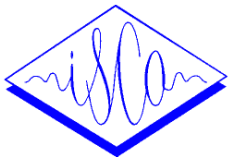

ETRW on Dialogue and Prosody
Veldhoven, The Netherlands
September 1-3, 1999

SYLLABLE BASED APPROACH TO AUTOMATIC PROSODY DETECTION; APPLICATIONS FOR DIALOGUE SYSTEMS

Ivan Kopeèek

Faculty of Informatics
Masaryk University
Botanická 68a,
602 00 Brno
Czech Republic
e-mail: kopecek@fi.muni.cz
http://www.fi.muni.cz/~kopecek/

ABSTRACT

Syllable based approach to prosody of Czech and an application for enhancing the communication in the developed dialogue programming system DIALOG is presented and discussed in the paper.

1. INTRODUCTION

Syllable based approach to speech processing is an interesting alternative to the diphone (triphone) - based approach, especially for the syllable-timed languages (see, e.g., [1] - [10], [15]). This is based both on the fact that more coarticulation aspects are included in syllable segments in comparison to diphone units, and on the fact that the main prosodic parameters (pitch, fundamental frequency, duration, and intensity) are closely connected to syllables. In what follows we discuss this approach in connection with applications for dialogue systems.

2. SEGMENTAL STRUCTURE OF THE USED CZECH PROSODY MODEL

Here we briefly mention the considered Czech prosody model (it is used also for speech synthesis in the DEMOSTHENES system - this synthesizer is incorporated in the dialogue programming system DIALOG - see Sections 6, 7). The model consists of the following basic types of segments:

- Syllables; as the basic prosodic elements, they carry prosodic attributes of pitch, duration and intensity (energy). They may also carry stress (accent).
- Stress units; they are segments of one rhythmical peak (in Czech containing typically one or two words) and of a varying number of syllable segments. The stress units form the first layer of the rhythmical structure of speech.
- Rhythm units; they consist of stress units and are delimited by a specific intonational pattern. (see, e.g. [14]). The juncture between two neighboring rhythm units is realized by intonation, stress and pause (the most prominent juncture element for many cases). The rhythm units form the second layer of the rhythmical structure of speech into which a connected portion of speech is phrased. Rhythm units are also called prosodic phrases. An algorithm for segmentation of the Czech text into rhythm units can be found in [6].
- Sentences; they are more grammatical segments then phonetic ones and include them into prosodic hierarchy may seem to be controversial, but from the point of view of speech synthesis or dialogue systems problems it is often reasonable.

3. SYLLABLES AND SYLLABLE SEGMENTS

Specifications of syllables are mostly more or less vague. This follows from the fact that the feeling of syllable boundaries, although very strong in most cases, is subjective and in many cases not unique. However, we need exactly defined segments to be able to handle them.

The specification of such segments is unfortunately complicated by the fact that the syllabification depends on such factors, as e.g. the pronunciation quality and the speech rate.

For instance, the Czech word *odstranit* (*remove, get off, take away*) is normally syllabified *od-stra-nit*. At low speech rate and accurate pronunciation the corresponding segments are *ot-stra-nit* but for common speech rate the phonetic transcription is *octranit* and possible segmentation is either *o-ctra-nit* or *oc-tra-nit* (neither the first nor the second possibility is consistent with the original syllables and both the segments "*o*" and "*oc*" do not respect the fact that "*od*" is a prefix).

Hence, it is reasonable to distinguish between "syllables" and "syllable segments". The example also shows, that speech rate

and pronunciation quality should be taken into account when reconstructing speech from speech segments .

Another reason that complicates the determination of an appropriate set of syllable segments is the necessity to respect the coarticulation effect between the adjacent syllables and to keep the number of segments in reasonable limits.

Naturally, syllable-based approach face the problem of relatively great inventory of syllables. Great number of syllables implies two main difficulties that have to be solved. First, how to handle coarticulation effect of the adjacent syllables; second, how to build up the database of syllable segments (clearly, it cannot be assembled manually).

The papers [9], [5] present methods to overcome these difficulties. The idea lies in exact defining of special syllable segments that are used for speech production and can be picked up so that good coarticulation of adjacent segments in the produced speech is preserved. These segments need not always be syllables in the classical sense, but boundaries between them can be well determined from phonetic point of view (enabling semi-automatic generation of the segment database), and the number of segments is reasonable.

Some problems of a concept of prosody model for syllable-timed languages are described in [8]. The discussed concept is implemented in DEMOSTHENES speech synthesizer [5].

4. ASSIGNING BASIC PROSODIC ATTRIBUTES TO SYLLABLES

The first thing we have to do is the segmentation into syllable segments, which simultaneously determinates the value of the duration of segments.

The semi-automatic segmentation procedure for syllable based speech synthesizer Demosthenes ([5], [9]) was based on detecting local maximums of the function of sonority decrease.

This method works satisfactorily for speech synthesis, where the estimation of the syllable segment boundaries may be effectively supported by the estimation of the segment length. However, for speech recognition the method is not sufficiently precise and reliable.

