ProGmatica: a Prosodic and Pragmatic Database for European Portuguese

Daniela Braga, Luís Coelho, João P. Teixeira, Diamantino Freitas

Universidade da Coruña, ESEIG-Inst. Politécnico do Porto, Inst. Politécnico de Bragança, Fac. Engenharia da Univ. Porto

dbraga@udc.es, luiscoelho@eseig.ipp.pt, joaopt@ipb.pt; dfreitas@fe.up.pt

Abstract

In this work, a spontaneous speech corpus of broadcasted television material in European Portuguese (EP) is presented. We decided to name it ProGmatica as it is meant to combine prosody information under a pragmatic framework. Our purpose is to analyse, describe and predict the prosodic patterns that are involved in speech acts and discourse events. It is also our goal to relate both prosody and pragmatics to emotion, style and attitude. In future developments, we intend, by this way, to provide EP TTS systems with pragmatic and emotional dimensions.

From the whole recorded material we selected, extracted and saved prototypical speech acts with the help of speech analysis tools. We have a multi-speaker corpus, where linguistic, paralinguistic and extra linguistic information are labelled and related to each other.

The paper is organized as follows. In section one, a brief state-of-the-art for the available EP corpora containing prosodic information is presented. In section two, we explain the pragmatic criteria used to structure this database. Then, we describe how the speech signal was labelled and which information layers were considered. In section three, we propose a prosodic prediction model to be applied to each speech act in future. In section four, some of the main problems we went through are discussed and future work is presented.

1. INTRODUCTION

In recent developments on speech processing the expressions “attitudinal prosody”, “emotional speech” and “affective computing” became more and more common. Once the segmental quality of speech synthesis systems is studied and more mature models are implemented the next step is the development of the suprasegmental level in order to reach the so desired naturalness of synthetic voices. This upper level is not only defined by linguistic parameters (lexical, semantic and syntactic) but also by prosodic and pragmatic parameters. This way, the prosodic dimension can inform about the speaker’s pragmatic/communicative intention, about his degree of certainty or uncertainty about the propositional content of his message and even about the degree of familiarity that the speaker has with the listener. However, it seems that there is no specific European Portuguese database available to the scientific community addressing the study of the prosodic/pragmatic interface.

In this work, it is our aim to present a prosodic database that we decided to name ProGmatica, which was developed bearing in mind pragmatic criteria. This database is a resource for the study, description and future prediction of prosodic patterns that define speech acts and discursive events related to emotion and attitude.

It is also important to emphasize the lack of resources existing for the Portuguese language so far, when for instance compared to the thousands of recording hours of speech available for the Japanese language[1].

For European Portuguese language, the most relevant projects on speech collection are the following:

1. *Corpus de Referência do Português Contemporâneo* (Nascimento, 2000) developed in 1997. This is mostly a written corpus (with 77.3 million words and about 1.7 million words for the spoken variety). This corpus includes documents published since the beginning of the XIX century until nowadays and several varieties of the Portuguese language around the world are included;

2. *Português falado, variedades geográficas e sociais* (University of Lisbon, 1997). It is composed by 4 CDs with 83 Portuguese language recordings in formal and informal style, collected during the last 25 years in Portugal, Brazil, Portuguese speaking countries in Africa, Macau and Timor;

3. *FEUP-IPB Database* (Teixeira, 2001). This corpus has prosodic labelling at phrase and accent levels, and it is composed by 100 minutes of high quality speech produced by a professional speaker. The text material was obtained from newspaper articles. The recordings were made in a studio with professional equipment and annotated at phoneme, word and phrase levels. F0 information was also included;

4. *REDIP - Rede de Difusão Internacional do Português: rádio, televisão e imprensa* (initiated on 2000), developed by Instituto de Linguística Teórica e Computacional, it is composed by audio and video recording of radio and television shows (Ramilo, 2002);

6. *Multimedia Prosodic Atlas for the Romance Languages* (AMPER) is a prosodic corpus developed by the University of Aveiro (official web page of the project http://www.ii.ua.pt/cidlc/gcl/AMPER-POR.htm/ in 01-03-2006) on the scope of a European project. The main goal of this project is to analyse the correlation of the word accent and its syntactic position in the sentence in Romance languages;

