INTEGRATING PROSODIC INFORMATION INTO A SPEECH RECOGNISER

Teresa López. Soto
Universidad de Sevilla

In the last decade there has been an increasing tendency to incorporate language engineering strategies into speech technology. This technique combines linguistic and mathematical information in different applications: machine translation, natural language processing, speech synthesis and automatic speech recognition (ASR). In the field of speech synthesis, this hybrid approach (linguistic and mathematical/statistical) has led to the design of efficient models for reproducing the acoustic features of natural language. However, the incorporation of language engineering strategies into ASR is only beginning. In this paper, we present a theoretical framework for the integration of linguistic information into an ASR system. The objective is to design a model which can detect the suprasegmental features of the speech input, mainly those related to the fundamental frequency (F0) that can clarify the functionality of pauses, intonation contour, and interruptions. This specification model has been designed in the framework of a dialogue system.

1. Introduction

ASR systems generate a speech hypothesis which shows an *n* similarity with the speech input. These ASR systems can be used in various applications. ASR is present in systems where the objective is to identify the individual who has uttered that piece of spoken language. In these systems the speech recogniser has to identify the general acoustic features associated with the speech input. Other applications integrate ASR into a natural language processing system, where the main objective is the extraction of the semantic structure. In these latter applications, the speech recogniser has to generate a literal transcription of the speech input.

The transcription of the words uttered by a speaker is not an easy task. Many different factors affect the recognition of the exact sequence: some are intrinsic to the speech signal and others can be said to be extrinsic.

Among the so-called intrinsic factors we can mention some of a phonological nature, such as juncture and intonational function that have an acoustic correlate in the form of phonemic transition and F0 analysis. Other intrinsic factors are linguistic: sentence division, pauses, repetitions, anaphoric constructions, etc. These intrinsic factors characterize the speech signal, but are seriously distorted by other "external" factors, elements which go beyond the nature of the speech signal itself and which derive directly from environmental forces. Among these environmental factors we can mention two big groups: environmental factors which directly affect the spoken situation (echo, background noise, coughs, low voice, etc.) and electronical factors which affect the system performance: transducer (microphone), robustness, etc.

The sum of all these factors makes it difficult to construct an efficient model of speech recognition. Even in very robust and powerful ASR systems the error rate may unexpectedly prevent an accurate and reliable extraction of the meaning of the speech input.

To cope with recognition errors different techniques can be integrated into the NLP module (Heeman, 1998; López-Soto, 1999). However, this approach does not really improve the ASR performance and can be considered to be a mere *ad hoc* solution. A different approach consists in the incorporation of parsing techniques into the ASR system (word lattice parsing, Van Noord, 1998). This technique consists of the application of syntactic and morphological information to construct a word hypothesis and complements other acoustic and statistical information.

It seems that the incorporation of language engineering techniques is necessary to improve ASR performance, but it is not the only solution: acoustic and statistical information is definitively important to cope with factors that distort the signal (background noise, channel distortions, etc.). Besides, the use of lexical and syntactic information only is not enough to deal with a wide range of discoursal and contextual information implicit in any communicative situation. We believe that part of this information can be easily processed if we take into account prosodic information. In section 2 we describe some general characteristics of spoken language paying special attention to those features that can be formalized using prosodic information. In section 3 we present our specification model for the description of prosodic information into an ASR system.

2. Some features of spoken language

In this section we analyze the nature of pause and interruptions in spoken language.

2.1. Pause

There are two main kinds of pauses: empty pause or silence and filled pause or lexicalized pause, which accomplish different functions:
In the case of empty pauses or silence, the processing of this information can determine word and sentence division. This information is very important to determine the meaning of the sequence.
A filled pause takes place when the speaker is trying to adjust the speaking rate to the cognitive processes that are going on. Filled pauses have a very complex functionality in a dialogue situation:
They help the speaker keep a dialogue turn (“I mean”, “well”, “er”, “um”, “you see?”etc.). Filled pauses also help to keep the listener alert, preventing undesired interruptions.
They are also common to express emotional states, such us hesitation (“er”, “um”, etc.), anxiety, anger, surprise, etc. (“come on!”, “sure!”, “but excuse me!”, etc.). They are frequently used to express cognitive and mental states (very often we incorporate filled pauses when we are looking for the appropriate word or expression or when we are trying to recall information).
Filled pauses frequently allow the listener to predict what comes next, as it happens in the following example:

(1) A: “I very seldom go out for dinner, just once, er...”
B: “Once in a blue moon”

In (1) B interrupts A and finishes A’s discourse.

