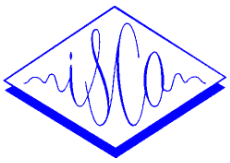

ISCA & IEEE Workshop on Spontaneous
Speech Processing and Recognition
Tokyo Institute of Technology, Tokyo, Japan
April 13-16, 2003

ANNOTATION AND ANALYSIS OF ACOUSTIC AND LEXICAL EVENTS IN A GENERIC CORPUS OF SPONTANEOUS SPEECH IN SPANISH

L.J. Rodríguez, I. Torres

Departamento de Electricidad y Electrónica
Facultad de Ciencias
Universidad del País Vasco
Apartado 644, 48080 Bilbao, SPAIN
e-mail: {luisja,manes}@we.lc.ehu.es

ABSTRACT

This paper presents the annotation and statistical analysis of spontaneous speech events in a series of broadcast news interviews drawn from the so called *Corpus Oral de Referencia de la Lengua Española Contemporánea*. The annotated corpus consists of 42 interviews taken from radio and television broadcasts, fully transcribed and lasting 6.41 hours. The corpus is intended primarily to compare frequencies and typologies of spontaneous speech events between task-specific and generic speech, but also to train acoustic and language models and carry out recognition experiments. The annotation process involved two steps: (1) filtering the initial transcriptions, and (2) augmenting the filtered transcriptions with acoustic and lexical events. Filtering was applied not only to adapt the orthographic conventions and the mark-up format but also to discard some of the marks, which were irrelevant from the point of view of speech recognition. Besides human and non-human noises, annotation included acoustic events: lengthenings, silent pauses and filled pauses; lexical events: cut-off words, mispronunciations and guttural affirmations; and speech overlaps, which rarely appear in human-computer dialogues. Statistics show that the probability of finding one of such events at each word is 0.19.

1. INTRODUCTION

New continuous speech recognition and understanding applications, such as speech-to-speech translation or dialogue-based information access, require handling spontaneous speech, which involves hesitations, repetitions, self-corrections, non linguistic sounds, etc. We will refer to these phenomena as *spontaneous speech events*. Though the term *disfluency* is also widely used, it has a more restricted meaning, involving only a subset of the events we are interested in. Formally, we define a *spontaneous speech event* —hereafter, SSE— as *any feature, at any level (acoustic, lexical, syntactical or even pragmatic), specific to spontaneous speech, i.e. present in spontaneous but not, or very rarely, in read speech*. These features arise from various conditions of spontaneous speech: environment (noises), type of interaction (simultaneous speech) and speech modality (hesitations, repetitions, etc.). They must be all identified, annotated and adequately modelled in the framework of a speech recognition system.

This work was partially supported by the Basque Country University, under a generic grant for research groups, and the Spanish MCYT, under projects TIC2001-2812-C05-03 and TIC2002-04103-C03-02.

The most successful methodologies applied nowadays in speech recognition require a large number of samples to train acoustic and language models suitable for spontaneous speech. The effort involved in acquiring and annotating such databases is enormous. As a result research focuses on short-term practical applications with a better return on investment. Various projects are currently in progress in Spain aimed at developing interactive dialogue-based systems that automatically perform very specific tasks (with not very complex syntax and a limited vocabulary). All of these projects must deal with spontaneous speech in Spanish language. A task oriented database —called INFOTREN— was recently acquired and annotated as part of one of these projects [1]: a spoken dialogue interface that provides information about train schedules, prices, etc. The annotations included a wide range of SSEs: noises, filled pauses, cut-off words, self-corrections, discourse markers, etc. [2]. Additionally, acoustic events were modelled and integrated in the speech recognition, resulting in improved performance [3].

However, the spontaneity found in a human-computer domain-restricted task may differ both in typology and intensity from that found in casual speech. Studying how much and why speech recognition performance degrades as a result of that change in modality —with more noticeable SSEs— might suggest ways of modelling spontaneous speech. Additionally, it will provide a reference rate for completely unrestricted spontaneous speech. To accomplish such a study, a database of casual speech should be acquired and annotated with SSEs, and eventually a speech recognition system tested on it.