7. C-ORAL-ROM is a European project that was started in 2000 (http://lablita.dit.unifi.it/coralrom/ in 06-03-2006). It addresses the study of the prosodic and syntactic levels and has comparable recordings of spontaneous speech in four different Romance languages (Italian, French, Portuguese and Spanish). The Portuguese part of the project is described in Nascimento et al. (2002). Although this resource is very important, it is commercially

[1] We can mention for example the JSP/ CREST ESP Project, composed by one thousand hours of natural daily conversational speech. in http://feast.atr.jp.

available at a high price and it still has some points that need to be developed in order to allow an immediate applicability to speech processing.
The Humaine Portal (http://emotion-research.net/ in 04-03-2006), dedicated to research on emotions and human-machine interaction, has several references to emotional databases but none of them addresses the Portuguese language.
This way, considering what is documented in scientific literature, none of the above databases seems to be completely oriented to the pragmatic dimension of speech. Therefore, we believe that ProGmatica is a resource that can highlight the correlation between the prosodic and the pragmatic events of speech communication.

2. PROGMATICA

ProGmatica is a spontaneous speech corpus containing different types of speech acts[2] selected from natural verbal interactions[3]. These speech acts were recorded from television programs where spontaneous speech material could be found, such as interviews, political debates and informal conversations performed in reality shows and soap operas. Although most of the studies done on speech are based on read speech corpora, it is our belief that spontaneous speech is the best source to observe the interaction between pragmatic purposes and prosodic events. Besides, as stated by Campbell (2000), read speech corpora are fake and artificial and are not recommend whenever the goal is studying speech from a communicative point of view.
The speech material was analysed and some utterances were classified according to the speech acts considered and described in section 2.1. As far as possible, we tried to select prototypical examples of each speech act. Syntactically speaking, we have preferred simple and small sentences in order to prevent superposition of different speech acts. Two main dimensions were considered when analysing a speech act: a segmental level, where glottal, phonetic and morphosyntactic dimensions were included; and a supra-segmental level, where prosody, emotion and style are considered.
ProGmatica is presently composed by 20 hours of television material recorded between 2003 and 2005 and later converted to digital format. From that material we had extracted ten prototypical utterances for each speech act. Labelling was performed with the help of PRAAT software (http://www.fon.hum.uva.nl/praat/).

[2] The term speech act was first used by Austin (1962) and later developed by Searle (1969). It has ever since been largely used in Discourse Analysis and Pragmatics, but recent contributions on this subject are scarce. We consider a speech act as an utterance whose limits are not necessarily coincident with the limits of a sentence. A speech act is also an utterance that is produced within a certain verbal interaction and that involves a certain communicative purpose .
[3] The term *verbal interaction* is defined by Kerbrat-Orecchioni (1990) under the French Discourse Analysis framework of Interactionism. It means any communicative exchange performed by at least two participants that are put face to face and that keep sending linguistic and paralinguistic information (prosodic, kinetic and proxemic signs) to each other. These signs assure their mutual influence, their mutual engagement and the management of their communicative exchange.