A method based on a similar idea as the method mentioned above is described in [13], but the main difference consists in more complex and more precise determination of the segment boundary characteristics. The method is based on knowledge of the boundary types, which is obtained by creating of a database of the boundary representatives.

The fundamental frequency (F0) is the most important prosodic attribute related to intonation. It is often obtained by the method of the short-period autocorrelation function. This method, however, may in some cases cause some difficulties, determining fundamental frequency 2 or 1/2 multiple of the proper value.

The method we have used for the determination of the fundamental frequency is based on the resonation principle ([11]). The lowest dominant frequency resonation in a defined region is detected and taken as the fundamental frequency. The method seems to be more stable and reliable then the autocorrelation method.

The intensity (energy) may substantially differ inside the syllable and depend on the presence or absence of stress. For the whole syllable we use the mean value of the intensity function inside the syllable.

5. INTONATION, SYLLABLE SEGMENTS AND DIALOGUE SYSTEMS

Intonation of a language is a very complex phenomenon related especially to semantics and pragmatics. We do not aspire to deep analysis of this very difficult problem, but we mention some aspects that can be used in dialogue strategies of dialogue systems.

Typically, intonation conveys information that is not obtained in the other language tiers. This is even more evident in the languages with free word order, like Czech. This property can be used to enhance the communication between a human and a dialogue system, but can be also a source of difficulties.

In what follows, we will not analyze the possible intonation schemes in Czech (which is a very complex problem not yet fully satisfactorily solved), but we restrict ourselves to the following intonation contours, that are of obvious interest from the point of view of applications in dialogue systems:

- Rising final intonation; this intonation is typical of question (query) or (especially on the boundaries of the prosodic phrases) also for indicating incompleteness of the information.
- Level final intonation; this type of intonation typically indicates a kind of incompleteness of the pronounced information.
- Falling final intonation; this intonation is typical for indicating completeness of the information.

It is worth noting that the intonation scheme may not be used if the information, which should be conveyed, is obtained directly in the word meaning. (For example the pronunciation of the Czech interrogative sentences that are introduced by the question words like *kde - where, kdy - when, kolik - how much*, etc., may lack the corresponding question intonation.)

Although the corresponding complete intonation schemes are various, in many cases we are able to recognize the type of intonation from the fundamental frequency and duration of the last segment. This is very regular, especially for

statements in the form of one-word sentences. In what follows, we will describe an application for the dialogue system DIALOG.

6. THE DIALOGUE PROGRAMMING SYSTEM "DIALOG"

The dialogue programming system DIALOG is developed at the Faculty of Informatics, Masaryk University Brno ([12]). The system is primarily built as a facility for the blind and visually impaired students and programmers. The system is framed to solve some typical problems that blind programmer meets, for example:

- Great number of statements and instructions that must the programmer (using a standard programming language as C++ or Pascal) know for developing a program.
- Problems with syntactical debugging an (partially) semantical debugging;
- Problems with editing of the developed program;
- Other difficulties that come from the fact that the present developing tools are graphically oriented.

To overcome or decrease these difficulties the system is developed as a natural language dialogue system (although the present versions has substantial restrictions). The system consists of the following basic modules:

- Communication module;
- Dialogue control module;
- Program generator and interpreter;
- Configuration module;
- Problem and user modeling module;

The communication module has the following functions:

- Speech recognition; the present version uses a statement (=isolated word) speech recognizer. In future, a continuous speech recognizer should be integrated. Speech is considered to be the main input of the system.
- Prosody detection, evaluation, and estimation; this function is closely connected especially to speech recognition. Some parts of it can be, however, used both by speech recognition and speech synthesis.
- Speech synthesis; synthesized speech is the main output. The speech synthesizer Demosthenes ([9]) is integrated in the communication module.
- Manager of standard input and output. Standard input (especially keyboard and mouse input) and output (e.g. graphics) can be combined to achieve as efficient communication as possible.

7. APPLICATION OF THE PROSODY DETECTION IN "DIALOG"

The present version of the system is based on speech recognition of isolated words. The standard type of the voice input has the form

statement argument_1 argument_2 ... argument_n

(Both the statement and arguments are isolated words). The number of the arguments may be also zero, some of the arguments can be omitted or have some default value. Even though the user modeling module tries to decide what form of help should be applied (if any) to the user, the particular optimal solution depends typically on the non-predictable knowledge of the user. Hence, it is reasonable to let this decision on the user.

Let us denote by '?' the question intonation (rising intonation), by '-' the level intonation (incompleteness) and by '.' falling final intonation (completeness). Then the following schemes explain the reaction of the dialogue system (U = user, S = system):

case 1:

U: statement?

S: starts the help dialogue, which explains the meaning of the statement and offers the assistance dialogue for setting the arguments.

case 2:

U: statement-

S: expects that the user knows the syntax and meaning and waits for a predefined time interval for setting the first parameter; if there is no reaction within the time interval the system starts the assistance dialogue for setting the arguments.

case 3:

U: statement.

