

PROZED: A MULTILINGUAL PROSODY EDITOR FOR SPEECH SYNTHESIS.

Daniel Hirst

1. Introduction

It is generally agreed today that the single most important progress which needs to be made to improve the quality and naturalness of synthetic speech is towards a better understanding and control of prosody. This is true even for those languages which have been the object of considerable research (e.g. English, French, German, Japanese, ...) - it is obviously still more true for the vast majority of the world's languages for which such research is either completely inexistant or is only in a fairly preliminary stage. Even for the best-researched languages, there is still very little reliable and robust data available on the prosodic characteristics of dialectal and/or stylistic variability.

It seems inevitable that the need for prosodic analysis of large speech data-bases from a great variety of languages and dialects as well as from different speech styles will increase exponentially over the next two or three decades. In this paper I present an overview of *ProZed* an aid for developing prosody rules for speech synthesis using the *MOMEL* and *INTSINT* [19], algorithms and interfaced with the *MBROLA*, [12]*MBROLIGN* [23] and *Praat* [4] programs. It allows the interactive editing of a symbolic representation of an utterance in any of the twenty languages and dialects for which an *MBROLA* diphone database is currently available.

ProZed defines a number of different levels of representation of varying abstraction. At the lowest level, representations are a specification of the identity of each phonemic segment together with its prosodic characteristics. More abstract representations allow the user to abstract away from speaker-specific characteristics in order to concentrate on the meaningful content of the utterance's prosody.

It is evident that no tool is entirely innocent of theoretical bias and the tools described here are no exception to this rule. *ProZed* has, however, been designed to be as theory-independent as possible so that it could be used to describe intonation patterns in a number of different frameworks with the ultimate aim of providing a framework that could be used to evaluate competing models of prosody.

The program is currently implemented as a set of Perl scripts. Like the other programs with which it is interfaced, *ProZed* will be freely distributed for non-commercial, non-military applications. It is hoped that this tool will greatly facilitate the task of phoneticians and linguists involved in developing efficient rules for speech synthesis for a wide variety of languages and dialects.

2. Levels of representation.

The number of factors which contribute to the prosodic characteristics of a given utterance are quite considerable. These factors may be language-specific, dialectal, individual, syntactic, phonological, semantic, pragmatic, discursive, attitudinal, emotional... The list is obviously far from complete.

Many approaches to the study of prosody attempt to link such factors directly to the acoustic characteristics of utterances. The approach I outline here is rather different. Following [20], I propose

CNRS Université de Provence, Aix-en-Provence France email: daniel.hirst@lpl.univ-aix.fr

to distinguish four different *levels of representation* : the *physical* level, the *phonetic* level, the *surface phonological* level and the *underlying phonological* level. Each level of representation must conform to an *interpretability constraint*, which states that each level must be interpretable in terms of adjacent levels of representation. The underlying phonological level is conceived of as the interface between the representation of phonological form and syntactic/semantic interpretation. I shall not be concerned here any further with this level although it is this level which is presumably conditioned by the different factors listed above. In the rest of this paper I describe the phonetic and surface phonological levels of representation together with their implementation in ProZed.

The aim of the research programme characterised by this approach is to develop automatic procedures which define a reversible mapping between acoustic data and phonetic representations on the one hand and between phonetic representations and phonological representations on the other hand. This programme aims, consequently, (at least as a first step) not to *predict* the prosodic characteristics of utterances but rather to *reproduce* these characteristics in a robust way from appropriate representations.

2.1 Phonetic representation

Following Trubetzkoy, I take the division between phonology and phonetics to be that between abstract qualitative characteristics on the one hand and continuous quantitative variables on the other hand. The first step (*phonetic representation*) is consequently to reduce the acoustic data to a small set of quantitative values from which it is possible to reproduce the original data without significant loss of information.

2.1.1 Duration

A phonetic representation of duration is obtained simply by the alignment of symbols from a phonetic alphabet. (*SAMPA* [32], [33]) with the corresponding acoustic signal. Traditionally such alignments have been carried out manually. This task is very time-consuming and extremely error-prone. It has been estimated that it generally takes an experienced aligner more than fifteen hours to align phoneme labels for one minute of speech (nearly 1000 times real time).