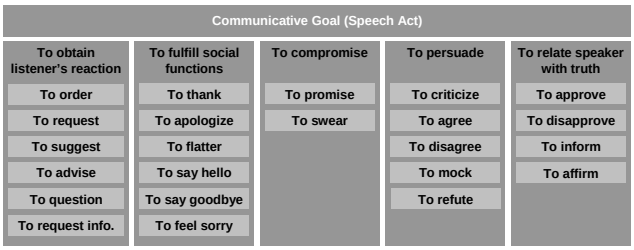

Figure 1: Pragmatic structure of ProGmatica in terms of speech acts

2.1. DATABASE STRUCTURE

ProGmatica is organized according to a revision of the Searle's (1969) well known speech acts typology. To this typology, we added what we can call the argumentative speech acts as suggested by Kerbrat-Orecchioni (1996). The author shows that speech acts expressing approval, disapproval, refutation or mockery can determine the speaker's prominence in the verbal interaction.
We have then expanded the former Searle's typology by including some more speech acts, such as agreement, disagreement, criticism and mockery, as shown in figure 1. Some speech acts were subdivided into categories, such as questions, whenever their syntactic and prosodic structure justified the distinction. At this stage, no declarative speech acts were considered because of their minor importance for TTS systems.
Five speech act communicative goals were considered in the collection of ProGmatica:
1. To obtain a listener's reaction. This category basically includes Searle's directive acts in which statements attempt to make the other person's actions fit the propositional content. With this purpose we consider orders, several types of requests, suggestions, advices and different types of questions.
2. To fulfil social functions. This category is a redefinition of Searle's expressive acts, since we think that these speech acts are more socially oriented than expressive, since the "sincerity condition" claimed by Searle doesn't have always to occur. When we apologize or feel sorry, we don't have to really mean it.
3. To persuade. This purpose was defined by Ducrot and Anscombre (1997). To these authors every speech act is argumentative because language is itself argumentative. Nothing that we say is meaningless in terms of influencing the other in some way. However, in our classification, we had a more narrow definition of this purpose, only considering those acts whose persuasive goal was more defined. Lexical items and their semantic meanings are decisive in this task.
4. To relate the speaker's propositional content with the truth. These statements may be judged true or false because they aim to describe a state of affairs in the world.
5. To compromise. In this category, commissive acts were considered as they were defined by Searle.

2.2. SIGNAL ANNOTATION

The recorded audio data is already a very valuable resource but it can be much more useful if a synchronized parallel annotation covering several levels is given. In the ProGmatica annotation the following informational levels were considered:

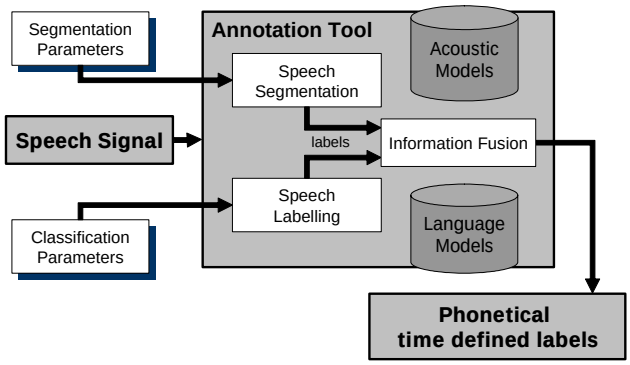

Figure 2: Annotation tool architecture.

1. *Narrow phonetic transcription*. The lowest annotation level is the phone level. Traditionally, the task of phonetic labelling is manually performed by a linguist. But in our case it was done automatically thanks to a tool developed by Coelho (2005). In figure 2, a diagram is displayed showing the working process of the tool. By using only acoustic data, this tool generates temporal boundaries that indicate phone start and phone end. For each time interval existing between boundaries it also classifies the phone type according to a previously defined lexicon. The lexicon is composed by 38 symbols representing the European Portuguese phonemes and 4 extra symbols for silence, aspiration, stop and stress. For phonetic annotation SAMPA transcription system was used. The language models and phonological rules were also adapted for European Portuguese in order to optimize the annotation performance. This tool gives an average error of 10% in boundary matching for 20ms error intervals and 5% on phone classification. When the speech samples have several voices or background noise some corrections may be needed. But in favourable acoustic conditions the error rate is close to the one attained by humans.
2. *Simple orthographical transcription*. In this level, a plain text conversion of the speech utterances was done for better handling of the tracks.
3. *Morphosyntactic annotation*. In this level, the words have been classified in terms of morphological category (noun, verb, adjective, pronoun, adverb, preposition, discursive marker and interjection).
4. *Prosodic labelling* with tonic accent (marked in the phonetic level before the tonic syllable), phrasing and focus. This information is correlated with the word level.
5. *Emotion annotation*. The emotion is classified according to seven types (Mozziconacci, 2002): neutral, joy, sadness, fear, indignation, surprise, anger. So far, we have considered these emotions because they are more stereotypical and thus more useful for speech processing subjects. The classification was subjectively performed by a linguist. For this task, both the semantic information of the utterance and the acoustic signal analysis were relevant. This was not an easy task, because in many cases there is more than one emotion transmitted. In the future, a perceptive evaluation performed by multiple listeners is foreseen in order to validate the initial classification.
6. *Style***.** In this level, two parameters were considered: the speakers' closeness and the linguistic politeness. The relationship parameter was classified in terms of [+formal], [+/-formal], [+/-informal], [+informal. The politeness parameter was classified in terms of [+polite]/[-polite].

