Timing restrictions on prosodic phrasing

Horne, Merle; Frid, Johan; Roll, Mikael

Published in:
Nordic Prosody IX

2006

Document Version:
Early version, also known as pre-print

TIMING RESTRICTIONS ON PROSODIC PHRASING

Merle Horne, Johan Frid, Mikael Roll

1 Background

Within our current speech technology research project (www.ling.lu.se/projects/ProSeg2.html), we have been investigating the role of function words in the processing of spontaneous speech (Horne et al., 2003, Horne et al., in press)). In this work, we have been influenced by speech processing models stemming from psycholinguistic research, in particular the work of Clark and Wasow (1998).

It has been pointed out for example that function words often occur "stranded" before hesitation pauses, during which time speech planning takes place. Furthermore, the phonetic form of function words before hesitations has been observed to be very marked in comparison to their form in fluent speech. Since it is is also known that speakers tend to resume production of the constituent that was being planned at the time of the hesitation (cf. Clark and Wasow's 'Commit and Restore' model), it is obvious that the salient phonetic form of function words together with following pauses can be used in the development of algorithms for segmenting speech into constituent-like units, since function words occur at the left-edge of constituents (i.e. conjunctions introduce clauses, prepositions introduce prepositional phrases, etc.). Thus one of our major goals in the project has been to investigate the segmental and prosodic form of function words before hesitations as compared to their form in fluent speech in order to be able to relate this difference to different discourse contexts. Moreover, following Clark and Wasow's 'complexity hypothesis', it is also assumed that the probability that speakers will hesitate in speech production is increased the more complex the constituent being planned is. Thus another goal of the project has been to compare the difference in syntactic complexity between speech fragments after function words in hesitation contexts and in fluent speech.

Complexity can be measured in terms of a number of parameters related to the lexico-syntactic structure of a given clause, e.g. number of words, number of nodes, number of phrases and the depth of the syntactic tree being planned. An example of an utterance containing three instances of stranded function words