Instead of creating a new database, we recycled and adapted an existing corpus in Iberian Spanish, the so called *Corpus Oral de Referencia de la Lengua Española Contemporánea* —CORLEC, hereafter—, recorded by the *Universidad Autónoma de Madrid* in 1991 for making theoretical studies of spoken language [4]. CORLEC includes both speech signals and orthographic transcriptions. The transcriptions are intended to reflect the acoustic content of speech signals, and therefore include many SSEs at the acoustic and lexical levels: silent pauses, filled pauses, cut-off words, mispronunciations, etc. It contains around one million words and approximately 100 hours of speech. However, from the point of view of acoustic modelling, CORLEC has one significant drawback: recording conditions were really poor. On most occassions, an audio tape recorder was used, which was placed on a table while people were speaking. Speech signals were stored on analog audio tape, and later transferred to single channel files. However, this

was not a problem for the research group carrying out the acquisition, since they were only going to use the signals for generating the transcriptions, and then studying phonological, morphological, syntactical and pragmatical aspects of spoken language. To summarize, CORLEC is a large and representative corpus of casual, spontaneous, completely unrestricted but —unfortunately— quite noisy and low-quality speech.

The rest of the paper is organized as follows: Section 2 sets out the criteria applied to draw the dialogues from CORLEC; Section 3 deals with the rules applied to filter the original transcriptions to fit our orthographic and format conventions; Section 4 explains how the dialogues were cut into turns and annotated; Section 5 presents the absolute and relative distribution of SSEs in the subcorpus, and briefly discusses them; finally, conclusions are given in Section 6, along with ongoing work and future lines of research.

2. DEFINING THE CORPUS

CORLEC is a spontaneous speech corpus in Spanish covering many semantic and pragmatic domains, composed of monologues and multiparty dialogues taken from radio and television broadcasts, daily conversations, academic lectures, round-table discussions/debates, etc. *Informants* (speakers) were drawn from various socio-cultural backgrounds, and dialogues were held in different situations, either formal or familiar (and all intermediate types). These features perfectly fitted our needs: a generic spontaneous speech corpus, large enough to train acoustic and language models, with a non restricted syntax and a large vocabulary.

Table 1. Number of word samples (S), vocabulary size (W) and average number of samples per word (S/W) in the original transcriptions, for the 17 blocks of CORLEC.

Block	S	W	S/W
Administrative	6322	1080	5.8537
Scientific	35172	4857	7.24151
Conversations	207748	14808	14.0294
Debates	81928	8557	9.57438
Sport	47165	5597	8.42684
Documentary	26779	4721	5.67232
Educational	59240	6429	9.2145
Interviews	147468	12813	11.5092
Humanistic	53432	7150	7.47301
Instructions	7175	1321	5.43149
Legal	34386	4247	8.09654
Games	50347	6356	7.92118
News	65373	8389	7.7927
Political	48604	5864	8.28854
Advertising	24896	3864	6.44306
Religious	11162	2298	4.85727
Technical	34687	4333	8.00531

CORLEC contains 941386 words (around 100 hours of speech), with a vocabulary of 39785 words. This was considered too large, so a smaller subcorpus was drawn from it. CORLEC comprises 17 blocks, defined according to either the semantic domain or the speech modality. Table 1 shows, for each block, the number of words, the vocabulary size and the average number of samples per word. Those blocks with the highest rate of samples per word were chosen: conversations (14.03 samples/word, over 207748 words) and interviews (11.51 samples/word, over 147468 words). Note that the larger the block the higher the rate of samples (Pearson's correlation coefficient $\rho = 0.9495$, two tailed t-test, t(15)=11.7183, p=0.0000). This is consistent, since the probability of unseen words reduces as the block size increases. Indeed, taking the full corpus, the average number of samples per word rises to 23.67, which is the same ratio found for the smaller but task-specific database INFOTREN [2]. The rate of SSEs in the original transcriptions was used as secondary information to help in making the decision. There were 92412 SSEs in the corpus, giving 0.098 SSEs per word on average. All the blocks showed similar rates, 13 of them —including conversations and interviews— between 0.08 and 0.12. Additionally, conversations and interviews jointly account for 41.63% of the SSEs, and therefore seem a suitable choice. Finally, after listening to the speech signals, noisy dialogues were discarded, obtaining a set of 132 dialogues: 67 interviews and 65 conversations (see Table 2) which we will call CORLEC-EHU.

Table 2. Subcorpus CORLEC-EHU: number of acoustically useful dialogues (and total), turns in useful dialogues and their accumulated duration (in seconds).