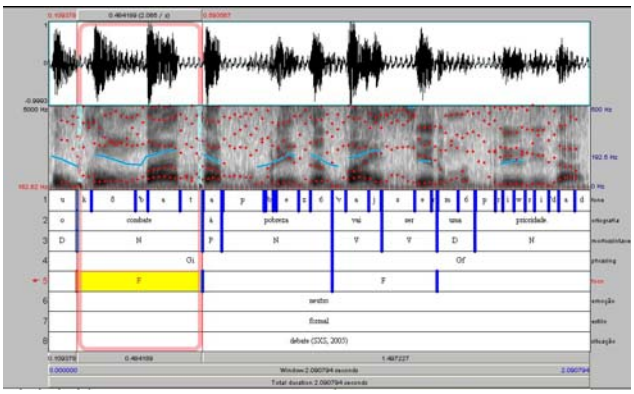

Figure 3: Example of annotation of a promise: "O combate à pobreza vai ser uma prioridade"(*The battle against poverty will be a priority*).

7. *Communicative situation*. This is the place where contextual information is given, such as broadcasting date, program from which the sample was extracted, speaker and verbal interaction type (debate, interview, conversation, etc).

In figure 3, an example of the described multi-level annotation is displayed. The annotation was performed using PRAAT software.

2.3. ACOUSTIC PARAMETERS EXTRACTION

The prosodic relevant acoustic parameters considered are the fundamental frequency (F0), segmental durations and, with less importance, intensity.

The durations and time intervals between begin and end of segments are automatically obtained from the annotation files. For large acoustic segments the duration is the sum of the phones' durations. The duration of the pauses between sentences are also considered.

For each phone the average intensity was calculated and registered.

3. PROSODIC PREDICTION

For each speech act it will be generated different models that allow synthesising the prosodic parameters from the text or from other linguistic mark-up. These models will be adapted from the segmental duration model and the F0 contour model developed by Teixeira (2004).

The duration model, already described by Teixeira (2004), was created with the help of an artificial neural network (ANN) with fixed structure and fixed features for the input vector. For each acoustic segment, feature vectors were prepared and the ANN was trained.

In what concerns fundamental frequency, the Fujisaki parameters (Fujisaki, 1997) were chosen after considering other options, like Tobi (Silverman, 1990), Intsint (Hirst, 1998) and the Tilt model (Taylor, 2000). This option was mainly due to the precise nature of the mathematical model and to some previous experience reported by Teixeira (2004). This way, two sets of parameters, phrase and accent, must be estimated. An ANN was also used and again different models for each speech act were adapted.

The feature vectors are extracted from the text.

In order to validate the calculated parameters some tests were performed. In figure 4 an example is shown of a F0 curve coded with Fujisaki parameters.

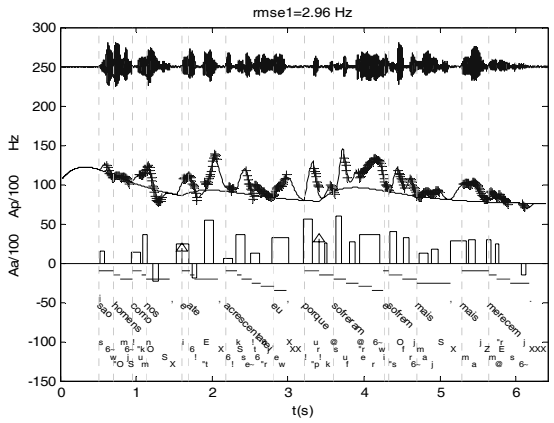

Figure 4: F0 contour codification example

From top to bottom it is represented: the speech signal, the F0 curve measured (with a plus '+' sign), the F0 curve produced by the model (Fb^4 + phrase components and Fb + phrase components + accent components), phrase commands and accent commands, syllable sequence aligned with speech signal, phrasal marks, orthographic transcription, phoneme sequence aligned with speech signal. The similarities between the F0 contour produced by the model and the original signal are obvious. In this special case the average quadratic error is 2.96 Hz.

4. CONCLUSION

In this article, it was described a spontaneous speech database for European Portuguese. This corpus, named ProGmatica, is structured in speech acts and labelled considering several linguistic levels (phoneme, orthography, morphosyntactic), prosodic information (accent, f0 e duration) and pragmatic events (emotion, style and communicative situation).

It was our objective to collect a sufficient number of utterance samples that can enable a consistent description of the prosodic events that can define a speech act. An ANN based model will be used for predicting utterances' durations. The Fujisaki model will be used for predicting speech acts' prosody produced in spontaneous speech.