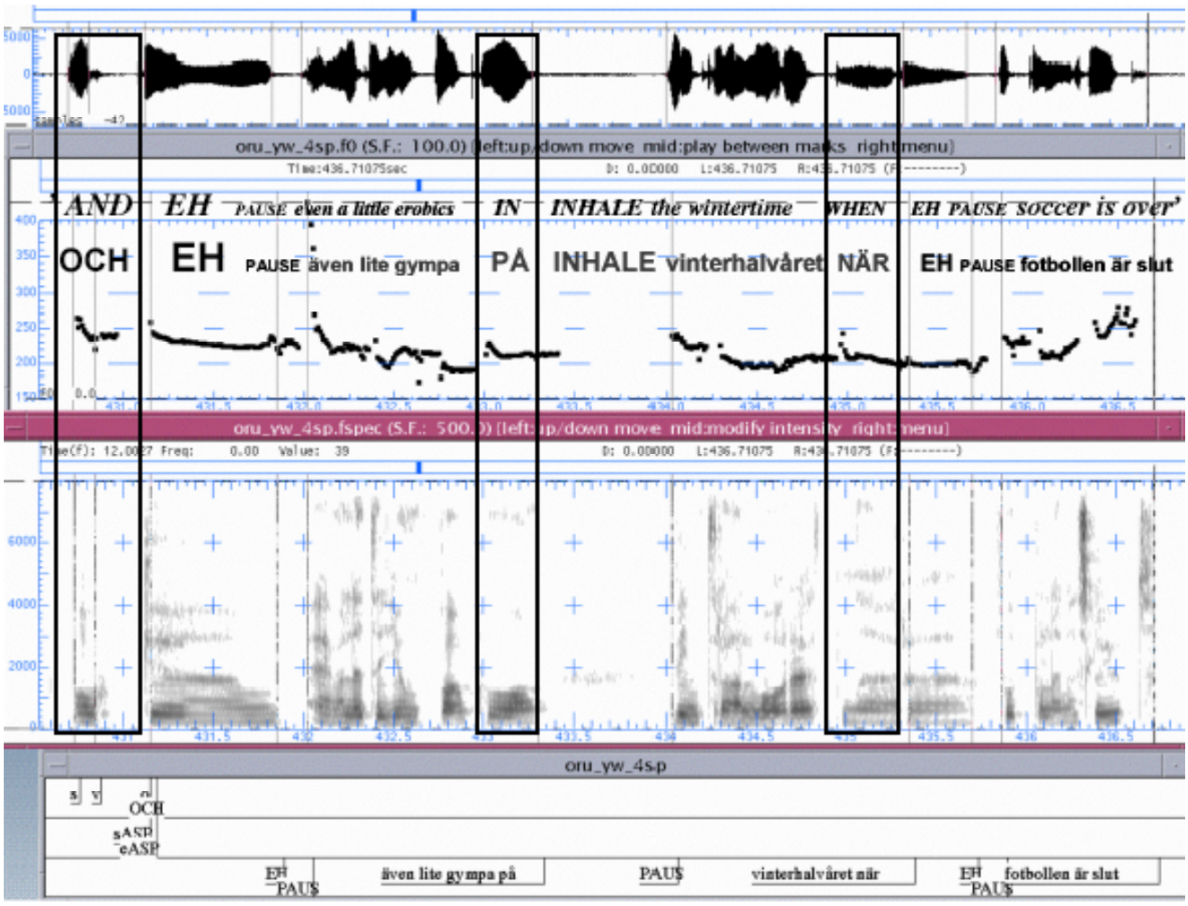

Figure 1. An example of an utterance containing three examples of hesitations after the function words och *'and',* på *'in' and* när *'when'. All three function words are segmentally unreduced and followed by pauses, either filled (EH), silent (PAUSE) or containing an inhalation (INHALE).*

(*och* 'and', *på* 'in' and *när* 'when') can be seen in Figure 1. According to Clark et al.'s 'Commit and Restore' model of speech production, stranded function words signal that the speaker intends to produce a constituent of the kind signalled by the kind of function word produced, e.g. a clause after a stranded conjunction, a prepositional phrase after a preposition, etc. Thus the recognition of stranded function words can be expected to be important for parsing algorithms.

2 Timing restrictions in speech production

During investigations into the linguistic form of the speech fragments following hesitations, it has also been observed that they seem to be grouped into units that have a relatively constant duration. It is this observation that has lead us to conduct a more detailed study of our spontaneous speech data in order to find support for this

observation. If shown to be a stable characteristic of spontaneous speech, it should be valuable knowledge in developing algorithms for spoken language processing.
Support for the hypothesis that there are in fact timing restrictions on the coding of speech has been seen in the literature from various areas of investigation. For example, in the area of memory research, Baddeley (1997), in his work on working memory, has claimed that the part of working memory where speech processing takes place ('inner speech') has a time limit of around 2 seconds. In his investigation, Baddeley used recall of digit spans as a memory task after performing a reasoning test; the trace decay time was found to be around 2 seconds. According to Baddeley, the number of items recalled will be a function of how long they take to articulate.

In the area of neurolinguistics, the notions of timing restrictions on speech planning has been reported for example by Ackerman and Hertrich (2003), who maintain that the temporal organization of inner speech is controlled by the cerebellum (the 'internal clock').

Discussions of timing constraints in linguistic research can be found in studies on speech rhythm; for example, Fant and Kruckenberg (1996) have focussed on the duration of inter-stress stretches of read speech. Empirical investigations on timing restrictions on larger speech production chunks/information units has not, to our knowledge, been the object of detailed investigation. Sigurd (1983), however, assumes speech chunks of one to two seconds length in his message-to- speech model and Chafe (1994) hints at some kind of timing restrictions on speech production when he notes that a speaker's 'focus of consciousness' is replaced by another idea at roughly one- to two-second intervals. We would like to build on these ideas and to hypothesize that speech planning units are between 2-2.5 seconds long. We further hypothesize that they can contain internal silent and/or filled pauses, but not pauses containing inhalations. The 2-2.5 second production units that we are invisaging can be thought to correspond to the ouput of the linguistic *Formulator* in Levelt's (1989) model of speech production. In this model, one can expect that pauses internal to production units can be related to e.g. lexical access time. (It should be pointed, however, that Levelt does not assume any timing restriction on speech coding in his model).