Software exists to carry out this task (or at least a first approximation) automatically ([9], [28], [31]). Such software, which generally uses hidden Markov modelling, requires a large hand-labelled training corpus. Recent experiments, however, ([11], [23]) have shown that a reasonably accurate alignment of phonemic labels can be obtained without prior training by using a diphone synthesis system (such as that described in [12]). Once the corpus to be labelled has been transcribed phonemically, a synthetic version is generated with a fixed duration for each phoneme and with a constant F0 value. A dynamic time warping algorithm is then used to transfer the phoneme labels from the synthetic speech to the original signal.

Once the labels have been aligned with the speech signal, a second synthetic version is generated using the duration defined by the aligned labels and using the fundamental frequency of the original signal. This second version can then be re-aligned with the original signal using the same dynamic time-warping algorithm. The process can be re-iterated until no further improvement is made.

2.1.2 Fundamental frequency

Fundamental frequency is modelled using a quadratic spline function. The algorithm MOMEL (**Erreur! Source du renvoi introuvable.**) factors the raw F0 curve into two components: a microprosodic component corresponding to short-term variations of F0 which are conditioned by the nature of the individual phoneme, and a macroprosodic component which corresponds to the longer term variations independent of the nature of the phonemes. The output of the modelling is a sequence of target points corresponding to the linguistically significant pitch targets.

Evaluation of the phonetic level of representation of F0 for English, French, Italian, Spanish and Swedish has been carried out within the MULTEXT project [30], [2] using the EUROM1 corpus [7]. The results show that the algorithm is quite robust. On the English and French recordings (all together about one and a half hours of speech) about 5% of the target points needed manual correction. Many of these corrections involved systematic errors (in particular before pauses) which an improvement of the algorithm should eliminate.

2.2 Surface phonological representation

At this level, an utterance is represented as a sequence of phonemic segments. The relative duration of each phoneme is specified on a five-point scale. Each phoneme may also be accompanied by an abstract *INTSINT* tonal symbol together with a specification of its alignment with respect to the phoneme. **INTSINT** (an **IN**ternational **T**ranscription **S**ystem for **IN**Tonation) was developed during the preparation of a study of the intonation of twenty languages and was used to transcribe examples of intonation patterns for nine of these: British English, Spanish, European Portuguese, Brazilian Portuguese, French, Romanian, Russian, Moroccan Arabic and Japanese. It aims to capture the surface distinctions used in different languages for building distinctive intonation patterns. Unlike many other transcription systems, including in particular ToBI INTSINT is entirely concerned with the representation of prosodic form rather than of prosodic function. In this sense it can be thought of as a prosodic equivalent of a narrow IPA transcription system for segmental transcriptions.

INTSINT represents an intonation pattern as a sequence of 'tones', coding the relative height of the significant "target points" of the pattern. Three of these tones: *Top*, *Mid* and *Bottom* are assumed to be defined globally with respect to the speaker's pitch range. Three other tones: *Higher*, *Same*, *Lower* are defined locally with respect to the preceding tone. Two further tones *Upstepped* and *Downstepped* are similar to *Higher* and *Lower* but imply a smaller interval with respect to the preceding tone. Typically, *Upstepped* and *Downstepped* are used in iterative sequences whereas *Higher* and *Lower* will generally correspond to peaks and valleys. Table 1 (from [20]) shows the orthographic and iconic symbols used in INTSINT.

Table 1: Orthographic and iconic symbols for the INTSINT coding system.

<i>ABSOLUTE</i>	T ↑	M ⇒	B ↓
<i>RELATIVE NonIterative</i>	H ↑	S →	L ↓
<i>Iterative</i>	U <	•	D >

The choice of tonal symbols implies a quantification of the frequency domain. The alignment of the two sets of symbols, however, has not yet been the object of similar quantification. Instead, typically,

the symbols are aligned graphically and analogically as in the following, a transcription of the French utterance "Il faut que je sois à Grenoble Samedi vers quinze heures." (I have to be in Grenoble by Saturday 3 p.m.):

(1) [ilfok@Z@swazagR@nObI][samdivERk~Ez9R]
 [↑ ↓ ↑ ↓ ↑] [↑ ↓ ↑ ↓]

The fact that INTSINT codes prosodic form rather than prosodic function means that it can be used as a data-driven tool for automatically extracting 'linguistic-like' information from acoustic data. In conjunction with MOMEL, which provides an automatic stylisation of F0 curves INTSINT has been used as a reversible coding system for F0 curves for a number of languages . and the representation system developed has now been implemented in two text-to-speech systems for French.

