not been tabulated.

| | Spontaneous [h] | | |
|---|---|---|---|
| | None | Speech | Music |
| Studio | 7:02 | 0:41 | - |
| Telephone | - | - | 0:05 |
| Outside | 0:01 | - | - |

Table 2: Duration breakdown regarding recording environment and background conditions for Spanish segments

## 4. Speakers

The set of speakers for each broadcast is formed by the invited participants and the anchorman, and is extended where applicable by people offering a short opinion on the topic via telephone. Each transcription document encloses a set of speaker descriptions, each with a unique identifier, full name (where identifiable), gender, and primary language. In case a speaker converses in multiple languages, the record is duplicated with a new identifier. As speaker overlap is a typical phenomenon during the debates, a combination of overlapping speakers receives an additional record solely with its identifier. The speaker distribution for Catalan and Spanish segments is shown in tables 3 and 4 respectively. The unknown gender category refers to segments of identifiable overlapping speakers but does not consider background speech (further see section 6). Likewise contributions of speakers via telephone as those received gender and language with their speaker record.

| Gender | # Speakers | Duration [h] | # Segments |
|---|---|---|---|
| male | 441 | 24:33 | 25335 |
| female | 113 | 3:51 | 3848 |
| unknown | 317 | 0:40 | 623 |

Table 3: Catalan Speaker Distribution

| Gender | # Speakers | Duration [h] | # Segments |
|---|---|---|---|
| male | 83 | 6:19 | 6869 |
| female | 29 | 1:20 | 1327 |
| unknown | 45 | 0:13 | 174 |

Table 4: Spanish Speaker Distribution

The speaker distributions indicate a clear gender misbalance among the data. The segments of unknown speaker are short and accumulate only a small contributions to the total duration.

## 5. Speech Events

Speech events comprise ordinary spoken words, incomplete words and hesitations. Table 5 shows the amount of speech events for Catalan and Spanish segments, whereas the total number of elements denotes the sum of all speech and non-speech events. The ratios in parenthesis refer to the total number of elements. Hesitations, i.e. a single syllable speech disfluency not consistent with the grammatical structure of a sentence, and incomplete words emphasizing the spontaneity of the language.

| # elements | Catalan | Spanish |
|---|---|---|
| total | 401701 | 95575 |
| running words | 373886 (93%) | 88614 (92%) |
| hesitations | 6007 (1.5%) | 1703 (1.7%) |
| incompletes | 2290 (0.6%) | 530 (0.6%) |
| words | 21908 | 8854 |

Table 5: Speech Events

Utterances primarily spoken in Catalan also contain 2600 Spanish words, and vice versa primarily Spanish utterances contain 142 Catalan words. For the sake of completeness, occurances of words of other languages across all segments: English 350, Arabic 132, French 62, German 35, Mandarin 23, Portuguese 21 and Italian 10. Words with 'unknown' spellings add up to 1963, those featuring a poor intelligibility accumulate to 11, and those being pronounced incorrectly 1392.

## 6. Non-Speech Events

The database annotation distinguishes several non-speech events as follows:

- throat - coughing, clear one's throat
- breath - audible breath noise
- voice - untranscribed overlapped or background speech
- laugh - laughing
- artic - non verbal articulatory noise of the speaker, e.g. smack, swallowing, etc.
- pause - silence, i.e. long speaker pause ($>$ 1 second)
- sound - non-articulatory harmonic noises, e.g. short music parts, beeps, other sound effects, etc.
- rustle - rustling such as with paper or microphone rustle
- noise - any other noise not particularly identified above, inharmonic noise, non-articulatory events like, e.g. knocking, babble of voices, machines, etc.