Our evidence for this assumed timing restriction on speech production comes mainly from observations of prosodic phenomena associated with units of speech that are ca. 2-2.5 seconds long. The following observations are based on the analysis of about 25 minutes of speech from two speakers:

- There is very often an inspiration after 2-2.5 sec speech
- There is very often a pause after 2-2.5 sec speech
- There is often a F0-declination pattern spanning over 2-2.5 sec.
- There is often an intonational phrase/tone unit boundary tone (H%/L%) after 2-2.5 sec speech
- There is often final lengthening/laryngealization after 2-2.5 sec speech

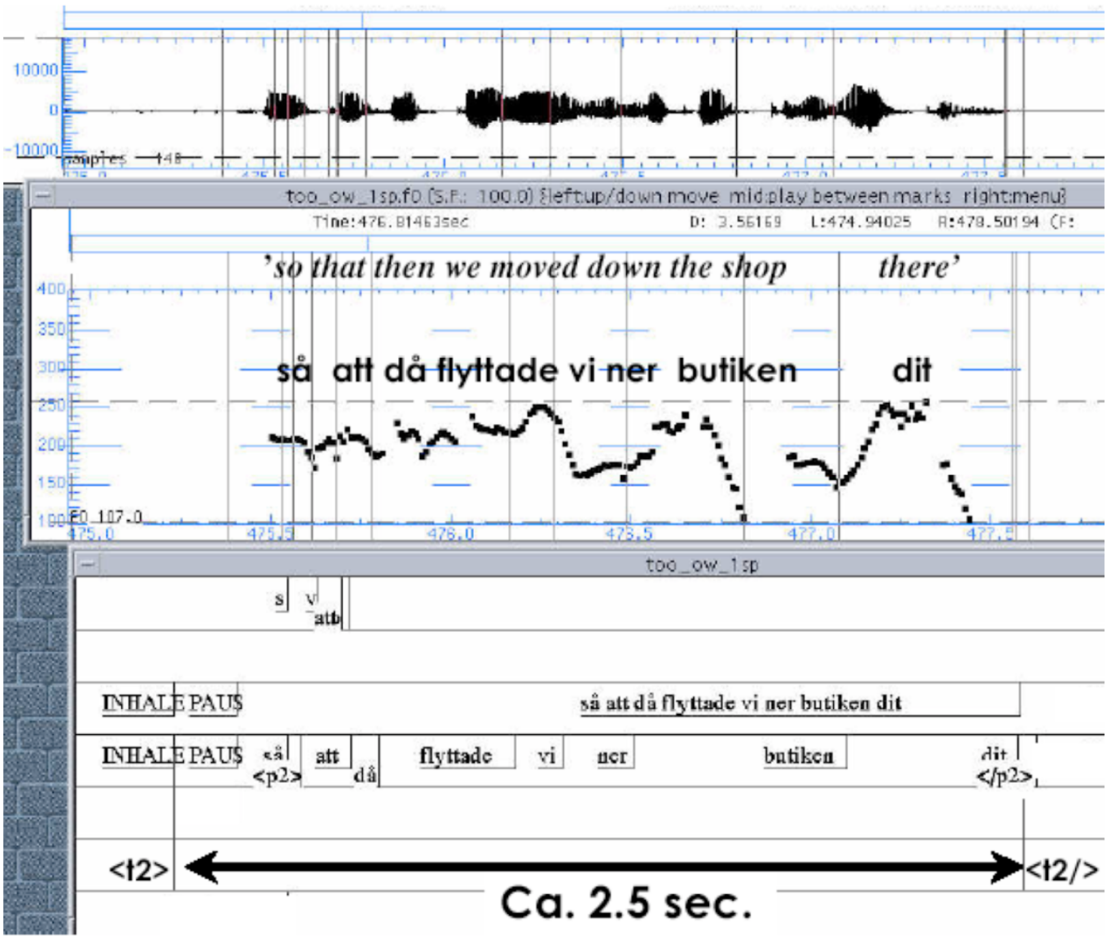

Figure 2. An example of a spontaneous speech chunk consisting of a complete clause which has a duration of about 2.5 seconds.

