Looking for the presence of linguistic concepts in the prosody of spoken utterances

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

Abstract

This paper describes an experimental method for detecting prosodic functions. We assume that the first step towards content driven synthetic prosody generation (Concept-to-speech) is invariably to determine the perceptually relevant prosodic features. The proposed method has been applied to the detection of syntactic structure, dialogue acts and given/new distinction. First results are being discussed.

1 Motivation

Within the framework of Concept-to-speech, our aim is to integrate additonal information such as structuring, focus, dialogue act, speaker attitude into the prosodic modelling of synthesized utterances. The first step into this direction is to find out which additional information is reflected in the prosodic structure. An investigation as to what information about the content of an utterance is actually encoded in the prosody and how this coding is realized by the natural speaker is inherently necessary [Lib74]. The aim for more natural prosody generation can only be determined by an adequate description of human prosody and its interaction with content information. In this paper we propose a method determining which linguistic concepts have a functional influence on prosody. Prosody is known to be of a very complex nature, yet we cannot per se suggest that every communicative function is relevantly encoded it in. At least we have to distinguish between functions that necessarily pertain to the prosodic structure, and those that are not identifiably located within prosody. Once the relevant concepts have been found, their influence on the acoustic parameters related to prosody can be investigated.

2 Methodological description

2.1 Idea

Prosodic function has been discussed frequently, e.g. [Bar81,Leo70,Koh87]. One major problem is the separation of prosodic and segmental influences. In applications with 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, we use specially designed perception tests. The segmental information in the stimuli is removed, in order to make sure that all information is carried by the prosody alone.

2.2 Choice of stimuli

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. Also, the stimuli can be extracted from a read text or from a natural dialogue situation, as long as the quality of the recording is not too degraded.

2.3 Stimuli manipulation

A stimulus is constructed on the basis of the points of glottal excitation (pitchmarks) of the original signal while preserving the energy. 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 PSOLA [Mou90]. Different stimulus manipulation methods have been compared in the validation test series (3).

2.4 Test procedure

Depending on the aim of the investigation, the manipulated stimuli are presented either with or with-

out the original sentence in writing, and either with or without visual representation. The proposed method is not tied to a specific test setting. Various examples of successful test procedures are reported in this paper, and more settings can easily be developed. 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.

2.5 Results

The reliability of the test results does not depend on the listener's 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 Validation test series

Several methods for speech delexicalisation can be found in the literature [Kre82,Pas93,Leh79,Mer96, Oha79,Pij94,Sch84]. The aim of all these manipulations is to render the lexical content of an utterance unintelligible, while leaving the speech melody and temporal structure intact. We think that the ideal stimulus manipulation for prosodic perception tests should meet three main requirements:

- it should clearly convey the primary prosodic functions (i.e. accentuation, phrasing and sentence modality)
- the detection of these phenomena should not require too much listening effort from the test subject
- the manipulation procedure should be simple and quick

We compared six methods of delexicalisation according to these criteria. Subjects had to complete four different tasks. They were questioned after each task which of the six different stimulus versions they found easiest for the task, most difficult for the task, most pleasant and least pleasant. Learning effects are negligible because the presentation order was changed for each subject.

3.1 Stimuli manipulation

All the stimuli referred to in this paper were digitally recorded in an anechoic chamber with 16kHz and 16bit. The following six manipulation methods were compared:

CCITT The extracted pitchmarks of the original signal were filled with an excitation signal proposed by the CCITT [CIT89], and also low-pass filtered.

F0 filt The original signal was low-pass filtered using a time variant filter with a cut-off frequency just above F0. At unvoiced segments within the signal the cut-off frequency was automatically set to zero.

inv A combination of spectral inversion and filtering proposed by [Kre82]. After high-pass filtering at 600Hz, the signal is spectrally inverted, then low-pass filtered at 4000Hz and then added to the residual of the original signal low-pass filtered at 200Hz. The resulting signal preserves the voiced / unvoiced distinction and is the most intelligible of the versions compared.

lfm The extracted pitchmarks of the original signal were filled with the Liljencrants-Fant model [Fan85] of glottal flow.

saw A simple sawtooth signal was inserted into the extracted pitchmarks.

sin The pitchmarks were filled with a sinus with a first harmonic of 1/4 of the amplitude and a second harmonic of 1/16 of the amplitude.