Block	Useful (Total)	Turns	Duration
Conversations	65 (126)	9691	38383
Interviews	67 (79)	4502	38907

The block of conversations is composed of open dialogues involving two or more speakers, with a large number of overlaps, since turns are not given but freely taken. The block of interviews consists of more formal dialogues, in most cases between two speakers, one acting as the interviewer and asking questions and the other answering them, sometimes in the form of long monologues. The conversations were recorded at home, in family meetings, or while travelling, so they were very noisy, with echo, etc. (only 52% included in the subcorpus). On the other hand, the interviews were all taken from TV or radio broadcasts, so most of them (85%) were acoustically useful.

3. FILTERING THE ORIGINAL TRANSCRIPTIONS

The work carried out for INFOTREN led to a first set of SSEs suitable for human-computer interactive dialogues. But human-human dialogues contain a wider range of phenomena. In particular, the basic inventory defined for INFOTREN was expanded with speech overlaps and guttural sounds, the latter commonly used as pseudo-words for accepting or rejecting a previous statement. Markup conventions for speech overlaps (and also for noisy segments) included a special mark to indicate that the turn was continued afterwards. Also, special marks were defined to account for acronyms and foreign words. The former were transcribed according to their pronunciation, either with all or with only some of the letters spelled. The latter were given an approximate orthographic transcription in Spanish. The low quality of recording media forced the definition of a special mark for cuts, which occurred either because the audio tape ran out, or because of an error in transferring analog audio signals to digital format. Finally, due to noisy environments or speech overlaps, some segments were unintelligible and thus not transcribed, but instead assigned a special mark. Cuts and unintelligible segments both implied the end of the turn.

The same simplified markup format defined for INFOTREN was used to express the expanded inventory of SSEs of CORLEC-EHU. Simplified marks, suitable for the annotation task, were subsequently translated to a definitive and more portable XML-based format, and put into headerless files, with the same orthographic conventions defined for INFOTREN. The original CORLEC transcriptions, on the other hand, had been stored in SGML files —partially in compliance with TEI guidelines—, with the typical structure of a header containing information about speakers, annotator, recording conditions, etc. followed by the transcriptions. Since both the orthographic conventions and the inventory of marks differ from ours, a translation tool (a Perl script) was designed to convert the original CORLEC transcriptions into simplified format. Among others, the following rules were applied:

- The header is deleted and each turn is preceded by a speaker identifier and a turn index (the latter was not present in the original transcriptions).
- Marks describing speech modality, such as double quotes (marking indirect speech), *<singing>*, *<read>*, etc. are all deleted.
- Punctuation marks are reduced to commas and full stops (colons and semicolons are converted into commas), and are separated from words by blank spaces.
- Phoneme deletions are treated as mispronunciations.
- Ellipsis (suspension points) immediately following a word were used in CORLEC to mark a pause. If the word ends in a vowel, 'n', 'l' or 's', this event is translated as a lengthening plus a silent pause. Otherwise it is translated as a silent pause.
- Phatic sounds marking affirmation or negation are translated as the corresponding guttural sounds. The remaining phatic sounds, as well as the sequence *"eh..."* are translated as filled pauses.
- Laughter, music, applause and other noises are all translated as generic background noise.

4. THE ANNOTATION PROCESS

The transcriptions resulting from the above transformations were taken as input by the annotator, for two main purposes: (1) to cut the speech signal corresponding to each dialogue into as many files as there were useful turns for that dialogue, and (2) to correct and augment the annotations of SSEs at the acoustic and lexical levels. We define *useful turns* as those that can be used to train acoustic models. Turns or turn segments affected by speech overlaps and/or background noise were annotated with SSEs but their signals were discarded. Annotations include the beginning and end of noisy/overlapping segments, so that we can easily eliminate them, leaving only the transcriptions of those segments actually cut. In fact, noisy/overlapping segments may appear either as the final part of a turn, as the initial part of a turn or as the whole turn, but never in the middle of a turn. The cuts were made so that each turn index was assigned at most one speech file.

SSEs were marked inside noisy/overlapping segments because: (1) we wanted complete and detailed counts of such phenomena; (2) the language models would be trained with the full transcriptions, corresponding either to *clean* or noisy segments, and even joining consecutive continued turns; and (3) future research might allow noisy/overlapping segments to be handled.