In dialogue systems, we can find two main approaches to deal with pauses:

In some systems, filled pauses are included in the lexicon.
Sometimes lexicalized pauses are considered unknown terms and do not go beyond the initial level of analysis.

In any of the cases above though, we miss the information that these sequences may have for a correct analysis and understanding of the speech signal.

The analysis of pauses in ASR systems has usually taken place in the acoustic module. The acoustic analysis of pauses will be more effective when they are lexicalized ("sure", "well", etc.), less so when they are not lexicalized ("um", "er", etc.). On the other hand, language modeling cannot be efficient to analyse filled pauses because they can appear at any position in the sentence.

So the question would be: Can filled pauses be analysed in a reliable way? We can state that only prosodic information can determine the occurrence of pauses in the speech signal. We can illustrate this with one example. In the following sentence, the term "well" is used as a filled pause and as an adverb. The meaning varies according to the pitch change associated with the word "well" when it is used as a filled pause.

"I don't like my steak well done" (but maybe I like it just done)
"I don't like my steak, well, done" (I like it rare)

A filled pause has no meaning in itself, its main function is to keep the dialogue turn. However, a filled pause may affect the general semantic structure of the speech sequence. Filled pauses occur when the speaking process is in a "stand-by" state, therefore articulatory features remain intact until discoursal and cognitive processes are kept equal. Our model is based on the analysis of F0 transitions and spectral deformation to detect filled pauses. Filled pauses usually occur when the tension of the vocal folds remains invariable under constant articulatory parameters and the F0 of voice stays practically constant but the spectrographic analysis is notably altered. The shape of the vocal tract does not vary under constant articulatory parameters, nor does the spectral slope deformation. With this system we can detect when a filled pause is introduced in the flow of the speech signal.

2.2. Interruptions

Interruptions and other speech disfluencies (as described in Heeman, 1998) also play an important role for the interpretation of the discourse. In this section we will analyze the characteristics and functions of interruptions in dialogue systems.

Interruptions may have the following functions:

They can be used to change the topic of the dialogue
Very often they confirm a message
Interruptions can repair or emphasise a message
They are very frequently used to obtain information

Interruptions can be defined as those situations where a person intends to continue speaking but is forced by another person to stop speaking, at least temporarily, or the continuity or regularity of speech is broken for any other reason.

Interruptions can be classified into two general groups:

Competitiveness: The listener interrupts the flow of speech to express urgency, degree of importance of the topic, interest. Competitiveness also takes place when the listener wants to express strong opinion or disagreement. The flow of speech is diverted after the interruption.

Cooperation: The listener interrupts the flow of speech to confirm or strengthen the speaker’s point of view. The flow of the speech remains constant after the interruption.

The prosodic features of interruptions of class (1) reflect the necessity and urgency of the listener to receive information, as well as the necessity of including the interruption in the discourse so that it shows a high degree of relevance. On the other hand, interruptions of class (2) are originated when the person who produces the interruption shows a greater degree of confidence and certainty in the dialogue.

Interruptions must be analysed after considering the general intonational structure of the sequence, the amplitude of the waveform and the speech rate. Interruptions of class (1) are usually associated with unexpected pitch changes (higher pitch levels), show a greater amplitude and the speech rate accelerates. In less traumatic interruptions (class 2) the pitch level is usually low, because the person who is interrupting merely expresses his/her agreement. However, the amplitude increases to reflect emphasis. The speech rate remains unaltered.

3. A model for the specification of prosodic information based on predictions

In the previous sections we have described the functionality of pauses and interruptions in a dialogue system. There are obviously several other phenomena that can be analysed in the general context of discourse in order to discover the information they supply to the general meaning structure: repetitions, false starts, etc. However, in our present study we are only concerned with the detection and analysis of pauses and interruptions for a Spanish corpus.