The major difficulty found during the performance of this work was the correct classification of each utterance due to a continuous overlap of the communicative objectives. We sometimes are forced to eliminate good samples of speech acts because of bad acoustic backgrounds, such as music or voice overlaps.

Another problem was the classification of the common non-verbal signs such as laugh, grunts, interjections, aspirations that can provide communicative information and emotional content (Campbell, 2005).

As future work, we intend to collect more spontaneous speech utterances in which linguistic politeness, as well certainty/uncertainty and doubt dimensions can be better studied and characterized.

We hope that this work can enlighten a bit more not only human interaction but also the way we build man-machine interaction systems.

[4] Fb represents a reference fundamental frequency value in the mathematical description of the model.

5. REFERENCES

Austin, J. (1962) *How to do things with words*. Oxford: Oxford University Press.

Campbell, N. (2000) Databases of Emotional Speech. In. *Proceedings of ISCA ITRW on Speech and Emotion : Developing a Conceptual Framework*, Newcastle, N. Ireland, pp. 34-39.

Campbell, N. (2005) No Laughing Matter. In, *Proceedings of Eurospeech 2005*. Lisboa, Portugal.

Coelho, L. (2005) *Etiquetagem Automática de Sinais de Fala*. Dissertação de Mestrado. Faculdade de Engenharia da Universidade do Porto.

Ducrot, O.; Anscombre, J. C. (1997) *L'argumentation dans la langue*. Liège : Mardaga.

Fujisaki, H. (1997) Prosody, models and spontaneous speech. In, Sagisaka, Y., Campbell, N., Higushi, N., *Computing Prosody*. Springer-Verlag: New York, inc.pp 27-42.

Hirst, D. and Di Cristo, A. (1998) *Intonation Systems – A Survey of Twenty Languages*. Cambridge: Cambridge University Press.

Kerbrat-Orecchioni, C. (1990, 2002) *L'énonciation. De la subjectivité dans le langage*. Paris: Armand Colin.

Kerbrat-Orecchioni, C. (1996). *La conversation*. Paris: Seuil.

Mozziconacci, S. (2002) Prosody and Emotions. In. *Proceedings of Speech Prosody 2002*. Aix-en-Provence, France.

Nascimento, M. F. Bacelar (2000) O corpus de referência do português contemporâneo e os projectos de investigação do Centro de Linguística da Universidade de Lisboa sobre variedades do português falado e escrito. In. Gärtner, E.; Hundt, C.(ogs.) *Estudos de Gramática Portuguesa I*. Frankfurt am Main: TFM, pp. 185-200.

Nascimento, M. F. Bacelar, Cresti, E., Moneglia, M., Sandoval, A., Veronis, J., Martin, P., Choucri, K., Mapelli, V., Falavigna, D., Cid, A., Bloom, C. (2002) The C-ORAL-ROM Project. New methods for spoken language archives in a multilingual romance corpus LREC in Rodrigues, M. C. and Suarez Araujo (a cura di), *Proceedings of Third International Conference on Language Resources and Evaluation*, Paris: ELRA, vol. 1, pp. 2-10

Ramilo, M. C.; Freitas, T. (2002) A linguística e a linguagem dos media em Portugal: descrição do projecto REDIP. In. *XIII Congreso Internacional da ALFAL*. San José, Costa Rica.

Searle, J. (1969) *Speech Acts. An essay in the philosophy of language*. Cambridge: Cambridge University Press.

Silverman, K. and Pierrehumbert, J. (1990) The Timing of Prenuclear High Accents in English. In. J. Kingston and M. Beckman (eds) *Papers in Laboratory Phonology I*. Cambridge Univ. Press: Cambridge UK, pp. 72-106.

Taylor, P. (2000) Analysis and Synthesis of Intonation using the Tilt Model. *Journal of the Acoustical Society of America*. vol 1073, pp. 1697-1714.

Teixeira, J. P et al. (2001) Phonetic Events from the Labeling the European Portuguese Database for Speech Synthesis, FEUP/IPB-DB. In, *Proceedings of Eurospeech 2001*. Aalborg, Denmark, pp.1707-1710.

Teixeira, J.P. (2004) *A Prosody Model to TTS Systems*. Dissertação de Doutoramento. Faculdade de Engenharia da Universidade do Porto.