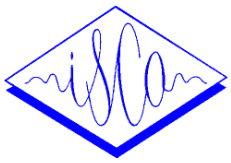

Second ESCA/IEEE Workshop
on Speech Synthesis
Mohonk, New Paltz, NY, USA
September 12-15, 1994

"Coding fundamental frequency patterns for multi-lingual synthesis with INTSINT in the MULTEXT Project"

Daniel Hirst, Nancy Ide, Jean Véronis
LABORATOIRE PAROLE ET LANGAGE
CNRS & Université de Provence
Avenue Robert Schuman
Aix-en-Provence Cedex 1 (France)

e-mail: hirst@fraix11.univ-aix.fr

1. MULTEXT Project Overview

MULTEXT (Multilingual Text Tools and Corpora) is the largest project funded under the European Commission's LRE (Linguistic Research and Engineering) Program. Intended to contribute to the development of generally usable software tools to manipulate and analyse multi-lingual text and speech, and to annotate multi-lingual text and speech corpora with structural and linguistic markup, it will attempt to establish conventions for the encoding of such corpora, building on and contributing to the preliminary recommendations of the relevant international and European standardization initiatives. MULTEXT will also work towards establishing a set of guidelines for linguistic software development, which will be widely published in order to enable future development by others. The project consortium, consisting of eight academic and research institutions and six major European industrial partners, is committed to making its results, namely corpus, tools, specifications and accompanying documentation, freely and publicly available.

At the outset of the project, the consortium will (in cooperation with the European Advisory Group on Language Engineering Standards, EAGLES) undertake to analyse, test and extend the SGML-based recommendations of the Text Encoding Initiative (TEI) on real-size data, and gradually develop encoding conventions specifically suited to multi-lingual corpora and the needs of NL and Speech corpus-based research (Ide & Véronis 1993a, 1993b).

By using the emerging software tools, the consortium plans to produce a substantial annotated multilingual corpus, including parallel texts and spoken data, in six EC languages (English, French, Spanish, German, Italian and Dutch). The entire corpus will be marked for gross logical and structural features; subsets of the corpus will be marked and hand-validated for sentence and sub-sentence features, part of speech, alignment of parallel texts, and prosody. All markup will have to comply to the TEI-based corpus encoding conventions established within the project. The corpus will also serve as a testbed for the project tools and a resource for future tool development and evaluation.

2. Prosody as the intersection of NL and Speech: Prosody

The place of speech in the overall project is intended to explore the possibilities of integrating NL and speech processing by attempting to harmonize tools and methods from both areas. MULTEXT will explore the possibilities for integration of NLP and speech by attempting to harmonize tools and methods from both areas. MULTEXT will pay special attention to phenomena at the intersection of the two domains, in particular prosody, whose supra-segmental nature invites research into the complex relationships it holds with morphology and syntax.

Research in this area is very important for applications such as high quality text-to-speech synthesis. High quality text-to-speech systems are needed for a wide range of applications: access to files by telephone, aid to handicapped persons, talking books, multi-media man-machine communication, etc. However, two main problems hamper their broad diffusion:

- most systems are oriented towards English synthesis, or a few major EU languages;
- most systems suffer from a lack of synthetic voice quality, in particular a low quality of the generated prosody.

MULTEXT will contribute to the field by providing prosody tools that will automatically derive a symbolic representation of the intonation from the speech signal. An automatic modeling system is highly desirable for a number of reasons. First, an efficient tool will be extremely useful for collecting data that can be used to improve both speech synthesis and speech recognition. Second, a symbolic coding will enable vastly reducing the amount of data stored (a few symbols instead of the complete acoustic curve). Most importantly, such a tool will be extremely valuable for testing models of intonation and their relationship with morphology and syntax, as well as examining variability in prosodic parameters across individuals and languages.
The aim of the prosody tools developped within MULTEXT is to provide the means for deriving a symbolic representation of the intonation of a spoken text directly from the speech signal. This symbolic representation should contain sufficient information so that a synthetic version of the intonation pattern can be generated from it without too great a loss of information.
Within the project the tools developped will be tested on a subset of the multilingual speech corpus EUROM-1 (developped within the ESPRIT-2 project SAM). The material to be annotated consists of 40 short passages of 5 thematically connected sentences, each recorded by several native speakers, for all six MULTEXT languages (the recordings for Spanish have recently been added to the original EUROM-1 recordings by the SAM-A project).
MULTEXT will enhance the EUROM-1 corpus with markup for prosody, segmentation, and POS. The prosody markup will consist of two levels: F0 curve modeling and symbolic coding. This markup will be accomplished using the prosody tools, and hand-validated. The orthographic transcriptions will be marked for level 2 and POS, and hand-validated.
The marking of alignment for several levels of language analysis (signal, phonemic transcription, orthographic transcription, F0 modeling, intonation symbolic coding, word boundaries, POS tagging) will provide a robust test of the TEI alignment mechanisms as well as a challenging problem for the corpus retrieval tools.

