BMBF

Verbmobil
Verbundvorhaben

# A framework to evaluate and verify the presence of linguistic concepts in the prosody of spoken utterances

Gerit P. Sonntag
Thomas Portele

IKP Universität Bonn

Vm Report 177
September 1996

September 1996

Gerit P. Sonntag
Thomas Portele
Institut für Kommunikationsforschung und Phonetik
Universität Bonn
Poppelsdorfer Allee 47
53115 Bonn

Tel.: (0228) 7356 - 44
Fax: (0228) 7356 - 39
e-mail: {gso}@ikp.uni-bonn.de

**Gehört zum Antragsabschnitt:** 4.3

Die vorliegende Arbeit wurde im Rahmen des Verbundvorhabens Verbmobil vom Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMBF) unter dem Förderkennzeichen 01 IV 101 D 08 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei dem Autor.

# A framework for evaluating and verifying the presence of linguistic concepts in the prosody of spoken utterances

**Gerit P. Sonntag, Thomas Portele**
**Institut für Kommunikationsforschung und Phonetik (IKP), Universität Bonn**
**email: sonntag@ikp.uni-bonn.de / portele@ikp.uni-bonn.de**

## 1. Motivation

With recent developments in controlling the prosodic output of speech synthesizers[1], the quality of synthetic speech has improved considerably. However, determining the prosody required to convey specific linguistic concepts is still a largely unsolved problem. Concept-to-speech systems seem the most promising: additional information (structuring, focussing, affirmation/negation, quotation, enumeration, time/date, salutation, speaker attitude, etc.) is available to the prosody generation algorithm. This paper describes a method for determining which linguistic concepts are present in the prosody of a spoken utterance and which should therefore be taken into account when modelling prosody.

## 2. Methodological description

### 2.1. Idea

The discussions about prosodic functions are numerous[2,3,4]. One major problem for researchers is the separation of prosodic and segmental phenomena. In applications where there is no control over spectral qualities, such as time-domain concatenative synthesis systems, only prosodic parameters can be modified to convey linguistic concepts. To qualify and quantify the information contained in the prosody alone, we propose specially designed perception tests. The segmented information in the stimuli is removed, hence we can be sure that all information is carried by the prosody alone.

### 2.3. Choice of stimuli

With the proposed manipulation (described in section 2.4.), the stimuli to be used need not have a neutral content. The stimuli can be chosen in order to evaluate a specific linguistic concept without the possibility of the listener obtaining crucial information not from the prosody but from the content transported by the segmental information. Many previous experiments on prosody have been forced to employ ambiguous test sentences or words which is clearly suboptimal. With our method the semantic content of the stimuli becomes irrelevant to the test results and the optimal stimuli for a given task can be used.

### 2.4. Stimuli manipulation

The pitchmarks of the stimuli are used to construct an *excitation signal* preserving the energy of the original signal (see Fig.1). The manipulated stimuli contain only prosodic information: F0 contour, temporal structure and energy distribution. Thus, they reflect exactly the parameters that can be varied using the PSOLA method[1]. Similar ways of reducing the speech signal to its prosodic content are given in [5,6,7].

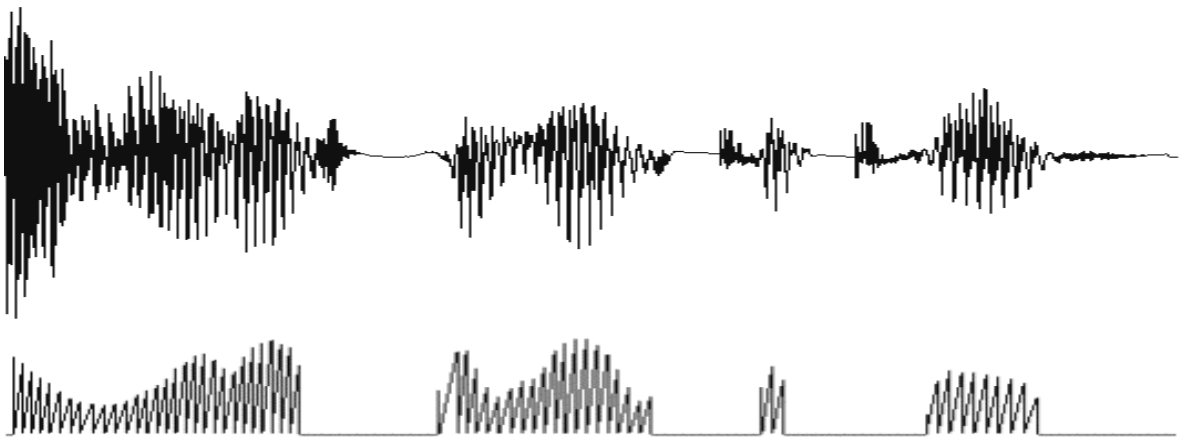

Fig.1 Example of *excitation signal* (bottom) extracted from original (top)

