Polish Synthesis and Representation Levels of Intonation

Durand, P.† Durand-Deska, A.† and Gubrynowicz, R. ‡
† LPL, CNRS UMR 6057, Phonetic Dpt., Université de Provence, France
‡ I.F.R.T., P.A.N., Poland
E-mail: durand@lpl.univ-aix.fr, rgubryn@ippt.gov.pl

ABSTRACT

In this paper, we describe the basic levels of prosodic representation applied to "Text-to-Speech" synthesis of Polish sentences by concatenation of diphones. The goal is to supply the synthesis device used in this project prosodic information in a way to fit natural sentences one. This work is devoted to sentences with "Czy" interrogative clause which show specific Fo contours. In MBROLA synthesis system, its necessary to supply Fo and duration information for each segment. On the upper level, INTSINT labeling is used because of its simple formal melody coding. At the intermediate level, it is necessary to find a scale that focuses on perceptually relevant melodic variations for different voices. For this purpose the semi-tone scale is used for it takes into account melodic variation and is independent of pitch absolute value. Given the melodic span of a given voice, it's simple to get contour in Hz to apply in MBROLA system.

1. INTRODUCTION

This paper is a part of a large task build out of a corpus of Polish sentences collected in the BABEL frame[1], the first results of which are presented in earlier communications [2-3-4-5] or forthcoming one [6]. We intend to manage "Text-to Speech" synthesis of Polish using MBROLA diphones concatenation program[7].

From the written text, we have to supply the synthesis system with three kind of data:

- the SAMPA transcription of the text to be synthesized.
- durational information of each segment
- Fo information on relevant segments.

Polish has conversion rules between oral and written text that make it easy[8]. The questionable points for this conversion are not relevant for our work.

Durational information was judged and still is judged as not significant for cueing stress[9-10-11]. Nevertheless, we show that hearers are sensitive to changes in this signal dimension [4-5].

Melodic information to be given is restricted to relevant voiced segments, the program linking the melody between two successive or not successive segments.

2. THE CORPUS

The data we are going to use to establish the melodic transcription are extracted from sentences of various size included in 40 short monologues out of Polish version of the BABEL database. The 60 speakers included a range of female and male voices as well as vocal strategies. They were asked to read monologues in a dramatized fashion in order to introduce context sensitive cues, and avoid those specific to reading.

Questions to be produced by the subjects were placed at beginning of the monologues, or were preceded by an assertion or another question. This ensured that a wider range of context sensitive prosodic cues would be used. Out of this corpus, 13 interrogative sentences with "Czy...?" were extracted. Out of them six typical sets of recordings were selected, reflecting casual pronunciation of modern Polish. The selection was to serve as a basis for a hypothesis on the prosodic cues involved in "czy ..?" questions to be tested against the remaining data. The results of this study are compatible with the analysis of the reduction of consonantal groups in fast speech [3], performed on the same acoustic data, or in spontaneous speech[2].

From the corpus composition one can deduce that neither serial effect, nor corpus induced regularity. If the reading task can give "laboratory speech", recording rules, and selection of speakers out of the 60 recorded ensure the selected items to be modern casual Polish representative. In fact, from such a database, the drawback remains the number of parameters to take into account even if the speakers selection, introduces a kind of normalization. The cues selected as a result of the analysis will allow the program to choose different strategies for generating the same type of sentences in order to avoid mechanical and artificial manner of "speaking".

3. THEORETICAL BACKGROUND

Various aspects of the phonological system of Polish ('a phonologist's paradise') have so far attracted the attention of different phonological schools. Most of them dealt with theoretical issues and were mainly concerned with applying the theory to specific languages. In metrical phonology, stress assignment constitutes an important

 ISBN 1-876346-48-5

topic and has received a great deal of attention from researchers. These studies are mainly devoted to the assignment of word-stress in single or two-word sequences.

Parallel to the interest in the theory, there is a well established tradition of acoustic analyses, with seminal influence of W. Jassem. The prosodic description is, however, inadequately represented and few reliable sources on this topic are available. Although Miko_ [13] analyses the intonation of all kinds of questions in Polish, his analysis is performed on a very limited body of questions read out in isolation, and does not take into account the problem of stress.