3. Prosodic labelling and notation.

A number of prosodic labelling systems exist at present but none of these can be considered an accepted standard for cross-language prosodic notation.
• The latest revision of the International Phonetic Alphabet includes a number of symbols for indicating prosodic characteristics of utterances but these have not met widespread acceptance in studies on the intonation of different languages (Bruce 1989).
• One recent attempt to define a standard for prosodic notation is the TOBI (=TOne and Break Index) system (Silverman et al. 1992) which has been designed for labelling the prosody of American English. This system makes use of two sets of labels:
- break indices - an integer value (0-4) is used to reflect the relative degree of separation between any two adjacent words in an utterance, and
tones
- a set of symbolic symbols are used to represent the inventory of pitch patterns found on accented syllables and at intonational phrase boundaries in American English.
Although there have been some attempts to apply similar systems to other languages, the TOBI system is not in our view usable in its present state as a multi-language prosodic notation system since it presupposes that the set of relevant pitch patterns for a given language is already known. For the majority of languages represented in the MULTEXT project, this information is not available. Even in the case of English the inventory of tonal configurations assumed in TOBI is not uncontroversial.
• A system which has been specifically developped for transcribing the prosody of several different languages is the INTSINT system (Hirst & Di Cristo forthcoming). This assumes that pitch patterns can be adequately described using a limited set of tonal symbols, (T,M,B,H,S,L,U,D standing for :Top, Mid, Bottom, Higher, Same, Lower, Upstepped, Downstepped respectively) each one of which characterises a point on the fundamental frequency curve . Unlike the TOBI system, INTSINT does not assume that the inventory of pitch patterns of a given language is already known.
We have decided to incorporate ideas from both TOBI and INTSINT in our prosodic notation. In order to make our data as compatible as possible with the TOBI system we shall carry out

the alignment of the word-boundaries and the onsets of the stressed syllables with the speech signal. The pitch patterns will be described using the INTSINT system after modelling the fundamental frequency curves as a quadratic spline function interpolating between a sequence of target points.

4 . Prosody tools

We have developed a number of tools for analysis which will be used as a starting point in MULTEXT. These tools will be adapted to the general MULTEXT software philosophy and integrated into the general toolkit. The tools will perform three tasks:

(i) Automatic modeling of the F0 curve from the speech signal.

This tool will implement the method described in Hirst et al. (1991) and Hirst & Espesser (1993), using a technique called "asymmetrical modal quadratic regression". The output of this tool is a sequence of target points (Hz, ms), which consitute a stylisation of the F0 curve. This output is of great interest in itself, since it has been shown that fundamental frequency synthesis by a quadratic spline function interpolating the target points is virtually indiscernible from the original.

(ii) Automatic symbolic coding of intonation from the sequence of target points.

This tool will take the sequence of target points generated by the previous tool, and produce a symbolic coding of intonation in the INTSINT system (Hirst & Di Cristo forthcoming). The output will be a sequence of INTSINT symbols, coupled with their place of occurrence on the temporal scale (in ms).

(iii) Alignment of INTSINT coding and phonemic or orthographic transcription.

If the alignment of the phonemic or orthographic transcription with the speech signal is known, at least in terms of word boundaries and stressed syllables, a third tool can easily align the symbolic coding and the transcription.

Since MULTEXT provides tools for marking morphology and POS in the orthographic transcription, this Task makes possible a complete chain of alignments for several levels of language, from the speech signal up through the level of POS.

Figure 1 illustates the French sentence "J'ai des problèmes avec mon adoucisseur d'eau" (I have problems with my water-softener), an extract from one of the passages of the French