Other ways of rendering an utterance unintelligible, such as [Pij94,Pag96], were not included as we tried to keep the effort for stimuli manipulation as low as possible.

3.2 Counting of syllables

In the first test session 18 subjects were asked to count the number of syllables of 12 short sentences aurally presented in the different manipulated versions. The stimuli were chosen out of five different sentences (5-8 syllables of length) spoken by a female speaker and manipulated with the six different procedures described above. Out of these stimuli two sentences per version were used for syllable counting while the rest was used for the accent assignment task. As this was an open response task, there is no referential chance level as in the other tests. The results show that the syllable number of nearly 60% of all stimuli can be determined exactly with the proposed method, at least at sentence level (Fig. 1). In 86% of all cases, the correct number of syllables plus/minus one were detected.

3.3 Phrase accent assignment

The same subjects then listened to the other 18 sentences (six versions in three different sentences) to

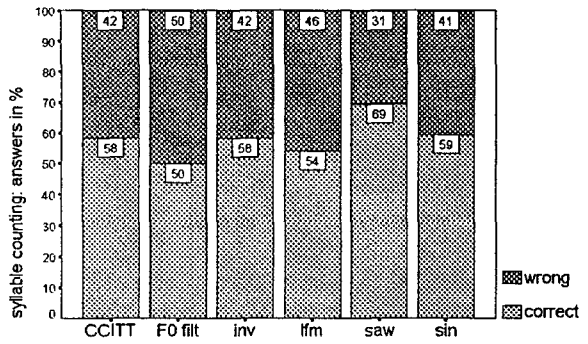

Figure 1: *Results of the syllable counting task for the differently manipulated stimuli.*

assign a phrase accent to a syllable. Again presentation order differed from subject to subject. Now, they could see a cursor moving along an oscillogram of the current phrase, where each syllable boundary was marked. This combination of aural and visual presentation was chosen to make sure that the subjects' ability to count syllables was not tested again. To avoid any influences of the visual amplitude differences between the syllables on the subject's choice, the stimuli had been adjusted to have a more or less equal energy distribution over the whole phrase. We thus reduced the intonational information by the energy factor. The results appear to confirm that this is the least important factor [Fry58] within prosodic perception. In 73.4% of all cases the phrase accent was correctly assigned (Fig. 2). Some of the subjects reported that the possibility of relating the perceived accent to a visual cursor position helped a lot. Others, who seemed to have no problems with the syllable counting task, said that they were rather confused by the visualization.

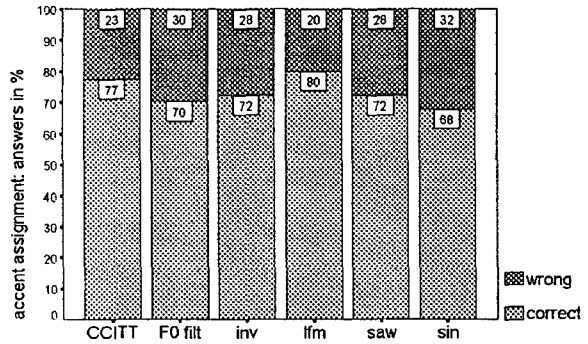

Figure 2: *Results of the accent assignment task for the differently manipulated stimuli.*

3.4 Recognition of phrase modality

16 subjects were presented with three phrases recorded from a male speaker and pronounced in three different modalities: terminal, progredient (i.e. continuation rise) and interrogative [Son96a]. Each subject listened to 32 stimuli chosen randomly from these nine phrases manipulated by the six procedures and decided on one of the given modalities. The result was highly significant: 84% of the stimuli were correctly recognized (Fig. 3).