The annotation task was assisted by a simple text editor —configured so that simplified marks appeared in different colours— to correct and augment the transcriptions, and XWAVES to listen to the signals (16 kHz, 16 bits, linear, 1 channel), to select and cut the segments and store them in ESPS sampled data files. Besides marking SSEs, some other convenient changes were made: numbers, ordinals and dates —which sometimes appeared as numbers in the original transcriptions— were all orthographically transcribed just as they were pronounced.

To date, 47 of 67 interviews have been processed, 5 of which have been discarded (due to noisy conditions) and 42 (lasting 6.41 hours) fully annotated. The process has taken around 180 hours, giving an average of 30 hours of annotating work per hour of speech. The annotations have been subsequently passed through a parser that located annotation errors, and then corrected and fully reviewed by a second annotator. We will refer to this subcorpus as CORLEC-EHU-1. Although this is just one part of CORLEC-EHU, including only acoustic and lexical events, we consider it large enough to draw statistically significant conclusions about the typologies and frequencies of such phenomena in generic spontaneous speech. In the following section we present absolute and relative frequencies of acoustic and lexical events.

5. DISTRIBUTION OF EVENTS

CORLEC-EHU-1 contains 2090 useful turns, amounting to 20197 seconds (5.61 hours, 87.41% of the speech signals). This gives an average of 9.66 seconds/turn, with a standard deviation of 14.40, revealing the high variability of turn durations. Indeed, the histogram of turns with regard to their duration showed a sharp peak of very short turns (between 0 and 5 seconds), with populations of more than 100 turns, and a long tail which reaches durations of more than 60 seconds, with almost null populations. In particular, there are 1090 turns lasting less than 5 seconds (52.26% of the turns), amounting to a total of only 1994 seconds (9.87% of the useful signals); on the other hand, there are 33 turns lasting more than 60 seconds (1.58% of the turns), totalling 2845 seconds (14.09% of the useful signals). Duration data reflect the typical interaction schemeof interviews, with relatively short questions on the part of the interviewer, followed by long monologues on the part of the interviewee. The scheme will probably be different when dealing with conversations, since none of the speakers should be dominant.

The counts of SSEs, computed over the 42 broadcast interviews of CORLEC-EHU-1, are shown in Table 3, with the average number of SSEs per 100 words, taking into account that the number of words (excluding SSEs) was 64905.

Firstly it is remarkable the large number of overlaps: 1808 of 2090 useful turns. This is the main difference between human-human and human-computer interactions, since the latter rarely exhibit overlaps: users request something and wait for the system to answer; they do not really interact. However, in this case note that: (1) many of those overlaps correspond to completely overlapped turns, which are not included in the set of useful turns; and (2) sometimes the same turn included two overlaps (one at the beginning and other at the end of the turn). The large number of breathings is also remarkable: 3005, one every 20 words, almost 94% of human noises. Speakers need to take air periodically, so breathings act as a sort of technical pause, which might not appear at linguistic boundaries. Annotating breathings will make it possible to distinguish these phenomena from silent pauses, which accomplish more important tasks in spoken language, marking either linguistic boundaries or more complex SSEs such as self-corrections.

Table 3. Absolute counts of SSEs (#SSE) and average rate of SSEs per 100 words (SSE/100W), for the various categories of events annotated in CORLEC-EHU-1.

SSE category	SSE subcategory	#SSE	SSE/100W
Noises	Breathing	3005	4.62984
	Lip smack	161	0.248055
	Cough	40	0.0616285
	Generic background noise	530	0.816578
Acoustic events	Silent pause	1945	2.99669
	Filled pause /a/	25	0.0385178
	Filled pause /e/	800	1.23257
	Filled pause /m/	323	0.49765
	Filled pause /?/	616	0.949079
	Lengthening	3638	5.60512
Lexical events	Mispronunciation	1013	1.56074
	Cut-off word	212	0.326631
	Guttural affirmation	295	0.45451
	Acronym	36	0.0554657
	Foreign word	187	0.288113
Other	Speech overlap	1808	2.78561
	Continued turn	440	0.677914
	Unintelligible	71	0.109391
	Cut	9	0.0138664