In the rest of this paper, rather than the iconic symbols used in (1), I use the orthographic symbols (T, M, B etc.) which (with the exception of D and L) have been integrated into the SAM phonetic alphabet SAMPA [32], [33].

Two types of INTSINT representations are possible: linear representations and "tiered" representations. In tiered representations, segments and tones are coded in different tiers or streams (columns, fields...). There is consequently no ambiguity between the two sets of symbols used. For linear representations, however, there is overlap between the segmental symbols and the tonal symbols /D/ = IPA /D/, /T/ = IPA /T/ /S/ = IPA /S/ etc. In order to distinguish tones and segments I follow Wells' (1995) proposal (originally suggested by Dafydd Gibbon) to include tonal symbols in angled brackets <> acting as 'tier escape' symbols.

For the relative alignment of tones and segments, different levels of representation can be assumed. At a surface level Hirst (forthcoming) proposed to align tones with segments. The relative timing of the tone to the preceding segment can be specified by means of 4 symbols '[', '+', '-', and ']' corresponding to *beginning*, *early*, *late* and *end* respectively. When there is no diacritic it is assumed that the tone symbol is aligned with the *middle* of the preceding segment.

In example 2 below the utterance "Like this" (/l a I k D I s /) is pronounced with a **M**id tone at the beginning (actually this tone will be interpreted as occurring in the middle of the pause preceding the utterance), the pitch then rises to a **T**op tone late in the diphthong /ai/, drops to a **D**ownstepped tone late in the vowel /I/ and then falls to the **B**ottom of the speaker's pitch range in the middle of the final pause. All the segments are of average length except the final vowel, which is lengthened, and the final consonant, which is extra-lengthened.

–	M
l	
AI	T+
k	
D	
I+	D[
s++	
–	B

(2) A tiered INTSINT representation of a reading of the utterance "Like this".

Using a linear transcription the same utterance would be coded:

$_ <M> | A I <T> + k <D> I + <D> [s + + _ $

(3). A linear INTSINT representation of the same utterance as in (2)..

In the preliminary version of *ProZed* presented here only one tonal symbol per phoneme is allowed but more complex contours can be obtained by concatenating a sequence of appropriate phonemes. Phoneme durations are calculated from a table of means and standard deviations following [5], tonal symbols are converted to F0 targets and the relative alignment of the target is specified as a percentage of the phoneme duration. Temporal, pitch and alignment parameters can be modified from default values dynamically by a statement of the form:

$<parameter \quad attribute=value>$

INTSINT can be used to test prosodic models for speech synthesis. In order to provide auditory feedback, the system has been interfaced with the MBROLA diphone speech synthesiser by means of a script written in Perl (Wall et al. 1991) which takes as input a tiered INTSINT file (.int) (like that of example 3 above) and provides as output an MBROLA file (.pho) with appropriate durations and pitch values. If the user has the MBROLA synthesiser, the appropriate set of diphones and a table of mean durations for the phonemes of the language (stored in the same directory as the program *int2pho* under the name *durations.txt*), *int2pho* can be used for any of the languages currently available for the MBROLA project.

The program calculates the segmental durations and pitch targets using a number of parameters. Parameters for duration are: *tempo*, *extrashort*, *short*, *long*, *extralong*. Parameters for F0 are *base*, *range*, *lower*, *higher*, *upstep*, *downstep*, *same*. Parameters for alignment are *beginning*, *early*, *middle*, *late*, *end*. All the parameters are initialised to standard values but any of them can be modified by the user in his transcription by inserting a line containing the symbol then the name of the parameter and its new value between angle brackets. Thus for example tempo is assigned the default value of 1, a line containing :

$<parameter \quad tempo=0.8>$

will increase the duration of segments from that point on by 20%. There are no restrictions as to the number of times parameters can be modified in a transcription. Lines containing the symbol ";" are treated as comments and together with empty lines are skipped.

Segmental duration is established by looking up the mean value for the phoneme in the file *durations.txt*, multiplying this by the appropriate lengthening or shortening ratio if necessary and multiplying the result by the tempo parameter.

The F0 target value is calculated by the following algorithm [20]:

$$ is assigned the value of the parameter *base*.

$<T>$ is assigned the value of the parameter *base* multiplied by the parameter *range*.

$<M>$ is assigned the mean of $$ and $<T>$.