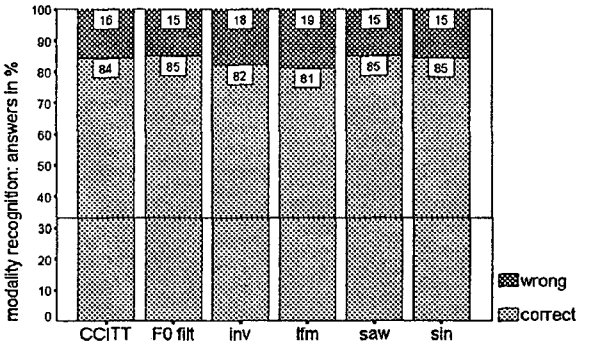

Figure 3: *Results of the modality recognition task for the differently manipulated stimuli. The line indicates chance level.*

3.5 Phrase boundary detection

12 subjects were asked to place two phrase boundaries in 20 manipulated stimuli with the additional help of visual presentation. Four different sentences (12-20 syllables) had been read by a female speaker, all containing two syntactically motivated prosodic boundaries. The visual signal contained markers at each visible syllable boundary which served as possible phrase boundary location. As there were 15 possible boundaries per sentence in the mean, chance level can be calculated as being around 6.6%. All stimuli were checked whether they contained a visually obvious pause at the boundaries. These pauses were manually eliminated. Even though this meant that the most important clue for boundary detection [Leh79] was eliminated the subjects managed a significantly correct detection in 66.6% of all stimuli (Fig. 4). One of the two boundaries was correctly placed in 90% of the cases.

3.6 Choice of stimulus manipulation

All four tasks yielded correct results. It was surprising that the error rate for the differently manipulated stimuli did not significantly differ, neither within a task nor over all. So the decision which manipulation procedure to prefer can only be

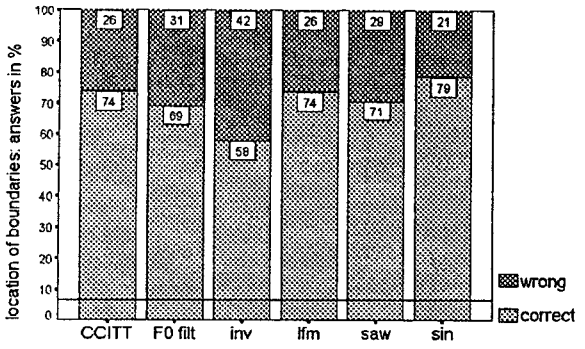

Figure 4: *Results of the prosodic phrase boundary location task for the differently manipulated stimuli. The line indicates chance level.*

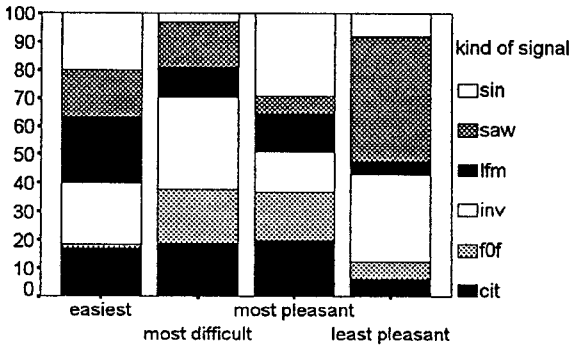

Figure 6: *Subjects' answers to the four questions: which of the signal did you find a) easiest? b) most difficult? c) most pleasant? d) least pleasant?*

based upon the subjective evaluation of the pleasantness. As the differences between the tasks are small enough, we compare the subjects' opininions over all tasks (Fig. 5). The least "easy" version was the one filtered at the fundamental frequency. The sinusoidal signal and the signal after the Liljencrants-Fant model were "not difficult". "Most comfortable" was the CCITT excitation signal, the signal filtered at F0 and the sinoidal signal. The spectrally inverted signal and the sawtooth excitation signal were judged "least comfortable". All these differences were significant ($p<0.05$). All in all we conclude that the sinoidal signal is the most appropriate one (Fig. 6). Our findings confirmed the results about the pleasantness of manipulated signals in [Kla97].