Among the acoustic SSEs, the large number of lengthenings is remarkable: 3638, one every 18 words. On the one hand, this reflects excessive commitment on the part of the annotator, who tagged as lengthenings segments which were barely stretched out. However, it also reveals a lack of attention, since sometimes emphasized segments were erroneously taken as lengthenings. In any case, most lengthenings were correctly identified, showing —as some authors have previously indicated [5]— that they must be considered as resources of spontaneous speech, performing the same function as filled pauses, either as turn holders or as markers of self-corrections. Filled and silent pauses were more reliably marked, but gave lower numbers: 1764 and 1945, respectively. This means that 26.74% of acoustic events were silent pauses, 24.00% filled pauses and 49.52% lengthenings. It is remarkable that a similar distribution was previously found for a task-specific database in Spanish [2]. This means that —at least in terms of use of these resources— there would appear to be no difference between task-specific human-computer dialogues and generic human-human dialogues. Internal distribution of filled pauses also confirms previous results for a task-specific database in Spanish: the realization /e/ is dominant (800 instances, 45.35%), followed by /m/ (323 instances, 18.31%) and /a/ (25 instances, 1.42%). The unidentified realization /?/ (616 instances, 24.92%) must be kept apart, because it usually appears as a distortion of a vowel sound at the end of lengthenings or other filled pauses.

Among the lexical SSEs, the large number of *mispronunciations* is remarkable (1013, 58.12% of lexical events). This is partly due to the strict conditions imposed on the annotator, who marked any deviation from the standard or *canonical* pronunciation; for instance, ”Madri” instead of ”Madrid”, ”pasao” instead of ”pasado”, ”desir” instead of ”decir”, etc. These strict annotations should lead to more accurate acoustic models, though explicit modelling of pronunciation variants should also be applied, since acoustic models would not absorb them. Finally, the presence of guttural affirmations is significant (no guttural negations were found): 295 instances, revealing that these phenomena must be modelled at both the acoustic and lexical levels.

6. CONCLUSION

Absolute and relative counts of SSEs in a generic corpus of spontaneous speech show the importance of modelling these phenomena in speech recognition systems: for every 100 words —out of a total of 64905, distributed in 2090 utterances— we found an average of 4.94 human noises, 3.00 silent pauses, 2.72 filled pauses, 5.61 lengthenings and 2.34 lexical events. There was therefore a probability of 0.19 of finding any of such events at any word. Future research will include the annotation of the whole corpus CORLEC-EHU defined in this paper, at both the acoustic/lexical and syntactic/pragmatic levels. This will allow a comparative study to be made of the typologies and frequencies of all kinds of SSEs between task-specific and generic corpora of spontaneous speech. Finally, speech recognition experiments are being caried out on CORLEC-EHU-1, which will yield a reference rate for completely unrestricted spontaneous speech in Spanish.

7. REFERENCES

[1] A. Bonafonte, P. Aibar, N. Castell, E. Lleida, J.B. Mariño, E. Sanchís, and I. Torres, “Desarrollo de un sistema de diálogo oral en dominios restringidos,” in *Actas de las I Jornadas en Tecnología del Habla*, University of Sevilla, Spain, November 6-10 2000, Project website: http://gps-tsc.upc.es/veu/basurde.

[2] L. J. Rodríguez, I. Torres, and A. Varona, “Annotation and analysis of disfluencies in a spontaneous speech corpus in Spanish,” in *Proceedings of the Workshop on Disfluency in Spontaneous Speech*, University of Edinburgh, Scotland, August 29-31 2001, pp. 1–4.

[3] L. J. Rodríguez, I. Torres, and A. Varona, “Evaluation of sublexical and lexical models of acoustic disfluencies for spontaneous speech recognition in Spanish,” in *Proceedings of the European Conference on Speech Communications and Technology (EUROSPEECH)*, Aalborg, Denmark, September 2-7 2001.

[4] Almudena Ballester, Carmen Santamaría, and Francisco A. Marcos-Marín, “Transcription conventions used for the corpus of spoken contemporary Spanish,” *Literary and Linguistic Computing*, vol. 8, no. 4, pp. 283–292, 1993.

[5] Robert Eklund, “Prolongations: a dark horse in the disfluency stable,” in *Proceedings of the Workshop on Disfluency in Spontaneous Speech*, University of Edinburgh, Scotland, August 29-31 2001, pp. 5–8.