The impression created by the two sets of papers is that stress and intonation belong to two separate, unrelated domains and that there is no direct link between theoretical consideration and acoustic analysis. To fill the existing gap in the current knowledge of the prosodic system of Polish, the assumption made in the present paper is that stress and intonation are two different faces of the same prosodic phenomenon. Since intonation units are divided into stress groups, it is necessary to include both of them in order to acquire a synthesis of Polish utterances of satisfactory quality.

Furthermore, at the sentence or at the ”passage” level, semantic aspect as well as emotion, emphasis communicative intentions take place[14-15], and can modify the canonical shape by specific focalization, inducing in some sentences changes in the place of perceived stress.

4. FROM SIGNAL TO REPRESENTATION

As the synthesis program has to be supplied with Fo information in Hz, it could be easier to use in sentences analysis identical scale. This ensures a good alignment on the physical signal giving information both on the speakers' characteristics and the Fo evolution (Figure 1)

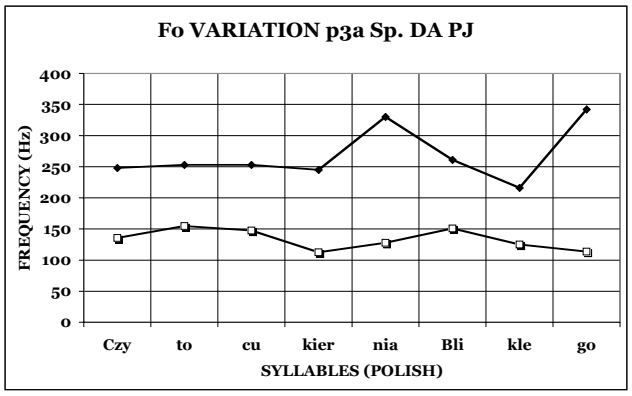

Figure 1: Melodic representation of Fo (Hz). Sentence p3a, speakers DA and PJ

The emphasis, given by this kind of representation on the absolute frequency value, make difficult inter-subject comparison, and uneasy the analysis of Fo variations. Moreover, identical Fo change produces different perceptive effects in accordance with initial Fo value.

For this reason, we used a log-frequency of Fo data, where 100Hz is the departure reference (St.=0 semi-tone). Conversion equations are:

$$Hz = 2^{st/12}\ 100 \quad (1)$$

$$St = 12[\ln (Hz/100)\ln 2 \quad (2)$$

As quotes G.Fant (15), the "span of modulation for males and females is closely the same on semi-tones scale"(Figure 2).

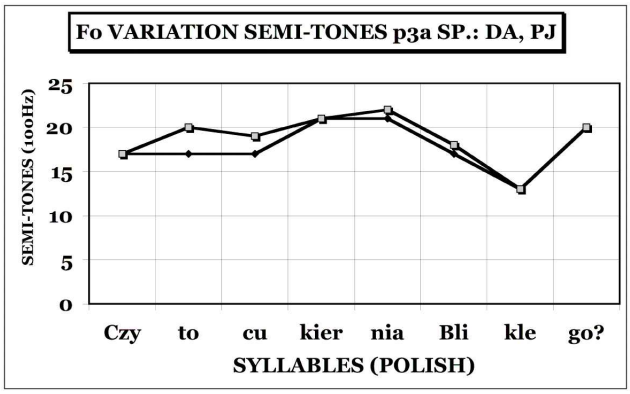

Figure 2: Melodic representation of Fo (Semi-tones). Sentence p3a, speakers DA and PJ

This semi-tones representation allows the comparison between different voices and different strategies in uttering the sentence (Figure 3)

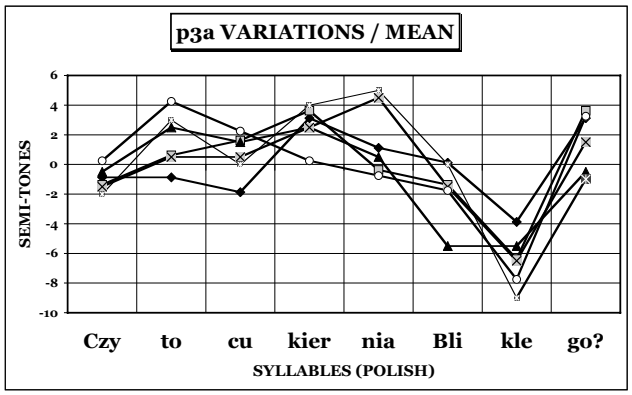

Figure 3 : Melodic representation of Fo (semi-tones). Sentence: p3a, Speakers: DA, DG, GA, PJ, Rtw and ZK. Variations are represented from Fo mean value of each speaker.