In addition to the above prosodically-based correlates, it has also been observed that there is often a constituent boundary after 2-2.5 seconds of speech. Thus there would seem to be a host of indications which would lead us to believe that there does in fact exist some kind of timing restriction on speech coding. If this is seen to be the case in more rigorous testing, it is a restriction that can be very useful in developing algorithms for the parsing of spontaneous speech.

In summary, in the investigation and analysis of timing restrictions on production units, we are thus making the following basic assumptions:

- A 2-2.5 sec speech production unit can contain internal pauses
- A 2-2.5 sec speech production unit does not contain internal inspirations, i.e. inspirations occur only at the edges of production units
- A 2-2.5 sec speech production unit optimally corresponds to a clause or a constituent

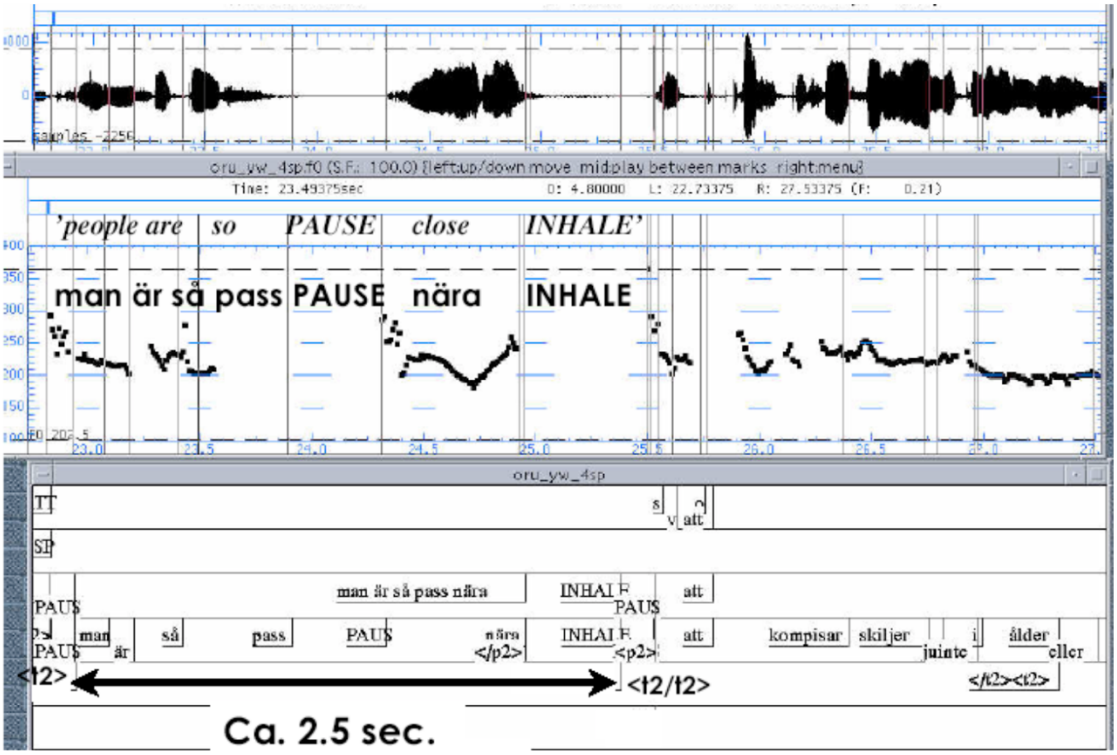

Figure 3. An example of clause containing a silent pause before the final focussed constituent. The whole fragment (clause + pause+ inspiration (labelled 'INHALE') has a duration of ca. 2.5 seconds.

2.1 Pauses, inspirations and the internal structure of production units

In Figures 2-6, we present some examples of 2-2.5 second long production units from our data that illustrate the varied structure of these units.

Figure 2 shows an example of a speech production unit that corresponds to a complete clause that has a duration of about 2.5 seconds. This kind of utterance can be thought of as an ideal example of the output of the speech planning mechanism, a complete clause realizing an underlying proposition.

Furthermore, the whole utterance constitutes a typical instance of an intonational phrase exhibiting F0-declination and a final phrase boundary tone.