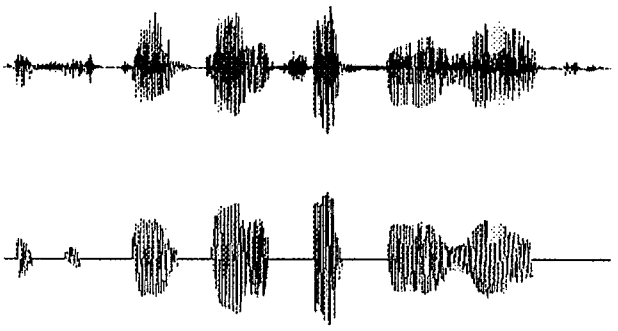

Figure 5: *Comparison of an original utterance (on top) and the manipulated sinusoidal signal (below).*

4 Examples of tests carried out to detect prosodic concepts

The first two tests described here (emotions and syntactic structure) took place before the comparison of stimulus manipulation methods. Therefore they have been carried out using the sawtooth excitation signal. In the latter two tests (dialogue acts and given/new), the sinusoidal signal manipulation described in 2.3 was used.

4.1 Emotions

In a test aimed at identifying the emotional content (e.g. fear, joy, anger, disgust, sadness) from the prosodic properties only, speech signals that were resynthesized with a concatenative system yielded the same poor results as the delexicalized stimuli [Heu96]. Both stimuli gave results that were at chance level. It is obvious that in this case, where the naturalness of an utterance depends on features that are not readily controllable by time-domain 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 delexicalized stimuli, because all the parameters that are used for the resynthesis are present in the delexicalized stimuli.

4.2 Syntactic structure

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 delexicalized stimuli [Son96b]. 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. 7). Asked to pick

Figure 7: *Example of a presented stimulus and the possible answers.*

example of a test item:
stimulus presented as excitation signal:
"Auf der alten Theke steht der Eintopf."
answering sheet:

Die kleine Katze liegt in der Truhe.
In der Truhe liegt die kleine Katze.
Die Katze liegt in der kleinen Truhe.
In der kleinen Truhe liegt die Katze.

out the sentence they were hearing, the subjects believed that what they heard was the written sentence, which shows that their decision was based solely on prosody. Stimuli of one male speaker were correctly classified in 80% of all cases. A professional male speaker with very elaborate speaking style yielded 67% of correct answers.

4.3 Dialogue acts

The motivation for this test was to decide whether different dialogue act types have a perceivable influence on the prosodic structure of an utterance. Within the VERBMOBIL project, dialogue act types from the domain of appointment scheduling dialogues are used [Rei95]. If these dialogue act types have specific prosodic forms, then the synthesis module should generate them accordingly.

For a first approach we chose to evaluate the four dialogue act types:

- affirmation
- negation
- suggestion
- request

For each dialogue act type, eight sentences were read by a male and a female speaker. For affirmation and negation, only statements were chosen (length: 1-10 syllables), and four questions and four answers for suggestion and request (length: 6-14 syllables). The resulting 64 sentences were manipulated and randomly presented to ten subjects who had to assign one of the four dialogue act types to each sentence. Although each subject remarked that this was a pretty difficult task, their answers were significantly ($p<0.001$) above chance level (Fig. 8). What seemed more difficult than relating the utterance to an abstract internal reference was the fact that the two speakers' utterances were presented in random order. They differed remarkably not only as to their fundamental frequency but also to their expressive strategies. Whereas the male speaker was more often thought to sound negating, the female speaker was mostly recognized as being requestive. Also, dialogue acts spoken by the female speaker were recognized significantly better as those spoken by the male. This indicates the degree to which the interpretation of a linguistic concept depends on the speaker's personality and should be taken into account whenever speaker adaptation of the synthetic output is desired. Perception tests should always take into account the subjects' comments on the completed task. This can yield very useful but often neglected extra information. The subject (no. 10 in Fig. 9) who scored better than the others explained his strategy. To distinguish between affirmation/negation on the one hand and suggestion/request on the other, he assumed that in the former, the focused part of the utterance lies at the very beginning of the utterance, whereas in the latter, the second half of the utterance should bear more focus. Whether this assumption can be generalized or not has to be investigated in further perception tests.