The relative tones $<H>$ and $<U>$ are calculated by raising the value of the previous target towards the value of $<T>$ by a proportion defined by the appropriate parameter *higher* or *upstep*. The relative tones $<L>$, $<D>$ and $<S>$ are calculated by lowering the value of the previous target towards

the value of by the appropriate parameter *lower*, *downstep* or *same*. Default values for *higher* and *lower* are 0.5, for *upstep* and *downstep* 0.25 and for *same* 0.1. The result of this algorithm is that a sequence of <D> tones will asymptotically decline (as observed by Pierrehumbert and Liberman 1984) and that a sequence of alternating H and L values will also asymptotically decline as in the figure2 without any need for a specific declination component.

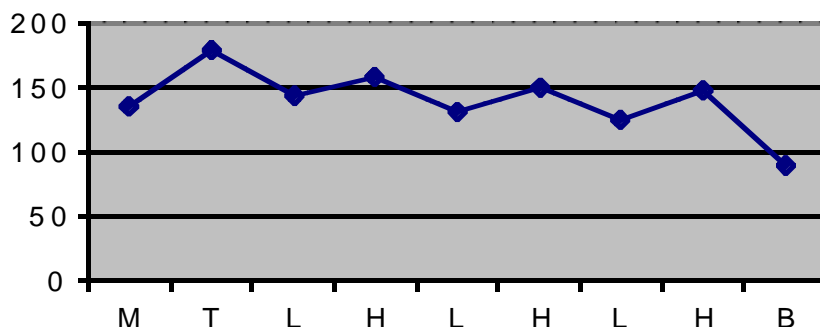


Figure 2. A sample sequence of F0 targets generated by the program int2pho with the parameter values *base*=90, *range*=2, *higher* = 0.4, *lower* = 0.4 illustrating the asymptotically downdrifting pattern which results.

The algorithm described above is applied by default on a linear scale. It is however possible to apply the same algorithm on a log scale by modifying the parameter scale (i.e. by including the line <scale log>). The script at present does not allow the user to specify other scales (Mel, Bark, ERB etc.) to be used but providing for these would be simple.

The relative alignment of the tone with respect to the preceding phoneme is determined by the presence or absence of an alignment diacritic. When there is no alignment diacritic, the tone is aligned by default with the parameter *middle*, which is initialised to 50%. The other parameters determined by the diacritics [, - , + ,] are initialised respectively to 0, 25, 75 and 100%.

MBROLA in its present form interpolates linearly between target points. INTSINT assumes in fact that between target points there is a curvilinear interpolation, which enables users to make a much sparser representation of the intonation pattern. It is hoped that a future revision of MBROLA will allow users the option of interpolation with a quadratic spline function as described by Hirst 1983, Hirst et al. (in press).

Most models of prosody, in fact, do not assume, as I have done above, that tones are aligned with phonemes. Instead some higher-level constituent is often taken to be the level at which tones and segments are linked. In order to specify for example that a tone is aligned in some way with a syllable or a foot or a syllable rime rather with a particular segment it is necessary to indicate the appropriate sequence. This could be done simply either by a tiered representation such as:

-	M
nV	T
TIN	BU
-	

(4). A tiered INTSINT representation of the word "nothing" aligned syllable by syllable with the sequences <M><T> and <BU>.

For a linear representation of the same structure it could be represented using parentheses (which are not used for any other purpose in the X-SAMPA proposal) to group the appropriate string of segments as in:

M(nV)<H>(TIN)<BU>

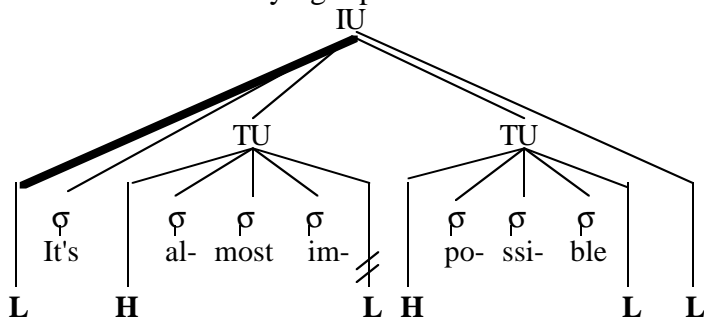
(5). A linear INTSINT representation of the word "nothing" aligned syllable by syllable with the sequence <MTBU>.