As is often the case in spontaneous speech, however, the output of the speech planning component does not necessarily constitute a whole clause but rather only contains parts of a clause. Central to the analysis presented here is the assumption that pauses can occur within a speech production unit, i.e. it will be maintained that the 2-2.5 second interval can contain pauses, both silent and filled as in Figures 3 and 4. In each of these cases, the pause, silent in Figure 3 and filled pause (*EH*) in Figure 4, precedes the final focussed constituent. The internal silent and filled pauses can be thought to be a reflexion of lexical access time.

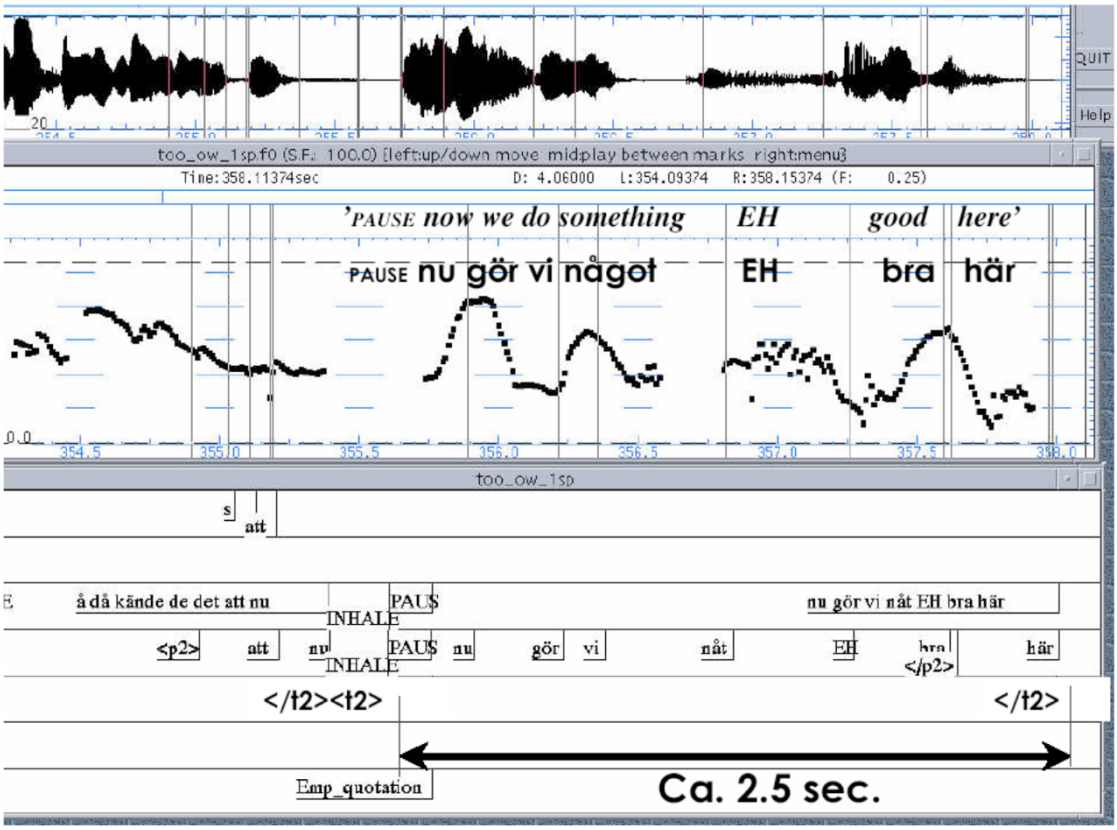

Figure 4. An example of clause containing a filled pause (EH). The whole fragment (clause + filled pause) has a duration of ca. 2.5 seconds and is characterized by a clear F0-declination pattern.

Breath pauses (i.e. inspirations) however, are assumed **not** to occur internal to the 2-2.5 sec. production units. Inspirations are rather assumed to occur only at the edges of speech planning units. In Figure 3, for example, the inspiration is at the right edge of the production unit, in this case, a clause. We are thus

attributing breathing an important role in the segmentation of spontaneous speech. Since inspirations occur only at the edges of production units, they can be thought of as anchor points for the division of speech into production units. In the context of automatic speech processing, the recognition of inspirations can be thought of as the first important stage in the chunking of speech into information units. This idea incorporates findings of e.g. Winkworth et al. (1995) and Hird and Kirsner (2002), who show that inspirations occur predominantly at grammatical boundaries.