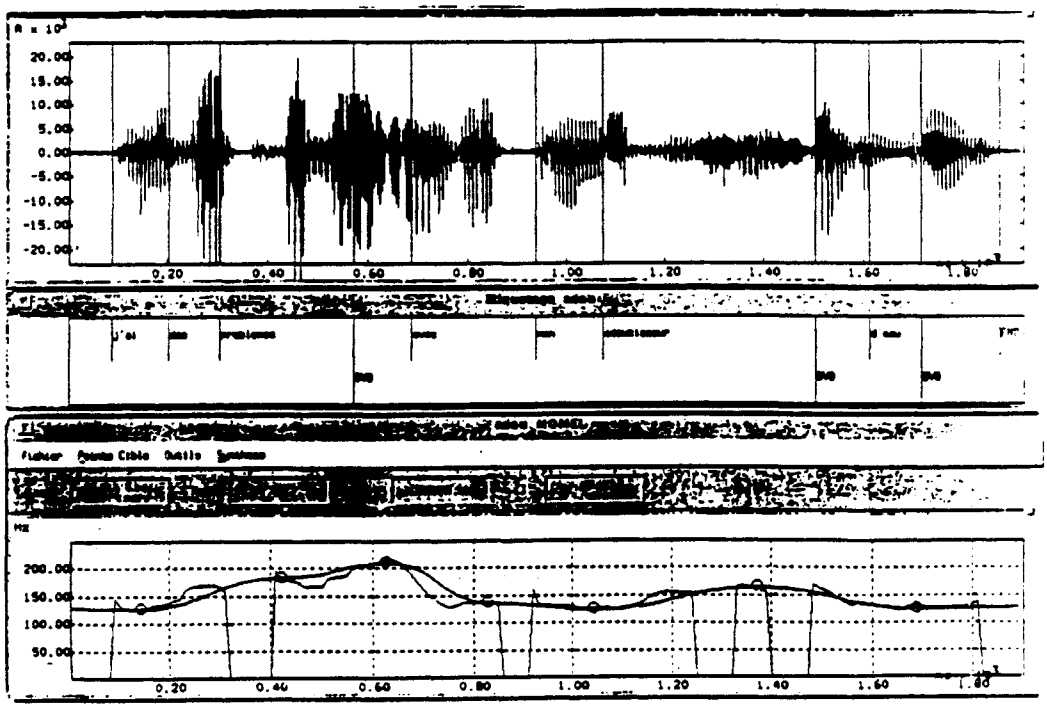

Figure 1 a sample sentence from the French recording of EUROM1 illustrating the prosodic labelling to be carried out in MULTEXT (cf text).

recordings from the EUROM1 corpus. In the top panel the wave-form is displayed. The panel beneath this displays two sets of labels. word onsets, and stressed vowel onsets. The end of the sentence is marked with the label END. The label cursors are synchronised with cursors on the wave-form panel. The word labels were entered by hand and synchronised with the wave-form using facilities for audio play back to check that they had been placed correctly. Stressed vowel onset was marked using automatic vowel-onset detection (Hermes 1990) and checked manually.
The bottom window displays both the raw fundamental frequency curve and the smoothed modelled curve interpolating between seven target points (represented by small circles on the display).
The output of the analysis is a set of three label files, word-boundaries, stressed vowel-onsets and F0 target-points, each of which is time-aligned with the speech signal. The third label set will later be converted to a set of symbolic labels such as the following : [M U T D L H B] also time-aligned with the speech signal.
The statistical values of the target points from one complete set of recordings for each language will be modelled either by the mean value of the symbol class (for the symbols M T and B) or by linear regression on the preceding target for the symbols (H S L U D). This will be used to generated synthetic F0 contours from the symbolic coding. An evaluation of the quality of the resulting synthetic speech for the six languages will provide a useful evaluation metric for the coding system as a cross-language analysis tool.

References

Bruce, G. (1989) "Report from the IPA working group on suprasegmental categories." *Working Papers* (Lund University) 35, 25-40

Hermes, Dik. J. (1990) "Vowel-onset detection." *J. Acoust. Soc. Am.* 87 (2), 866-873.

Hirst, D., Di Cristo, A. (forthcoming) "A survey of intonation systems." In Hirst, D., Di Cristo, A. (eds) (forthcoming) *Intonation Systems: a survey of twenty languages*. Cambridge University Press, in press.

Hirst, D., Espesser, R. (1993) "Automatic modelling of fundamental frequency." *Travaux de l'Institut de Phonetique d'Aix*, 15, 71-85.

Hirst, D., Nicolas, P., Espesser, R. (1991) "Coding the F0 of a continuous text in French : an Experimental Approach." *12eme Congres International des Sciences Phonetiques*, Aix-en-Provence, 5, 234-237.

Ide, N., Véronis, J. (1993a). "Background and context for the development of a Corpus Encoding Standard", *EAGLES Working Paper*, 30p.

Ide, N., Veronis, J. (1993b). "What next after the Text Encoding Initiative? The need for text software." *ACH Newsletter*, Winter 1993, 1-12.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., Hirschberg, J. (1992) "TOBI : a standard for labeling English prosody." Proc. Internal. Conf. Spoken Language Processing Vol. 2, 867-870.