It's possible to take into account perceived and not perceived Fo fluctuation. The grouping of syllables that show not relevant variation is possible. In this case, semi-tone notation can show in cases where different focalization are not done striking similarities (Figure 4)

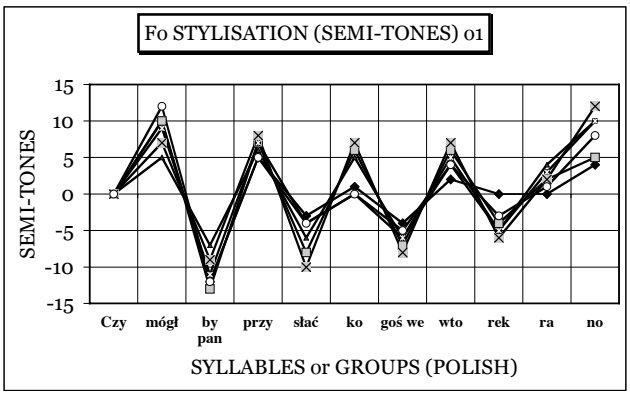

Figure 4 : Variation of Fo in semi-tones. Syllable with no relevant melodic change are grouped together. This figure show that stressed syllables wear a melodic prominence.

To allow a high quality synthesis with the help of a diphones library, it is necessary to use for melody a level of abstract description as phonetic transcription do for segments. As phonetic transcription can be used regardless of other parts of linguistics, it ought to be self-determined. For this reason, and for practical ones, we choose to use a language independent symbolic description, INTSINT, to label the higher level of prosodic description. This coding is not language dependant, and its application to Polish do not require a considerable amount of research to establish the intonation patterns of this language. It do not requires the help of skilled linguist to apply it, for its application is all but simple.

INTSINT description provides a purely formal encoding of prosodic events. It takes into account target points of melodic curve either by absolute labels, or by relative ones giving the inflexion of the curve from the previous labeled point.

Absolute levels (Top, and Bottom) measure the span of variation in a given item. For passages rules can modify these levels according to the place of the sentence in the larger unit[16]. An intermediate absolute label (Mid) is also proposed. It takes place in the middle of the range of Fo variation. If absolute level T and B can appear once in

a sentence, Mid can be helpful in resetting the Fo curve to prevent relative relative pitch notation to move away of the actual curve.

In INTSINT relative labels are used to show the local variations of the melody. They are divided in “non-iterative” ones, as High, Same and Low (H,S,L) and “iterative” ones as Upstepped and Downstepped (U, D). In our work, the use of U and D was for configurative reasons.

5. FROM ABSTRACT DESCRIPTION TO SPEECH SYNTHESIS

To allow a “text-to-Speech” synthesis of Polish using MBROLA program, it’s necessary to have a phonetic transcription using SAMPA. This transcription gives the sequence of phonemes that the program combine in diphones. As Polish has quite simple graphotactic rules, conversion rules of written text in a sequence of phones is easy and is done automatically.

The program asks for segmental durations, which contribute in the perception of prominence. This important point is to be dealed with elsewhere [5].

It also asks for Fo values in Hz. From the abstract description given in INTSINT, it is easy to derive a sequence of target points using semi-tones scale. To derive the Hz curve from the semitone one, it’s necessary to know the span of variation of the speaker’s Fo. From the same abstract description, it is possible to generate voices with different range of Fo.