2.5. Test procedure

Depending on the aim of the investigation, the manipulated stimuli are presented either with or without the original sentence in writing (cf. 4.). The questions the subject has to answer can be very simple, aimed directly at the linguistic function in question. There is no need to instruct the subject to listen only to the prosody, as he/she will hear nothing else.

2.6. Results

The reliability of the test results does not depend on the listeners ability to concentrate solely on the prosody as is the case when evaluating original utterances, nonsense sentences or utterances consisting of nonsense words. The results can be based on a large number of stimuli rather than be restricted to the particularities of only a few, because there are no semantical limitations to generating more stimuli.

## 3. Examples of tests carried out with the proposed method

To show how the proposed test method can be applied in various domains of interest and what has to be taken into account in each case we describe a few tests that have already been carried out.

3.1. Emotions

In a test aimed at identifying the emotional content (e.g. fear, joy, anger, disgust, sadness) from just the prosodic properties (namely F0, duration and energy), speech signals resynthesized with a concatenative system yielded the same poor results as the *excitation signals*[8]. It is obvious that in this case, where the naturalness of an utterance depends on features that are not readily controllable by the synthesis system (e.g. aspiration, creaky voice etc.). A test procedure with resynthesized speech will not improve the results that have been obtained with the *excitation signals.* All the parameters that are used for the resynthesis are present in the *excitation signal*.

3.2. Evaluation of an automatic parametrisation of F0 contours

A method to parametrize F0 contours[9] was evaluated by asking the subjects whether they could hear a difference between the original signal and a signal with an automatically generated F0 contour[10]. In one test, utterances with manipulated prosody were compared to the original utterances. In a second test, *excitation signals* of the original and the parametric F0 contours were compared. In both tests the original contour was compared to the same automatically generated F0 contour.

The results of the two tests were compared (see Fig.2). The comparison showed that the judgements made on the *excitation signals* correlated more closely to the acoustic parameters. The judgements on the manipulated utterances were partly mislead by some errors inherent in the manipulation process done with PSOLA (see Fig.3).