The prosodic structure associated with a speech signal has a very specific function. This information is as relevant as the syntactic and semantic information. The main obstacle we have to face is that intonational features are strongly associated with particular communicative contexts. For this reason, we propose a specification model which studies the functionality of prosody in a dialogue system. In our system the user can send messages to the machine and there exists a limited interchange of questions and answers in the negotiation of meaning (Álvarez et al., 1997). That is, we are dealing with a very restricted domain where we can find a limited number of linguistic constructions, and therefore, of prosodic structures. In the next sections we explain how the corpus was designed.

3.1. Corpus labeling

The model that we present here is based on human predictions (Tamoto et al. 1999). This model is based on the evaluation by human labelers, who identify and select the prosodic structures associated with the corpus. In this section we describe the corpus labeling process.

To determine the function of pauses and interruptions in the corpus, we have first analyzed the intonational structure in the domain taking into account pitch range, amplitude, F0 values, frequency, spectral slope, duration and speech rate. The result has been a corpus labeled by trained native speakers. The model has followed several stages:

The corpus is recorded and transcribed.
The transcriptions are then labeled by trained native speakers. The objective is to select the discourse structures which characterize the corpus following the model proposed by Carletta et al. (1997). That is, each complete discoursal sequence is labeled according to the communicative function it shows in the general context of the dialogue.
The recording is filtered to extract the F0. This new version is passed onto the labelers which determine which intonational structure corresponds to each discoursal label according to their knowledge of the language.

3.2. Prosodic labeling

Once discoursal labels had been assigned, the corpus was labeled once more, this time to show intonational structures. Prosodic labeling was done in two states:

The first labeling is done manually, after the labelers have listened to the recording from which the F0 has been extracted. These labels follow the notation described by Silverman et al., (1992).
The second labeling is done automatically using Waves TM. The

representation of the F0 of the speech signal is obtained and labeled following the same notation system.

Once the labeling process is over, important conclusions can be taken after contrasting the two labeled corpora.

4. Conclusion

The model that we present is based on the predictions made by native speakers assigning intonational labels to a spoken corpus. After comparing the two corpora (one manually labeled, the other one automatically labeled) we can get to more reliable conclusions. The corpus that we have obtained includes information that is necessary in order to process prosodic information, and, more specifically, in order to detect pauses and interruptions. This information is based on an analysis of the f0. Other information can be extracted from frequency, spectral slope, duration and speech rate. With these results we are now working on the development of an automatic model to analyze and process prosodic information in order to detect pauses and interruptions in a spoken corpus for a dialogue system.

Bibliographical references

Álvarez, J., D. Tapias, C. Crespo, I. Cortázar y F. Martínez. 1997. "Development and Evaluation of the ATOS Spontaneous Speech Conversational System". *International Conference on Acoustics, Speech and Signal Processing (ICASSP'97),* 1139-1142.

Carletta, J., S. Isard, G. Doherty-Sneddon, A. Isard, J. C. Kowtko y A. H. Anderson. 1997. "The Reliability of a Dialogue Structure Coding Scheme". *Computational Linguistics*, **23**(1):13-31.

Heeman, P. A. 1997. *Speech Repairs, Intonational Boundaries and Discourse Markers: Modelling Speaker's Utterances in Spoken Dialogue*. PhD dissertation. Dpt. of Computer Science, University of Rochester, N.Y.

López-Soto, M. T. 1999. *Estrategias de análisis gramatical y semántico para un sistema dirigido por voz*. PhD dissertation. Dpto. de Lengua Inglesa, Universidad de Sevilla, Sevilla.

Silverman, K., M. Bechman, J. Pitrelli, M. Osterdorf, C. Wightman, P. Price, J. Pierrehumbert, J. Hirschberg. 1992. "ToBI: a Standard for Labelling English Prosody". *Proceedings, 2nd International Conference on Spoken Language Processing,* Banff/Canada:867-870

Tamoto, M., M. Kawamori y T. Kawabata. 1999. "Integrating Prosodic Features in Dialogue Understanding". *Proc. of Eurospeech-99*. (CD-ROM)

Van Noord, G., G. Bouma, R. Koeling y M. Nederhof. 1998. "Robust Grammatical Analysis for Spoken Dialogue Systems". *Natural Language Engineering* **1**(1):1-48.