IMPROVEMENT ON CONNECTED NUMBERS RECOGNITION USING PROSODIC INFORMATION

Eduardo López (*), Javier Caminero, Ismael Cortázar, Luis Hernández (*)

Speech Technology Group, Telefónica Investigación y Desarrollo

Emilio Vargas 6, E-28043 Madrid, Spain

email: {eduardo, jcam}@craso.tid.es

ABSTRACT

In this paper we propose a strategy to improve the performance of a connected number recognition system in Spanish using prosodic information. Prosodic information is included as the detection of pitch movements between what some studies of intonation in Spanish called 'melodic units'. The basic linguistic background of our approach together with the specific strategies to detect and correct ambiguities and recognition errors are discussed. Experimental results show a 16% of reduction in recognition errors for our state-of-the-art connected number recognizer, and the possibility to solve ambiguities unable to be considered by the recognizer.

1. INTRODUCTION

The use of prosody in Automatic Speech Recognition (ASR) systems has been proposed for over thirty years, but there was only limited research results up until the second half of the 1980's. Prosodic characteristics of speech include a range of segmental and supra-segmental features including intonation (F0, pitch), intensity and temporal features (duration, pauses) to convey cues to syntax, semantics, and pragmatics, and different emotions and attitudes of the speaker. Although it is clear the important role of prosody in terms of speech production and perception, its use in the design of ASR systems is still a challenge for the speech recognition community, and until now there is no general and effective procedures to combine prosodic information with other acoustic features to significantly increase recognition performance [1], [2], [3].

In this paper, trying to contribute to the research on how to include prosodic information for the improvement of state of the art speech recognition systems, we focused on two major topics:

• To identify acoustic events related to basic prosodic units that could be useful to improve an automatic connected number recognition system in Spanish working in real applications over the telephone line. This is a complex state-of-the-art continuous speech recognition task where prosody shows a great potential to detect and solve many recognition errors and ambiguities. Moreover well-known prosodic events can be considered as an additional source of knowledge that in the particular task of connected numbers recognition can be easily taken into account to improve the acoustic information generally used in speech recognition systems. Furthermore the resulting methodology for this task can be extended to other connected speech systems.

• To propose particular algorithms to efficiently combine and process acoustic cues to prosodic information in a connected number speech recognizer. In particular we will consider two strategies directed to detect and solve two major difficulties in this kind of systems: ambiguities (a same utterance can be interpreted in different ways as different natural numbers, for example *dos cientos diez* as 200-10 or 210) and recognition errors. In our work, looking for simplicity and low computational cost we will propose processing algorithms working as post-processing steps of a baseline recognizer. We will also briefly describe the signal processing tool added to the ASR system: a robust pitch detector, that has been designed to work in a telephonic environment and with the same spectral (cepstral) characteristics than the recognizer for an easy integration with it.

In order to cover the previous two objectives the paper is organized as follows: Section 2 presents the theoretical prosodic background of our work through the discussion of classical intonation models for Spanish. Section 3 describes how we define some basic and simple acoustic cues related to the major prosodic events found in the pronunciation of natural numbers. Details of the ASR systems we use for natural numbers recognition and of the development of robust acoustic tools for a reliable measurement of prosodic features, specially for the estimation of pitch information, are presented in Section 4. Section 5 presents two strategies to improve the performance of the natural number recognizer in two different areas: ambiguity resolution and recognition errors correction. Experimental results using the Telefónica I+D natural number recognition system working on real tasks over the Spanish telephone network are given in Section 6.

2. BACKGROUND

In this paper we will use the term **intonation** from a 'phonetic' point of view [6]: *the evolution over time of the F0 contour of an utterance*. Therefore our work needs to start with a first approach to classical Spanish intonation models. Without trying to present a complete review of classical intonation models for Spanish, that can be found in [6] and [7], it is important to point out that due to the difficulty in representing different related phenomena in prosody, most of the proposed models are focused over the melodic curve. Different events over melodic curves have been studied for English and extended over other languages, however, obviously, there are noticeable differences in many cases. Most of the studies of intonation for Spanish, see the reviews in [6] or [7],

(*) E.T.S.I. Telecomunicación. Universidad Politécnica de Madrid. Spain.

are based on the work of Tomás Navarro Tomás [8] who observed intonation from the study and classification of melodic curves.