Rather than this I shall prefer a slightly more complex form using a notation called *polymetrical expressions* as recently proposed for the representation of simultaneous sequences in music [3]. A polymetrical expression (A, B, C...) is taken to represent the simultaneous realisation of the sequences A, B, C etc.. This notation has the advantage that it could be extended to cover any number of different tiers rather than just two. Example 4 using this type of representation would be coded:

<M>(nV, <H>)(TIN, <BU>)

(6). A polymetrical INTSINT representation of the word "nothing" aligned syllable by syllable with the sequence <MTBU>.

[18] describes a multi-linear underlying representation of an intonation pattern of the form:



where tones are linked to different higher level prosodic constituents. The double bar across the link to the second L tone is taken to represent the fact that this tones is 'floating', that is that it does not surface as a tonal target although it will influence the pitch value of the subsequent H tone.

With INTSINT this could be represented:

(Its(O:lm@UstIm, <HL>)(pQsIbl=, <H!L>), <L[L]>)

where the symbol '!' is taken as indicating that the following L tone is "floating".

3. Deriving 'linguistic-like' information from acoustic data.

The next step in this project will be to develop a reverse model deriving a 'linguistic-like' representation from acoustic data. For surface representations this is relatively simple¹. Preliminary results from the application of this system to the Eurom1 corpus **Erreur! Source du renvoi introuvable.** will be presented.

For more abstract representations a number of challenging and interesting problems of phonetic interpretation arise which will be addressed in future work.

References

- [1] Arvaniti, Amalia Ladd, D. R. and Mennen, Ineke 1998. Stability of Tonal Alignment: the case of Greek Prenuclear Accents. *Journal of Phonetics*
- [2] Astésano, C.; Espesser, R.; Hirst, D.J. & Llisterra, J. 1997. Stylisation automatique de la fréquence fondamentale : une évaluation multilingue. *4e Congrès Français d'Acoustique*, 14-18 avril 1997, Marseille 441-443.
- [3] Bel, B. 1992. Symbolic and sonic representations of sound-object structures. In M. Balaban, K. Ebcioglu, and O. Laske (eds.) *Understanding Music With AI*. Menlo Park: AAAI Press, pp.64-109.
- [4] Boersma, P. , Weenink, D. 2000. *Praat: a system for doing phonetics by computer*. <http://www.fon.hum.uva.nl/praat/>
- [5] Campbell, W.N. 1992. *Multi-level Timing in Speech*, PhD Thesis, University of Sussex.
- [6] Campione, E., Flachaire, E., Hirst, D.J. & Véronis, J. 1997. Stylisation and symbolic coding of F0, a quantitative approach. *ESCA Tutorial and Research Workshop on Intonation*. 18.-20.September. Athens.
- [7] Chan, D., Fourcin, A., Gibbon, D., Granström, B., Huckvale, M., Kokkinas, G., Kvale, L., Lamel, L., Lindberg, L., Moreno, A., Mouropoulos, J., Senia, F., Trancoso, I., Veld, C., Zeiliger, J. 1995. EUROM: a spoken language resource for the EU. *Proceedings of the 4th European Conference on Speech Communication and Speech Tecnology, Eurospeech '95*, Madrid vol. 1, 867-880.
- [8] Courtois, F., Di Cristo, Ph., Lagrue, B., Véronis, J. 1997. Un modèle stochastique des contours intonatifs en français pour la synthèse à partir des textes. *4ème Congrès Français d'Acoustique*. Marseille, avril 1997, 373-376.
- [9] Dalsgaard, P. Andersen, O. & Barry, W. 1991. "Multi-lingual alignment using acoustic-phonetic features derived by neural-network technique." *ICASSP-91*, 197-200.
- [10] Di Cristo, A., Di Cristo, P., Véronis, J. 1997. A metrical model of rhythm and intonation for French text-to-speech synthesis. *Proc. ESCA Workshop on Intonation*, Athens Sep. 1997, 83-86.
- [11] Di Cristo, Ph. & Hirst, D.J. 1997. Un procédé d'alignement automatique de transcriptions phonétiques sans apprentissage préalable. . *4e Congrès Français d'Acoustique*, 14-18 April, Marseille, 425-428

¹ The MacPerl script *int2pho* together with a preliminary version of the inverse script *pho2int* as well as example labelled files from the Eurom1 corpus **Erreur! Source du renvoi introuvable.** will be found at the address <http://www.lpl.univ-aix.fr/~hirst>