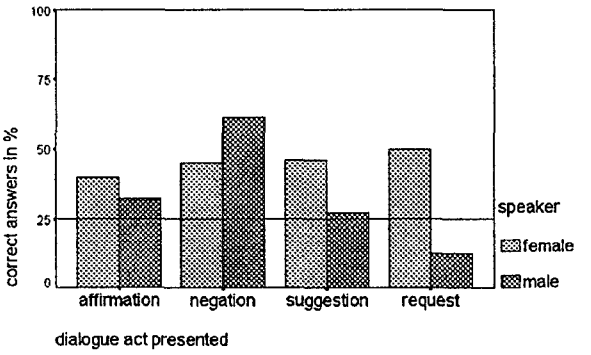

Figure 8: *Results of the dialogue act recognition task for each presented act. The line indicates chance level.*

4.4 Given/new

As an extension of the phrase accent assignment test we tested the accuracy with which subjects perceive differently focussed parts within a delexicalized utterance. The stimuli consisted of eight sentences of a new/given structure and eight sentences of a given/new structure of different length. They were read by a female and a male speaker as possible answers to a question, then manipulated and presented in random order. The 'given' part was always a rephrasing of a part of the question. Ten subjects were given a short explanatory text with an example and then asked to decide in which order the different

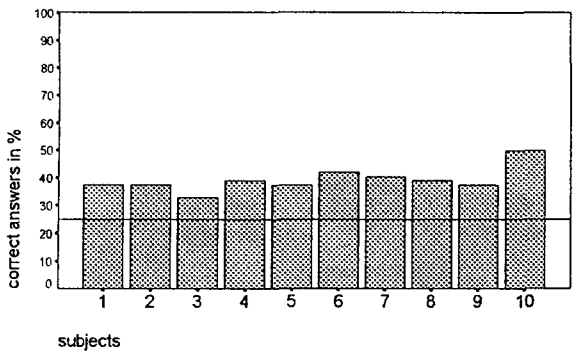

Figure 9: *Results of the dialogue act recognition task for each subject. The line indicates chance level.*

parts appeared within the utterance and where the boundary between the two parts was located. The task was supported by an oscillogram of the stimulus containing four marks as possible boundary locations. As in Section 3.3, the energy distribution over the whole sentence was smoothed. Some subjects claimed that the location task was easier than the order recognition task. The order recognition task was correctly completed in 78%, the boundary was correctly located in 62% (Fig. 10). Both tasks were significantly ($p<0.001$) completed over chance level, yet some inter-subject differences were also significant. The subjects located the 'new' part significantly ($p<0.002$) more often at the beginning of the sentence, which can be explained by intonational downstep.

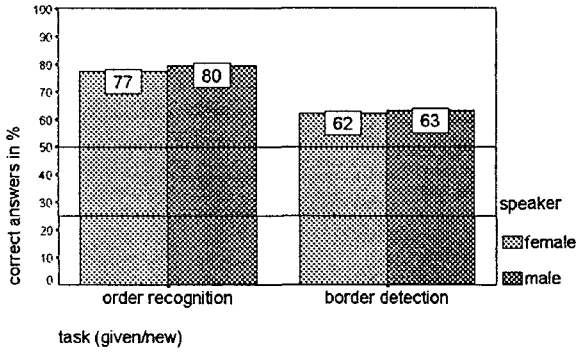

Figure 10: *Results of the order recognition task (chance level=50%) and the boundary location task (chance level=25%) for each speaker.*

5 Conclusion

We have shown that the proposed method stands up to the three requirements. It significantly conveys prosodic functions and no segmental information, a reasonably pleasant signal manipulation was found and the manipulation is easy, so that most preparatory effort can go into the choice of stimuli and the test design. The test design is variable and can be adequately set for the phenomenon under investigation. The problem of localizing a certain part of an utterance has been tackled by visual presentation. The visual presentation should still be improved so that it does not show pauses or energy distribution. The mixture of different voices within one test seems to degrade the results. It is desirable to check the findings with more different voices. A separate test run for each voice should faciliate the task as it enables the subject to get used to the individual speaker properties.

Some of the subjects' side comments have allowed an interesting insight into their listening strategies. We think that the proposed method is an efficient link between linguistic theory and practical application. On the one hand theoretical assumptions within Concept-to-speech have to be validated in an actual application. On the other hand perception tests of the kind we have described them can lead to new theoretical findings.

The method is being applied to detect prosodic content information in dialogue situations of the domains appointment scheduling, hotel reservation and tourist information within the German VERBMOBIL project. Once more reliable information about what can be perceived from the prosody has been collected, the interplay of the correlating acoustic parameters will be investigated. Finally the findings will be implemented and evaluated again.