Figure 5 illustrates an example of a clause whose articulation stretches over two production units. The final constituent in the clause, a long compound, *terapibiträdeskurs* 'therapy-assistent course' is preceded by a silent pause. Again, the silent pause can be thought to be due to the time needed to access the word from the mental lexicon and/or to generate it if the compound is not a lexical item. stored as a whole. Again, the inspiration ('INHALE') comes at the end of the clause.

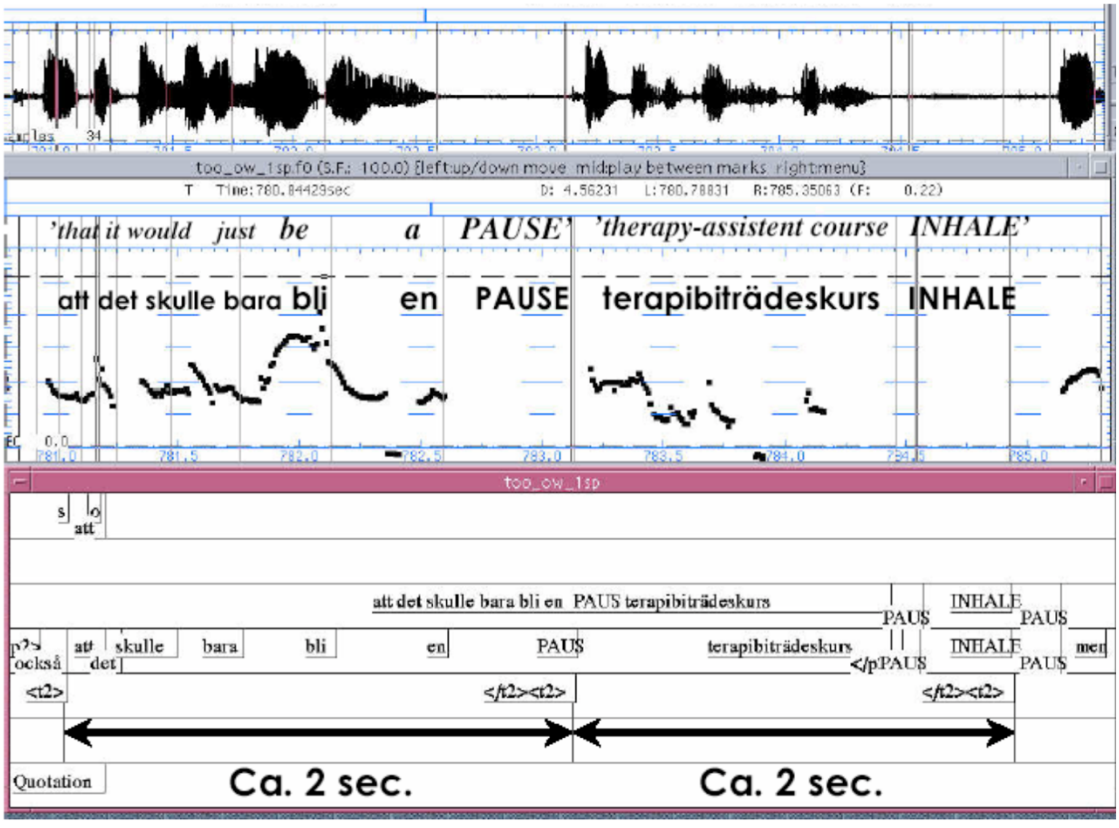

Figure 5. An example of clause which is produced during two production units. The final constituent, a compound, is preceded by a pause and ends in an inhalation.

