Speech Segmentation using Raw Sound

Merle Horne

Dept. of Linguistics

Center for Language and Literature

Lund University

Språk- och litteraturcentrum
Centre for languages and literatur
Centre de langues et de littérature

LUNDS
UNIVERSITET

Speech synthesis/Speech recognition/Speech understanding

↕

Psycholinguistic models of speech production/comprehension

VERKET FÖR INNOVATIONSSYSTEM

The role of function words in spontaneous speech processing

SPEECH TECHNOLOGY PROGRAM PROJECT

Funding: 2003-2004

MERLE HORNE
JOHAN FRID
GÖSTA BRUCE
BIRGITTA LASTOW
MIKAEL ROLL
ADINA SVENSSON

www.ling.lu.se/projects/ProSeg2.html_

Psycholinguistic hypotheses tested on spontaneous speech

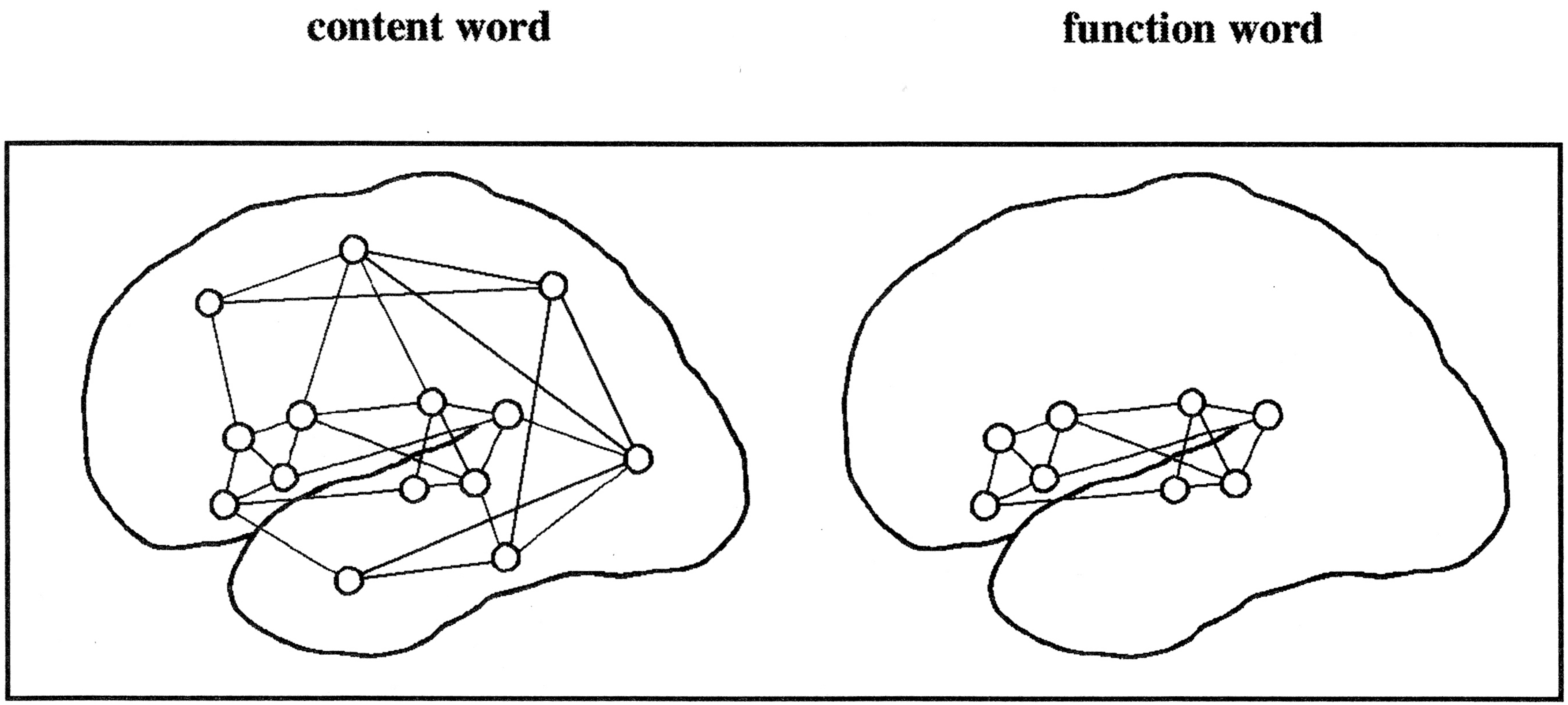
- **Commit- and- Restore hypothesis**: Stranded function word reflect ”syntactic commitments” (Clark & Wasow (1998)). I.e. they signal that the speaker intends to produce a constituent of the kind signalled by the function word produced