- [12] Dutoit, T. 1997. *An introduction to Text-to-Speech synthesis*. Kluwer Academic Press, Dordrecht.
- [13] Dutoit, T., Pagel, V., Pierret, N., Bataille, F., Van Der Vrecken, O. 1996. The MBROLA project. Towards a set of high-quality speech synthesisers free of use for non-commercial purposes. *Proceedings ICSLP '96* (Philadelphia) 3: 1393-1396.
- [14] Hirst, D.J. 1983. Structures and categories in prosodic representations. in A. Cutler & D.R.Ladd (eds.) *Prosody: Models and Measurements*. Springer, Berlin. 93-109.
- [15] Hirst, D.J.; A. Di Cristo, M. Le Besnerais, Z. Najim, P. Nicolas, P. Roméas (1993) Multi-lingual modelling of intonation patterns. *Proceedings ESCA Workshop on Prosody*. Lund, Septembre 1993, 204-207.
- [16] Hirst, D., Espesser, R. 1993. Automatic Modelling of Fundamental Frequency using a quadratic spline function, *Travaux de l'Institut de Phonétique d'Aix-en-Provence*, 15, 75-85.
- [17] Hirst, D.J. & Di Cristo, A. 1998. A survey of intonation systems. In D.J.Hirst & A. Di Cristo (eds) *Intonation Systems. A survey of Twenty Languages*. Cambridge, Cambridge University Press., 1-44.
- [18] Hirst, D.J. 1998. Intonation in British English. in Hirst & Di Cristo eds. 1998, 56-77.
- [19] Hirst, D.J. 1999. The symbolic coding of duration and alignment. An extension to the INTSINT system. *Proceedings Eurospeech '99*. (Budapest, September 1999)
- [20] Hirst, D., Di Cristo, A., Espesser, R. in press.. Levels of representation and levels of analysis for the description of intonation systems. In Horne, M. (ed.) in press.
- [21] Horne, M. (ed.) in press.. *Prosody: Theory and Experiment*, Dordrecht; Kluwer Academic Publishers.
- [22] Liberman, M., Pierrehumbert, J.,. 1984. Intonational invariance under changes in pitch range and length. in M. Aranoff and R. Oerhle (eds.) *Language Sound Structure: Studies in Phonology Presented to Morris Halle*. Cambridge, Mass.; MIT Press., 157-233.
- [23] Malfrère, F. & T. Dutoit 1997. High Quality Speech Synthesis for Phonetic Speech Segmentation, *Proceedings Eurospeech. '97*. 2631-2634.
- [24] Mora, E. Hirst, D.J., Di Cristo, A. 1997. Intonation features as a form of dialectal distinction. in *Proc. ESCA Workshop on Intonation*, Athens Sep. 1997, 247-250.
- [25] Pierrehumbert, J. in press. Tonal elements and their alignment. in M. Horne (ed.) in press.
- [26] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., Hirschberg, J. 1992. ToBI: a standard for labelling English prosody. *Proceedings ICSLP'92*, 2, 867-870, Banff, Canada.
- [27] Strangert, E. and Aasa, A. 1996. Evaluation of Swedish prosody within the MULTTEXT-SW project. *TMH-QPSR 2/1996 (Speech, Music and Hearing - Quarterly Progress and Status Report)*, KTH, Stockholm, Sweden, 37-40.
- [28] Talkin & C. Wightman (1994) The aligner. *Proceedings ICASSP 1994*.
- [29] Véronis, J., Di Cristo, P., Courtois, F. & Lagrue, B. 1997. A stochastic model of intonation for text-to-speech synthesis. *Proceedings Eurospeech '97* (Rhodes) 5: 2643-2646.
- [30] Véronis, J., Hirst, D.J., Espesser, R., Ide, N. 1994. NL and speech in the MULTTEXT project. *AAAI '94 Workshop on Integration of Natural Language and Speech.*, 72-78.

- [31] Vorsterman, A. Martens, J.P. & Van Coile, B. (1996) Automatic segmentation and labelling of multi-lingual speech data. *Speech Communication* 19, 271-293.
- [32] Wells, J.C. 1995. Computer-coding the IPA: a proposed extension of SAMPA. ms.
<http://www.phon.ucl.ac.uk/home/sampa/x-sampa.htm>.
- [33] Wells, J.C., Barry, W., Grice, M., Fourcin, A., Gibbon, D. .1992. Standard computer-compatible transcription. *Esprit project 2589 (SAM), Doc. no. SAM-UCL-037*. London, Phonetics and Linguistics Department, UCL