Figure 6 shows an example of spontaneous speech illustrating how inhalations can function as anchor points for the segmentation of speech into units corresponding to clauses and smaller constitutents. Assuming that inspirations can be detected and labelled at some initial phase of the speech recognition process, segmentation of speech could then proceed to the left and right of the INHALE- labels in 2-2.5 second intervals, searching for prosodic cues which are known to signal boundaries between clauses. For example, the segmentation algorithm could move 2 seconds to the right of the first inhalation in the speech signal in Figure 6; around that point, one would expect to find a prosodic boundary of some kind. The H% tone is such a cue which could in its turn be expected to correlate with a syntactic constituent boundary. At this point, then, one could insert a production unit boundary after the H% tone. In the labelling work done in the current project, we have used the label <t2/> to indicate the end of a time- restricted production unit and <t2> to indicate the beginning of such a unit.

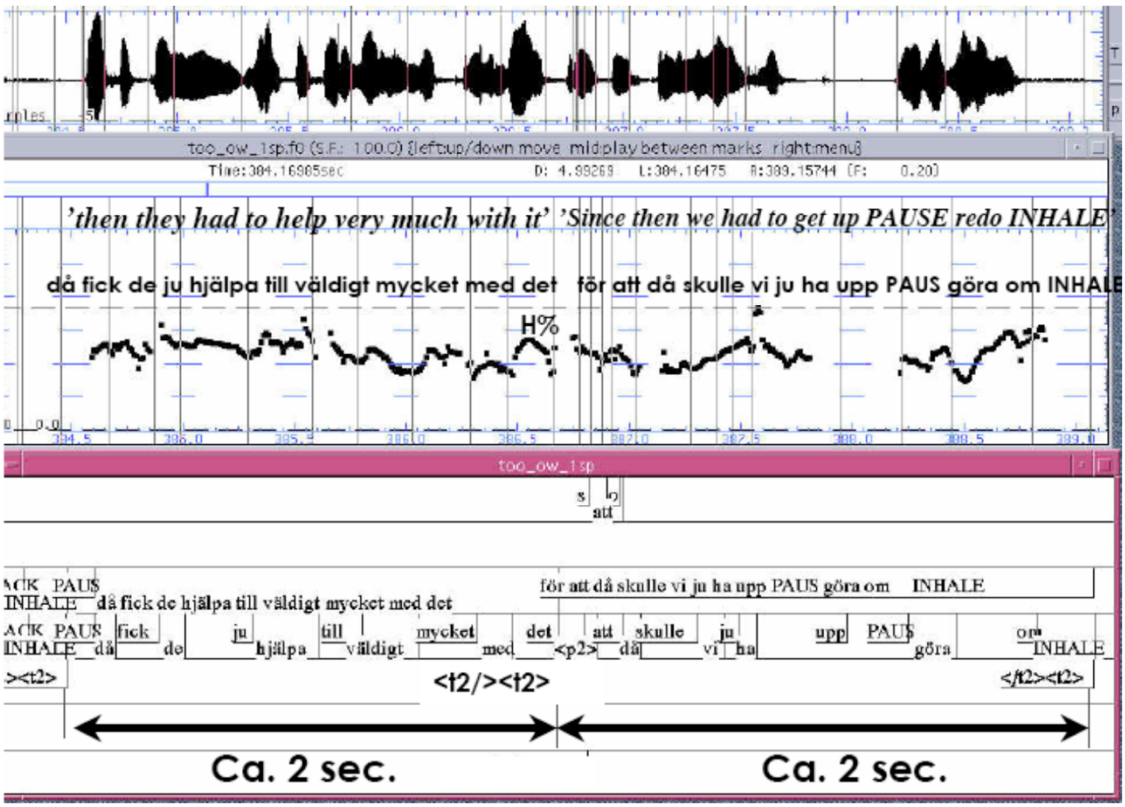

Figure 6. An example of spontaneous speech illustrating how inspirations (labelled INHALE) can be used as anchors in the segmentation of speech into time-restricted speech production units, The labels <t2/> (end point) and <t2> (beginning point) have been used as right and left edge labels for the production units.