S: generates the corresponding statement as a statement without parameters, or returns error message (if the statement requires arguments) and offers the assistance dialogue for setting the arguments.

In setting the arguments, the meaning of the intonation is analogous, i.e.:

case 1:

U: argument_i?

S: acknowledges, warns or explains the meaning of the argument according to the situation.

case 2:

U: argument_i-

S: provided the setting is correct, the system waits for a predefined time interval for setting the next argument; if there is no reaction within the time interval the system starts the assistance dialogue for setting the arguments.

case 3:

U: argument_i.

S: closes the arguments setting and generates the corresponding statement (provided the argument setting is correct, otherwise generates an error message and continues analogously).

This application of prosody detection makes the communication for the user more comfortable, even though the equivalent of the prosodic information could be conveyed to the system also via standard input (e.g. keyboard).

Technically, the determination of the relevant intonation is realized by segmentation into syllable segments and then by prosodic evaluation of the last segment. Here, the fundamental frequency and duration (which is obtained by segmentation) is used for evaluation, the mean intensity (energy) for normalization. The used method shows good results, it is however speaker dependent.

8. CONCLUSIONS AND FUTURE WORK

Using prosody features in dialogue systems is clearly a very effective way how to enhance the quality of the communication and also often the only possibility how to get the relevant information. For the continuous recognition and free dialogue in natural language the importance of the prosody is even higher, unfortunately the related problems are also much more complicated than in such relatively simple cases that were described in the paper.

In the next work we would like, as the first step, to free to some extent the syntax of the conversation of the developed dialogue system, and then gradually come over to free dialogue. Clearly, this is impossible without deeper understanding of prosody and without further progress in developing effective methods for automatic prosody detection.

ACKNOWLEDGEMENT

The author is grateful to K. Pala for reading the draft of the paper and valuable comments. The research has been partially supported by the Czech Ministry of Education under the Grant VS97028 and by the Grant Agency of the Czech Republic under the Grant 201/99/1248.

REFERENCES

1. Josifovski, L., Mihajlov, D., Gorgevik, D. "Speech Synthesizer Based on Time Domain Syllable Concatenation", *Proceedings SPECOM'97*, Cluj-Napoca, 1997, pp. 165-170.

2. Doddington, G. "Syllable Based Speech Processing", *WS97 Project Report, Research Notes No. 30*, J. Hopkins University, 1997.

3. Greenberg, S. "Speaking in Shorthand - A Syllable-Centric Perspective for Understanding Pronunciation Variation", *Proceedings of the Workshop Modeling Pronunciation Variation for Automatic Speech Recognition*, 1998, pp. 47-56.

4. Greenberg, S. "A Syllable-Centric Framework for the Evolution of Spoken Language", *Commentary on MacNeilage, P. The frame/content theory of evolution of speech production. Brain and Behavioral Sciences*, 21, 518.

1. Kopeèek, I. "Speech Synthesis Based on the Composed Syllable Segments", *Proceedings of the First Workshop on Text, Speech and Dialogue - TSD'98*, 1998, pp. 259-262.

6. Kopeèek, I.. "Automatic Segmentation into Syllable Segments", *Proceedings of First Int. Conference on Language Resources and Evaluation*, 1998, pp. 1275-1279.

7. Kopeèek, I. "Syllable Segments in Czech", *Proceedings of the XXVII. Mezhvuzovskoy naucznoy konferencii*, Vypusk 10,St. Petersburg, March 1998, pp. 60 - 64.

8. Kopeèek, I., Pala, K. "Prosody Modelling for Syllable-Based Speech Synthesis", *Proceedings of the IASTED Conference on AI and Soft Computing*, 1998, pp. 134-137.

9. Kopeèek, I., "Syllable Based Speech Synthesis", *Proceedings of the 2nd International Workshop SPECOM'97*, Cluj-Napoca, 1997, pp. 161-165.

10. Kopeèek, I. "Speech Synthesis of Czech Language in Time Domain and Applications for Visually Impaired", *Proceedings of 2nd SQEL Workshop*, Pilsen, 1997, pp. 141-145.

11. Kopeèek, I., "Modeling the Corti organ", to appear in *FI MU Report Series*.

12. Kopeèek, I., "Dialog Based Programming", in *Proceedings of the First Workshop on Text, Speech and Dialogue - TSD'98*, 1998, pp. 407-412.

13. Kopeèek, I., " Speech Recognition and Syllable Segments", to appear in *Proceedings of the Workshop TSD'99*.

14. Palková, Z. "Phonetics and Phonology of Czech" (in Czech); *Charles University*, Prague, 1994.

15. Shastri, L., Chang, S., Greenberg, S. "Syllable Detection and Segmentation Using Temporal Flow

Neural Networks" *in Proceedings of the International Congress of Phonetic Sciences*, San Francisco, 1999.