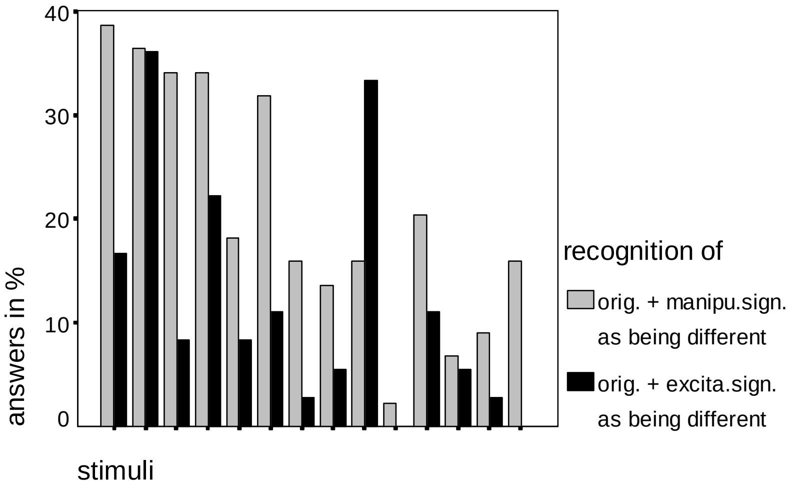

Fig.2 Signals perceived as different across stimuli

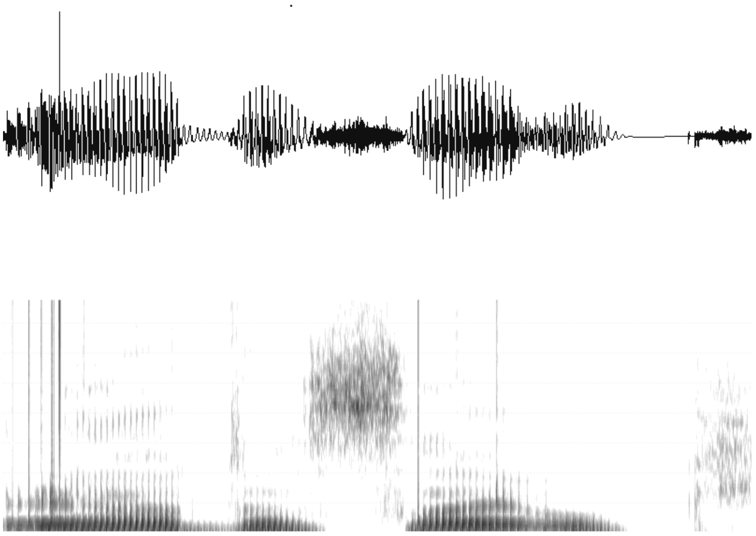

Fig.3 Example of the poor spectral quality in the manipulated utterances

## 3.3. Temporal structure

If the stimuli are carefully chosen, even the temporal structure of an utterance can be evaluated using the *excitation signal.* To avoid a continuous sound signal, the rhythm of an utterance can be made explicit by choosing stimuli in which all syllables are deliminated by voiceless segments, thus structuring the audible *excitation signal.* In two different tests, subjects were asked to assign perceived accents to a syllable within a short utterance. Both tests were carried out using the *excitation signal.* One test used only completely voiced utterances, the other only stimuli with voiceless segments between each syllable. After the first test, subjects complained that, even though they clearly perceived accents within the utterances, they found it impossible to place them on a specific syllable. They found the second test much easier.

## 3.4. Evaluation of naturalness of a speaker's prosody

An assessment of natural speech has been successfully carried out using the *excitation signal* alone[11]. The utterances of two male speakers were evaluated for naturalness. Applying the *magnitude estimation* technique[12], the *excitation signals* yielded results that significantly distinguished the two speakers (see Fig.4). The success of this method for natural speakers justifies the application with synthetic prosody, where the differences are usually much larger. Such a test was performed[13], and distinct differences could be measured.

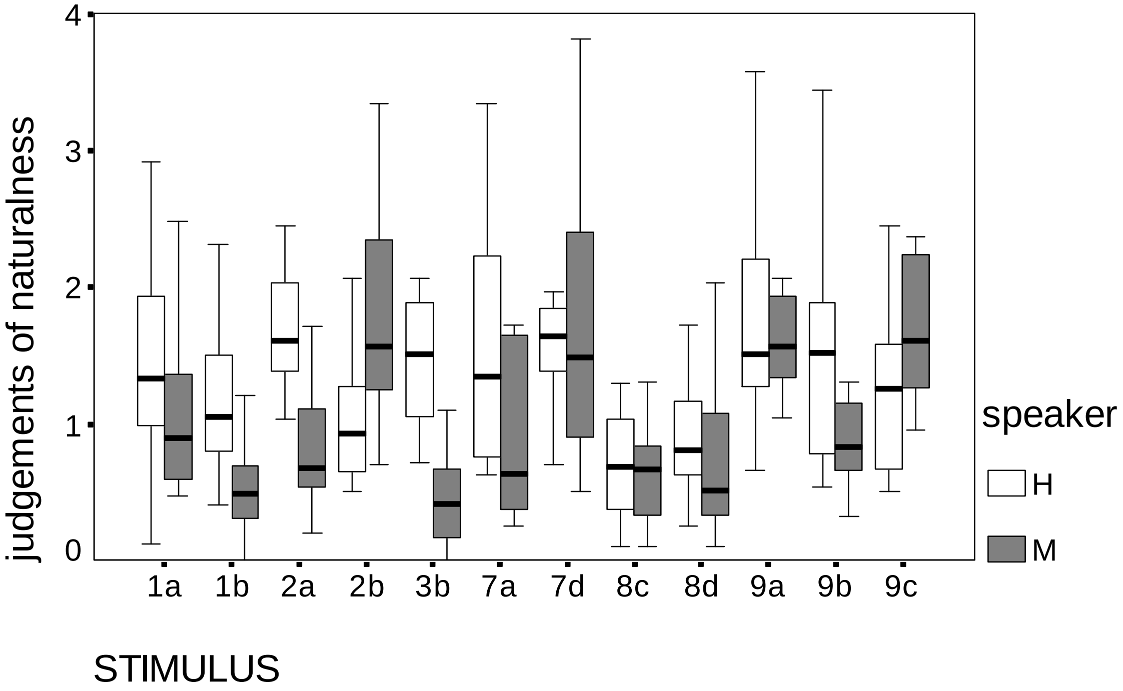

Fig. 4 Speaker H's prosody was considered to be significantly more natural than speaker M's prosody (Wilcoxon Rank Test: p<0.0001).