According to Navarro Tomás, we can describe intonation of Spanish as a hierarchical structure that applies over different-level domains or units, both local and global. Similar descriptions can be found for other languages, and several definitions of global and local units have been proposed [6]. The basic unit we will use in this work is what Navarro called 'melodic unit' (*unidad melódica*): the shortest part of a discourse with both proper meaning and melodic structure. This unit is similar to the **phonetic group** (part of an utterance between pauses) but melodic units may also be marked by changes in energy, duration or special movements in the melodic curve.

At this point we can say that our work on the use of prosodic information to improve an ASR system is concentrated around the melodic unit. Starting from the recognized hypothesis provided by the ASR system, we will identify melodic units. These melodic units will be classified in terms of their melodic curves. Then high-level linguistic knowledge and empirical knowledge, from the observation of melodic curves of a corpus containing task-dependent utterances, is used to derive procedures to detect and, in some cases, correct recognition errors. In some way this approach is similar to the use of a top-down generated pitch contour compared to the observed F0 contour for disambiguating recognition results proposed by Keikichi Hirose in [1], but in our work instead of using the whole pitch contour we will concentrate on its particular behaviour around the discontinuities between melodic units. Note also that other proposals to improve ASR systems using the F0 curve to detect word boundaries (see for example [3] [5]) are different from the one proposed here.

Then, just because our work is based on the recognition and classification of melodic groups, it is important to describe how we can recognize melodic units and its standard classification in Spanish.

• As we mentioned before, in many cases melodic units are separated by pauses, but in many others they are recognized only from changes in energy, duration and for a more or less noticeable jump o reset in the melodic curve. In our model, as we will detailed in the next section, melodic units are only detected looking for F0 movements from the last syllable of a unit to the first syllable of the next unit.

• Navarro considered a melodic unit as composed of three different parts: head, body and tail, that would roughly correspond to the British style division of tone groups into pre-head, head or body, and nucleus plus tail. The head of a melodic unit goes from the beginning of the group to the first stressed syllable, the body of the unit goes from the first stressed syllable to the last stressed one (excluded), and the tail of the group, called 'toneme' (*tonema*), from the last stressed syllable to the end of the group. From these parts of a melodic unit a general pattern of melodic curves in Spanish could be described as: an initial rising in the head (F0 rises between 2 or three semitones; 7 or 8 semitones in emotional intonation); an approximately sustained tonal level in the body; and a final movement of F0 in the tail, toneme, which is the most important part in terms of the classification of melodic units. Navarro described five different kinds of *tonemas* depending on the evolution of the melodic curve: falling (*cadencia*); half-falling (*semicadencia*); level (*suspensión*); half-rising (*semianticadencia*) and rising (*anticadencia*).

3. ACOUSTIC CUES TO MELODIC UNITS

The basic background presented in the previous Section has been applied to the particular case of spontaneously spoken natural numbers. Therefore apart from the linguistic knowledge we have observed several melodic curves for this particular case, and from these observations, and having in mind the linguistic background, we have derived a set of practical procedures to improve our natural number ASR system.

From a corpus of 519 utterances containing spontaneously pronounced natural numbers recorded in a real application through the telephone line, we have observed how a speaker reflects the grouping of numbers in the pitch contour of the different melodic groups. Generally natural numbers are pronounced by grouping numbers (that makes them easy to be remembered), and each group of numbers is pronounced as a single melodic unit. As we already said these melodic groups may be separated by pauses but not always.