3 Conclusion and follow-up studies

Prosodic evidence for the existence of isochronal 2-2.5 sec speech production units has been presented. Factors such as F0-declination patterns defined over these 2-2.5 sec. units, as well as boundary tones at the edges of these assumed planning units give support to the idea that prosodic structure serves as an important planning framework for an utterance. The findings provide support for the assumption of a 'Prosodic Planning Hypothesis' such as that proposed by Shattuck-Hufnagel and Turk (1996) and Shattuck-Hufnagel (2000), who assume that an utterance-specific frame "independent of its contents plays a role in production processing, and prosodic structure is a natural candidate for this structural frame". Similar ideas have also been presented by Wheeldon and Lahiri (1997) who claim that "articulation is preceded by the generation of an abstract prosodic representation of an utterance".

Breathing has been assumed to play an important role in delimitation of these production units: Inspirations only occur at edges and can thus function as anchors for the grouping of speech into 2-2.5 sec speech chunks. Local prosodic information (pauses, boundary tones (H%/L%) and the timing restriction, can be used to make a further segmentation of spontaneous speech into 2-2.5 sec production units. The existence of such a time restriction on speech planning can be used in the design of algorithms for the automatic segmentation of speech.

In follow-up studies, we plan to design an algorithm for segmenting speech using the timing restriction as well as other prosodic parameters. Further, in order to make use of the timing restriction, it is necessary to be able to distinguish pauses containing inhalations from silent pauses, since it is only the former that are assumed to occur exclusively at planning unit boundaries. Thus a study on the acoustic properties of inhalations is currently underway. Conducting neurophysiological experiments to look for external support for timing unit boundaries is also planned as a future study. Moreover, a better understanding of the relationship between pause type and type of memory process as well as the relationship between the planning of breathing in relationship to speech planning are further topics that we plan to investigate in future research.

Acknowledgements

This research has been supported by grant 2001-06309 from the *VINNOVA* (*Verket för Innovationssystem* 'The Swedish Agency for Innovation Systems') Language Technology Program.

References

Ackermann, H. and Hertrich, I. (2003). Cerebellar contributions to speech motor control and auditory verbal imagery: acoustic/kinematic analyses of ataxic dysarthria and functional magnetic resonance imaging in healthy subjects. *Proceedings of 15th ICPhS (Barcelona)*, 163-167.

Baddeley, A. (1997). *Human Memory: Theory and Practice*. Hove: Psychology Press.

Chafe, W. (1994). *Discourse, Consciousness and Time*. Chicago: University of Chicago Press.

Clark, H. and T. Wasow (1998). Repeating words in spontaneous speech. *Cognitive Psychology 37*, 201-242.

Fant, G. and A. Krukenberg (1996). On the quantal nature of speech timing. *Proceedings of ICSLP 96*, 2044-2047.

Hird, K. and K. Kirsner (2002). The relationship between prosody and breathing in spontaneous discourse. *Brain and Language 80*, 536-555.

Horne, M., J. Frid, B. Lastow, G. Bruce and A. Svensson (2003). Hesitation disfluencies in Swedish: prosodic and segmental correlates. *Proceedings ICPhS* (Barcelona), 2429-2432.

Horne, M., J. Frid and M. Roll. (In press). Hesitation disfluencies after the clause marker *att* 'that' in Swedish. *Working Papers* (Dept. of Linguistics, Lund University) 51.

Nordic Prosody IX, Ed. G. Bruce and M. Horne. Frankfurt am Main, 2006, pp. 117-126.

Levelt, W. (1989). *Speaking: From Attention to Articulation.* Cambridge, Mass.: MIT Press.

Sigurd, B. (1983) How to make a text production system speak. *Working Papers* (Dept. of linguistics, U. of Lund) 25, 179-194.

Shattuck-Hufnagel, S. (2000). Phrase-level phonology in speech production planning: evidence for the role of prosodic structure. In M. Horne (Ed.), *Prosody: Theory and Experiment. Studies Presented to Gösta Bruce*, 201-229. Dordrecht: Kluwer.

Shattuck-Hufnagel, S. and A. Turk (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research* 25, 193- 247.

Wheeldon, L. and A. Lahiri (1997). Prosodic units in speech production. *Journal of Memory and Language* 37, 356-381.

Winkworth, A., P. Davis, R. Adams and E. Ellis (1995). Breathing patterns during spontaneous speech. *Journal of Speech and Hearing Research 38*, 124-144.