Table 6 tabulates the number of occurances of above listed events (needless to say, although the table refers to Catalan and Spanish segments, the events are not anticipated to be different in nature).

| # events | Catalan | Spanish |
|---|---|---|
| breath | 16777 (4.1%) | 3992 (4.1%) |
| throat | 357 (.08%) | 97 (.10%) |
| rustle | 337 (.08%) | 81 (.08%) |
| voice | 1352 (.33%) | 270 (.28%) |
| laugh | 163 (.04%) | 38 (.04%) |
| pause | 152 (.03%) | 76 (.08%) |
| sound | 311 (.07%) | 36 (.03%) |
| artic | 1224 (.30%) | 319 (.33%) |
| noise | 1135 (.28%) | 349 (.36%) |

Table 6: Non-Speech Events

## 7. Discussion

The above described new Catalan (-Spanish) speech data contain a wide range of phenomena of natural language. With the spontaneous nature of the broadcast conversations, the data feature characteristic phenomena particularly occuring in spontaneous speech, e.g. non-verbal events, disfluencies, incompleteness, repetition and repair, and therefore encourage to be subject of a profound linguistic and acoustic analysis. The distinction into planned and spontaneous speech may motivate further investigation. Aside from the fundamental research, the database supports the development of speech processing systems, in particular speech recognition, acoustic event detection, as well as language and speaker identification and tracking.

Speaker and music overlap opens challenges to automatic literal transcription, and speaker and language identification tasks. A frequent speaker change and speaking rate acceleration induce grammatically incomplete sentences and orthographically incomplete words. The speaking rate acceleration in parts of the debate also accounts for a higher dynamic in allophone durations. Both may be subject to benchmarks of existing approaches in language and acoustic modelling as well as opportunity for refinement.

The speaker annotation and frequent speaker change facilitate studies with emphasis on identification and tracking across a particular broadcast or a collection. Moreover, gender information provided for each speaker may facilitate features and models for the distinction of speakers aside of a pure gender detection task. From the automatic speech recognition point of view, gender specific acoustic models can provide an increase of allophone discriminance.

As the database provides overlapped speech to a larger extend it may also become subject to overlapped speaker detection and as a consequence to speech recognition.

The language annotation provides sufficient references for language identification and tracking in a bilingual environment, e.g. in the autonomous regions in Spain, where aside of preferred use of the regional language, Spanish is well spoken and understood.

Although the database provides a rich annotation, it needs to be noted, that a rather precise segmentation of acoustic events, i.e. the provision of synchronization marker, is not available although desirable for some of the above mentioned research topics.

The gender misbalance as shown in table 3 and 4 should be considered when deriving conclusions for particular phenomena or methods.

## 8. Conclusions

In order to facilitate speech research and development activities in Catalan, a broadcast conversational speech database has been derived from recordings of the Catalan TV3 Àgora programme. The recordings were segmented according to its speakers, distinct environment and channel conditions, and received literal transcriptions as well as detailed annotations with respect to acoustic events, speaking mode and speakers. The article described major aspects of the database material and its annotation. Furthermore it provides qualitative measures to essential phenomena and elements of annotation. It finally identified and discussed major subjects of its application. However, it needs to be noted, that in particular for speech recognition system development, the amount of available Catalan speech data in the target domain is not competitive with other languages so far. As further collection efforts are recommended, a new collection of general broadcast news data is in progress.

## 9. Acknowledgements

The database development was funded by the Generalitat de Catalunya in the framework of the TECNOPARLA project [3]. The authors would like to thank Mateu Aguiló Bosch for the coordination efforts and Verbio Technologies S.L for performing the transcription.

## 10. References

[1] Max W. Wheeler, *The Phonology of Catalan*. Oxford, UK: Oxford University Press, 2005.

[2] C. Barras, E. Geoffrois, Z. Wu, and M. Liberman, "Transcriber: a free tool for segmenting, labeling and transcribing speech," in *First International Conference on Language Resources and Evaluation (LREC)*, 1998, pp. 1373–1376.

[3] H. Schulz, M. R. Costa-Jussà, and J. A. R. Fonollosa, "TECNOPARLA - Speech Technologies for Catalan and its Application to Speech-to-Speech Translation," in *Procesamiento del Lenguaje Natural*, 2008.