This work has partly 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.

References

Bar81] Barry,W.J. (1981): "Prosodic functions revisited again!" in: Phonetica 38, pp.320-340

CIT89] CCITT, Blue Book, Vol.V, Telephone Transmission Quality, Series P Recommendations, IX. Plenary Assembly, Geneva 1989, Recommendation P.50, pp.87-98

Fan85] Fant,G.; Liljencrants,J.; Lin,Q. (1985): "A four-parameter model of glottal flow." STL-QPSR 4/85,pp.1-13

Fry58] Fry,D.B. (1958): "Experiments in the perception of stress." in: Language and Speech 1, pp.126-152

Heu96] Heuft,B.; Portele,T.; Rauth,M. (1996): "Emotions in time-domain synthesis." Proc. ICSLP'96, Philadelphia, pp.1974-1977

Kla97] Klasmeyer,G (1997): "The perceptual importance of selected voice quality parameters." Proc. ICASSP'97, Munich, vol.3, pp.1615ff

Koh87] Kohler, K.J. (1987): "The linguistic functions of F0-peaks." in: Proc. ICPhS 11, Tullin, vol.3, pp.149-152

Kre82] Kreimann,J. (1982): "Perception of sentence and paragraph boundaries in natural conversation." in: Journal of Phonetics 10, pp.163-175

Leh76] Lehiste,I.; Wang,W.S-Y. (1976): "Perception of sentence boundaries with and without semantic information." in: Dressler,W.; Pfeiffer,O. (eds.), Phonologica 19, Innsbruck, pp.277-283

Leh79] Lehiste,I. (1979): "Perception of sentence and paragraph boundaries." in: Lindblom,B.; Öhman,S.(eds.) Frontiers of speech communication research, Academic Press, NY, pp.191-201

Leo70] Léon,P.R. (1970): "Systématique des fonctions expressives de l'intonation." in: Léon (eds.) Prosodic feature analysis, pp.57-74

Lib74] Liberman,M.; Sag,I. (1974): "Prosodic form and discourse function." in: Papers from the Tenth Regional Meeting, Chicago Linguistic Society, pp.416-427

Mer96] Mersdorf,J. (1996): "Ein Hörversuch zur perzeptiven Unterscheidbarkeit von Sprechern bei ausschließlich intonatorischer Information." in: Fortschritte der Akustik - DAGA'96, Bonn, pp.482-483

Mou90] Moulines,E.; Charpentier,F. (1990): "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones." in: Speech Communication 9, pp.453-467

Oha79] Ohala,J.J.; GilbertJ.B. (1979): "Listeners' ability to identify languages by their prosody." in: Léon,P./Rossi,M. (eds.), Problêmes de Prosodie, Studia Phonetica 18, pp. 123-131

Pag96] Pagel,V.; Carbonell,N.; Laprie,Y. (1996): "A New Method for Speech Delexicalization, and its Application to the Perception of French Prosody." in: Proc. ICSLP'96, Philadelphia

Pas93] Pascale,N.; Roméas,P. (1993): "Evaluation of prosody in the French version of a multilingual text-to-speech synthesis: neutralising segmental information in preliminary test." in: Proc. Eurospeech'93, Berlin, pp.211-214

Pij94] de Pijper,J.R.; Sandermann A. (1994): "On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues." in: Journal of the Acoustical Society of America 96 (4), pp.2037-2047

Rei95] Reithinger,N.; Maier,E. (1995): "Utilizing Statistical Speech Act Processing in VERBMOBIL." in: Proc. ACL 33, Cambridge, MA

Sch84] Schaffer,D. (1984): "The role of intonation as a cue to topic management in conversation." in: Journal of Phonetics 12, pp.327-344

Son96a] Sonntag,G.P. (1996): "Untersuchung zur perzeptiven Unterscheidung prosodischer Phrasen." in: ITG Fachtagung Sprachkommunikation, 17./18.9.96, Frankfurt am Main, pp.121-124

Son96b] Sonntag,G.P. (1996): "Klassifikation syntaktischer Strukturen aufgrund rein prosodischer Information." Fortschritte der Akustik - DAGA'96, Bonn, pp.480-481