Looking at the behaviour of the tail in the melodic groups or tonemes, we found that, as expected, melodic groups at the end of the utterance has a falling (*cadencia*) toneme. All the other tonemes of an utterance usually present rising (*anticadencia*) tonemes. However we also observed a few but important number of speakers that pronounce the sequence of groups of numbers as an enumeration and thus the tonemes corresponds to falling (*cadencia*) or half-falling (*semicadencia*) instead of rising (*anticadencia*). Figure 2 illustrates a typical intonation pattern during the pronunciation of a natural number: *veinticinco - tres cientos dos - ocho cientos noventa y seis* (twenty five - three hundred and two - eight hundred ninety six). In the figure the melodic curve represents the F0 movements in a normalized logaritmic scale (zero values corresponds to unvoiced sounds). As it can be seen in the Figure in this example all the melodic groups are separated by pauses and the first two groups have rising tonemes while the last one presents a falling toneme.

As a conclusion of our studies and analysis we realized that in the spontaneous pronunciation of natural numbers there are some cases where the presence of different melodic groups is mandatory for a proper understanding of it. For example, in the utterance of Figure 2, the labels at the bottom of the figure represents the recognition labels of our ASR system and a recognition error is present. In this case the third melodic unit *ocho cientos noventay seis* (eight hundred and ninety six) is recognized as *ocho ciento noventa y seis* (eight one hundred ninety six). But if the recognition result were correct *ocho* (eight) and *ciento* (one hundred) should necessary belong to different melodic units, and therefore an F0 discontinuity should appear between them, and as can be observed in the Figure there is no F0 discontinuity around the boundary proposed by the ASR system. Furthermore, the important point is that in this case the two possible melodic units (*ocho - ciento ...*) could be present without no pause in between, therefore only the pitch movement between the tail of a melodic unit and the head of the next unit could help to detect a melodic unit boundary. Therefore we designed and tested a procedure to detect and, in some cases, correct

recognition errors and ambiguities of our ASR system based on the evaluation of pitch movements around the boundaries of melodic units. This procedure is described in Section 5 after a brief review of the acoustic modules of our system in the next section.

4. ACOUSTIC MODULES

Two key aspects of our work are the characteristics of our natural number recognizer and of the pitch detector we used. The continuous natural number recognition system we used is based on semi-continuous HMM [4] with a Mel-cepstra parametrization using 18 features in three different codebooks: 8 cepstra, 8 delta-cepstra and the energy and delta-energy respectively. Originally this recognizer is based on sub-word units, but for this work we used syllable-like units to allow a segmentation in syllables needed to characterize the pitch movements between melodic units. Another important acoustic module is the always problematic pitch detector, specially in applications such as ours dealing with real telephonic speech. For compatibility with the recognizer and low computational cost we developed a reliable pitch detector working in the same cepstral domain than the recognizer. Starting from the cepstral estimation a linear weighting is applied to equalize the speech spectrum before a peak detection in the 63 to 500 Hz band. From the relative amplitudes of spectral peaks we obtain a confidence level and a first estimation of pitch which is used to estimate a possible range of pitch variability for the speaker. After that a pitch tracking algorithm combines all the available information to provide the final pitch estimation together with the voiced/unvoiced decision. To compare data from different speakers, resulting pitch values for an utterance are normalized using a linear transformation to have zero mean pitch value (across the utterance) and pitch variance one.

5. RECOGNITION IMPROVEMENT THROUGH STATISTICAL MODELING OF F0 MOVEMENTS

As we stated before, our procedure to deal with recognition errors and ambiguities is based on the detection of special pitch movements at the boundaries between specific melodic units. Then we have two major topics to discuss: in one hand, the specific units we considered, which are related to typical prosodic patterns during the pronunciation of ambiguous sentences or highly confusable strings. On the other hand, the statistical model we use to deal with the characterization of pitch movements taking into account speaker and context variability. These two major topics will be described in the next paragraphs for ambiguity resolution and error correction.

In both cases, ambiguities and recognition errors, F0 movements were characterized by obtaining the average pitch value for the last syllable of a melodic unit and the average pitch value for the first one in the next unit. Using these two values a transition from a falling (*cadencia*) or rising (*anticadencia*) toneme was detected by means of Linear Discriminant Analysis (LDA). To deal with the speaker and context variability LDA was designed using a training data base with automatic labelling of transition between melodic units, these labels were obtained by means of forced alignment using our natural number recognizer.