The sentences analysis process and the synthesis from the text are summarized in the following figure :

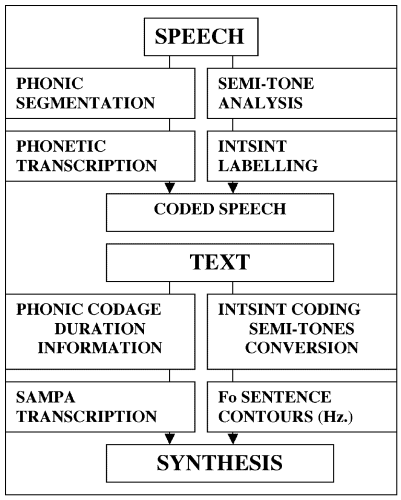

Figure 5 : Speech-analyze process and synthesis from the text using in analysis semi-tone scale to reach abstract labeling, in synthesis, using semi-tones as intermediate level between abstract labeling and the actual Hz fundamental curve, with the necessity to give the span of speaker’s Fo variation to get the sentence contour.

6. RESULTS

The synthetized sentences produced by this double way, from actual sentences to abstract labels by the semi-tone step in the two ways, show a great deal of naturalness. Speakers’voice characteristics remain close to the original.

ISBN 1-876346-48-5

CONCLUSIONS

Semi-tone scale and INTSINT abstract labeling show a good complementarity in the passage from substance to melodic form, and in the generation from abstract data of synthetized speech from written text. A light improvement of the conversion between semi-tones scale and INTSINT labels is in some cases necessary to get a higher quality.

Acknowledgement

This study was supported by the MAE/KBN Polonium 01424PA contract, the CNRS/PAN 10031 contract and received help from PACA Council

REFERENCES

[1] R., Gubrynowicz, "The Polish Database of Spoken Language." *Proc. First International Conference on Language Resources and Evaluation*, Granada, 1998, pp.1031-1037.

[2] R. Gubrynowicz, P. Durand, "The influence of speaking style on articulation of fricative-affricate clusters in Polish", Aix-en-Provence, *Proc. S.P.O.S.S.*, 1998, pp.39-42

[3] P. Durand, R. Gubrynowicz, "On the reduction of consonant clusters in Polish according to emotional situation", San Francisco, *Proc. XIVth I.C.Ph.S., Vol.III*, 1999, pp.2113-2115

[4] P. Durand, A. Durand-Deska, R.Gubrynowicz, B. Marek"Polish: Prosodic Aspects of "Czy" questions" *Prosody 2002*, Aix-en-Provence, 2002, pp.255-258

[5] P. Durand, A. Durand-Deska, B. Marek «Stress Cues in Polish : Evidence from speech synthesis », *XVII ème Congrès des Linguiste*s, Prague, 2003

[6] P. Durand, A. Durand-Deska, R. Gubrynowicz, B. Marek, «About stress in Polish: Experiments on "Czy…?" interrogative sentences» forthcoming

[7] K.Szklanny, K.Marasek, «Polish Diphones», Warszawa, *Polish Japanese Institute of Information Technology*, 1992

[8] Z., Folejewski, "The Problem of Polish Phonemes", *Scando-Slavica*, 2, 1961, pp.87-92

[9] W. Jassem, *Aksent Języka Polskjiego*, Wrocław: Ossolineum, 1962.

[10]J., Zborowska, 2001. *Rythmic Organisation of Language*, Unpublished Thesis, Poznań, 2001

[11]G. Dogil, G., "Stress patterns in West Slavic Languages". *Phonetik Word Stress*, Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung, Universität Stuttgart, 1995. 61-87.

[12] Jassem, W., Morton, J., Steffen-Batóg, M., The perception of stress in synthetic speech-like stimuli by Polish listeners, *Speech Analysis and Synthesis*, 1, 1968, pp. 289-308.

[13] M. J. Mikoś, "Intonation of questions in Polish", *Journal of Phonetics*, 4/3, 1976 pp.247-253.

[14]B. Marek,." Focus and intonation".: *Phonological Investigations, Literary and Literary Studies in Eastern Europe*,. In J. Fisiak and S. Puppel (Eds.) vol.38, 1992 pp.443-465.

[15]G. Fant, A. Kruckenberg, K. Gustafson, J. Liljencrants, "New Approach to Intonation Analysis and Synthesis of Polish", Aix-en Provence, 2002, *Speech Prosody 2002*, pp.283-286.

[16]E. Campione, D. J. Hirst, J. Véronis, "Automatic Stylisation and Modelling of French and Italian Intonation", in *Intonation Analysis, Modelling and Technology*, A.Botinis Ed., 2000 pp. 185-207.