3.5. Classification of syntactic structures

To show that prosody transports information about the syntactic structure of a sentence, subjects were asked to assign one of several given syntactic structures to the presented *excitation signal*[11]. The possible syntactic structures were represented by written sentences, one of which had the same syntactic structure as the stimulus. These sentences differed from the utterances that served as the source for the test stimuli (see Fig.5). Being asked to pick out the sentence they were hearing, the subjects believed that what they heard was the written sentence, showing that their decision was based solely on prosody.

example of a test item:
stimulus presented as excitation signal: *"On the ancient counter lies a journal ."*
answering sheet: [ ] The smallest baby is in the cabin.
[ ] In the cabin is the smallest baby.
[ ] The baby is in the smallest cabin.
[ ] In the smallest cabin is the baby.

Fig.5 Example of a presented stimulus and the possible answers

**4. General remarks**

It is important to mention that the test design may vary according to the phenomena under observation. It is easier for the subjects to read the sentence while they listen to the *excitation signal* so that they can mentally connect the visual and aural stimulus. However, for other test procedures like the example dealing with the temporal structure, the subjects must not know about the original sentence, because their judgements must not be influenced by the semantic context of the stimulus. The example of the test procedure on syntactic structures shows that combining the *excitation signal* with written sentences differing from the original ones opens a variety of interesting possibilities for an individual test design. But it also draws the attention to the importance of carefully chosing the stimuli.

The *excitation signal* used so far is a sawtooth signal, which is a crude approximation to the glottal waveform. We are currently working on improving the quality of the manipulated stimuli by using the Liljencrants-Fant model[14] in the manipulation procedure. Better stimuli quality will impose less listening effort on the subjects.

**Acknowledgements**

This work has been funded by the German Federal Ministry of Education, Science, Research and Technology in the scope of the VERBMOBIL project under Grant 01 IV 101 G. Further work in this project will aim at developing a multilingual concept-to-speech model. We would like to thank Jürgen Krämer for his help in programming.

## References

[1]Moulines, E.; Charpentier, F. (1990): "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones." Speech Communication 9, pp.453-467
[2]Barry, William J. (1981): "Prosodic functions revisited again!", Phonetica 38, pp.320-340
[3]Léon, P.R. (1970): "Systématique des fonctions expressives de l'intonation", Léon (eds.) pp.57-74
[4]Kohler, K.J. (1987): "The linguistic functions of F0-peaks", PICPS 11, vol.3, pp.149-152
[5]Ohala, J.J.; Gilbert, J.B. (1979): "Listeners' ability to identify languages by their prosody" in: Problèmes de Prosodie, Léon, P.; Rossi, M. (eds.), pp.123-131
[6]Nicolas, P.; Roméas, P. (1993): "Evaluation of prosody in the French version of a multilingual text-to-speech synthesis: neutralising segmental information in preliminary tests", Eurospeech'93, Berlin, pp.211-214
[7]Mersdorf, J. (1996): "Ein Hörversuch zur perzeptiven Unterscheidbarkeit von Sprechern bei ausschließlich intonatorischer Information", Fortschritte der Akustik - DAGA'96, Bonn, pp.482-483
[8]Heuft, Barbara; Portele, Thomas; Rauth, Monika (1996): "Emotions in time-domain synthesis", Proc. ICSLP'96, Philadelphia (to appear)
[9]Heuft, B. (1995): "Parametrich description of F0-contours in a prosodic database", ICPhS'95, vol.2, Stockholm, pp.378-381
[10]Heuft Barbara; Streefkerk, Barbertje; Portele, Thomas (1996): "Evaluierung der Parametrisierung von Grundfrequenzkonturen", Elektronische Sprachsignalverarbeitung VII, Berlin (submitted)
[11]Sonntag, Gerit P. (1996): "Klassifikation syntaktischer Strukturen aufgrund rein prosodischer Information", Fortschritte der Akustik - DAGA'96, Bonn, pp.480-481
[12]Pavlovic, C.V. (1990): "Use of the magnitude estimation technique for assessing the performance of text-to-speech synthesis systems", JASA 87(1) Jan.'90, pp.373-382
[13]Reuter, A. (1996): "Generierung prosodischer Parameter unter Verwendung eines neuronalen Netzes", thesis, Bonn university
[14]Fant, G.; Liljencrants, J.; Lin, Q. (1985): "A four-parameter model of glottal flow", STL-QPSR 4/85,pp.1-13