5.1. Ambiguity resolution

The use of intonation can solve some ambiguities in the recognized phonetic strings that cannot be solved by our baseline connected numbers recognition systems. In particular we studied ambiguity resolution for utterances containing the Spanish word *cientos* (hundred). The use of this word produces ambiguities in utterances like *tres cientos cinco* (three hundred five) that can be interpreted as 300-5 or 305, and only prosody can help to solve it. Then we try the use of our measure of F0 movements to detect typical rising (*anticadencia*) or falling (*cadencia*) movements between the two possible melodic units: the last syllable in 300 and the first syllable in 5. In the example, using the time boundary between 300 and 5 provided by the recognizer we measure the average pitch of the last syllable in 300 and of the first syllable in 5. These two values are combined and compared to a threshold to decide whether two melodic units (300-5) or only one (305) was pronounced.

5.2. Misrecognition of words belonging to different melodic groups

A second situation where prosodic information can help to improve the performance of a connected number recognizer is to detect recognition errors of some highly confusable words belonging to special melodic units. In this direction we studied three different cases:

- Misrecognition of words *cientos* and *ciento* ('hundred' - 'one hundred and'); for example: *tres cientos dos* (302) recognized as *tres ciento dos* (3-102)
- Confusions between endings *-enta* and *-enta-y* (-ty+digit and -ty); for example: *setenta-y-uno* and *setenta uno* (seventy one, seventy and one).
- Pair confusion of the word-pair *cien dos* (one hundred two) as *cientos* (hundred)

In all these cases we can detect recognition errors by looking for the presence of one or two melodic units following a statistical analysis based on LDA, similar to the one described for ambiguity resolution. As an example, Figure 1 represents average values of F0 for the last syllable in *enta* and *entay* versus the average value of F0 for the first syllable of the next melodic unit. In the Figure two clusters belonging to different F0 movements for *enta* and *enta-y* can be easily identified (values around zero correspond to utterances were our pitch detector can not find reliable pitch values; these cases are not post-processed). It can also be observed that the unit *enta* has a high F0 value, corresponding to a rising (*anticadencia*) toneme, which is not the case in *enta-y*.

6. EXPERIMENTAL RESULTS AND CONCLUSIONS

For experimental results we used a recently designed database of connected spanish numbers recorded by Telefónica I+D (NUMLIM), and the connected numbers recognizer also developed by Telefónica I+D [4]. Both correctly recognized

sentences and errors were used trying to derive safe rules that could fix the errors but do not change those sentences well recognized without the use of prosodic information. The total number of files was 5232 and the total sentence error rate is 9.92% (519 errors).

After the application of the proposed procedures for ambiguity resolution and correction of misrecognition of words belonging to different melodic groups the sentence error is significantly reduced to 8.35%. A closer analysis of our results revealed that 90% of the recognition errors were detected, most of them belonging to the word *cientos*; the confusion *-enta* and *-enta-y* is not very common in our spanish connected numbers recognizer but has been proved successful in correcting ambiguities in numbers pronounced in catalan language where there are no orthographic distinction between the two words. In the future, we will try to extend this method to other cases and languages.

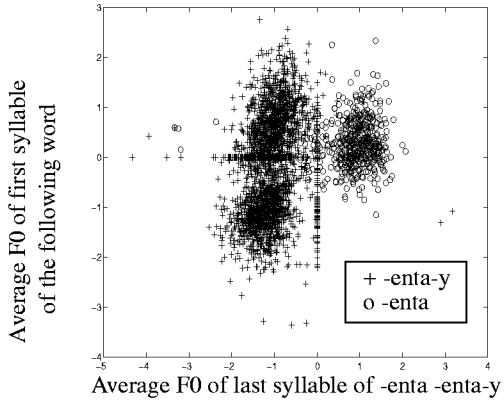

Figure 1: Average normalized pitch values for F0 movements in the *enta* and *enta-y* cases.

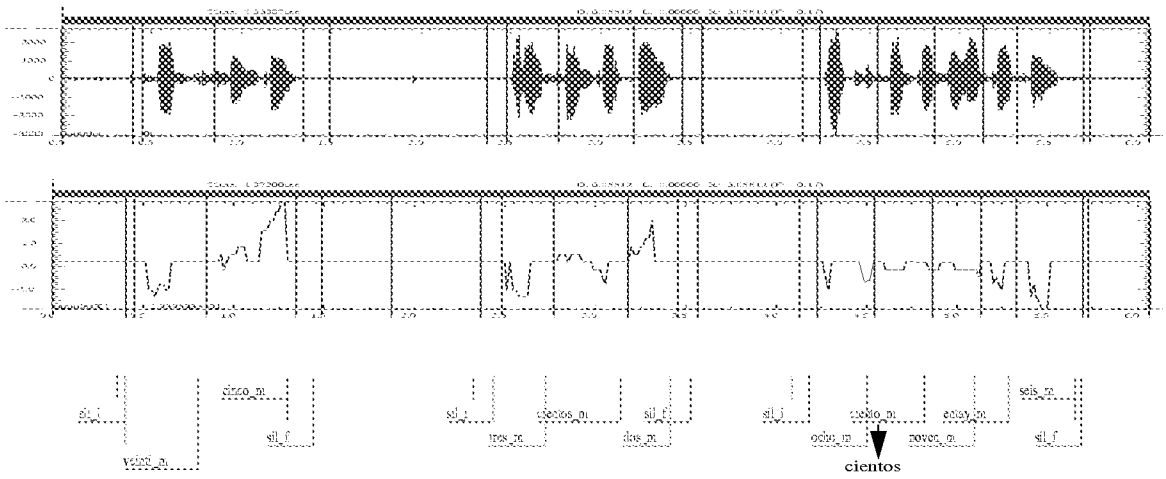

Figure 2: Speech wave form, pitch contour and recognition labels for the utterance of *veinticinco trescientos dos ocho cientos noventa y siete.*

Therefore we can conclude that the proposed characterization of F0 movements between melodic units is a useful way to improve a connected number recognition through the use of prosodic information. Both ambiguities and recognition errors can be detected and solved using this acoustic cue to prosody information. We are now considering the use of other acoustic characteristics such as duration and energy that could be useful to obtain more complete acoustic cues to prosodic information. Of course more general rules and cases should be studied to extend this methodology to more general ASR systems.

7. ACKNOWLEDGEMENTS

The authors from ETSIT - UPM are grateful to CICYT for their financial support under the contract no. TIC-96-0956-C04-03.

8. REFERENCES

[1] Y. Sagisaka, N. Campbell and N. Higuchi Eds, *Computing Prosody: Computational Models for Processing Spontaneous Speech*, Springer, New York, 1997.

[2] K. Bartkova and D. Jouvet, "Usefulness of Phonetic Parameters in a Rejection Procedure of an HMM Based Speech Recognition System", Proc. EuroSpeech-97, pp. 267-270.

[3] S. Rajendran and B. Yegnanarayana, "Word Boundary Hypothesization for Continuous Speech in Hindi Based on F0 Patterns", Speech Communication, vol. 18, 1996, pp. 21-46.

[4] C. de la Torre, et al., "Recognition of Spontaneously Spoken Connected Numbers in Spanish over the Telephone Line", Proc. Eurospeech-95, pp. 2123-2126.

[5] R. Ramana and J. Srichand, "Word Boundary Detection Using Pitch Variations", Proc. ICSLP 1996 Philadelphia, pp. 813-816.

[6] J.M. Garrido Almiñana, Modelling *Spanish Intonation for Text-to-Speech Applications*, Ph D Thesis, Universidad Autónoma de Barcelona, 1996.

[7] M.L. García-Lecumberry, *Intonational Signalling of Information Structure in English and Spanish: A Compartive Study*, Ph D Thesis, University of London, 1995.

[8] T. Navarro Tomás, *Manual de Entonación Española, Guadarrama*, Madrid, 1948