- **Complexity hypothesis**: the probability that a speaker will hesitate in speech production will increase, the more complex the constituent being planned is (Clark & Wasow (1998))

Related Neurophysiological hypotheses (Pulvermüller 1995, 2003)

- Function words stored in the perisylvian cortex (Broca’s region) - lateralized

- Content words have a more widely distributed storage - not lateralized- support from imaging studies (Pulvermüller 1995, 2003)

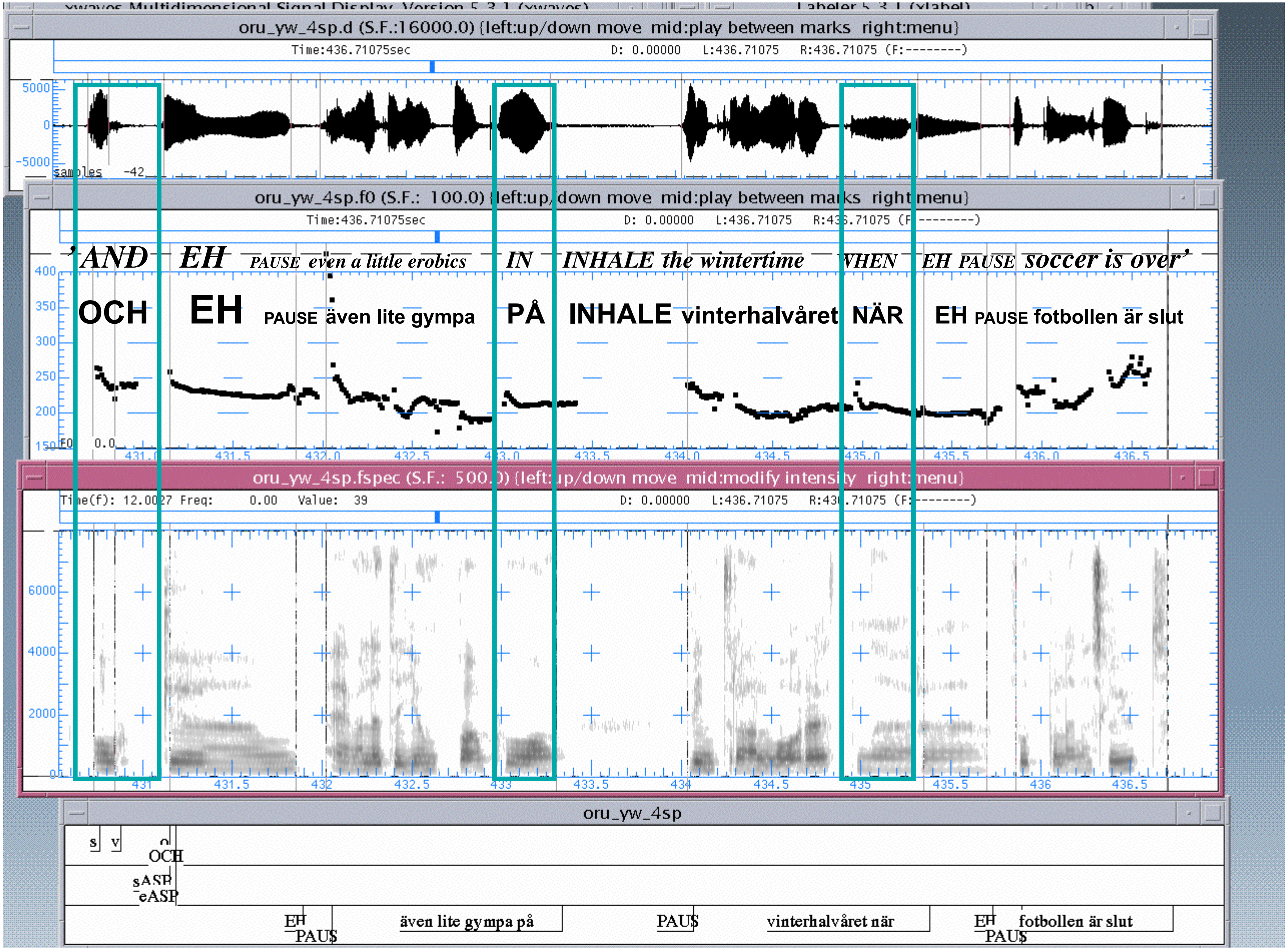


content word

function word

Figure 6.7. Schematic illustration of left-hemispheric distributions postulated for cell assemblies representing high-imageability content words (left) and highly abstract function words (right).

(From Pulvermüller 2003)



xwaves Multidimensional Signal Display Version 5.3.1 (xwaves)
Labeler 5.3.1 (xlabel)
oru_yw_4sp.d (S.F.:16000.0) (left:up/down move mid:play between marks right:menu)
Time:436.71075sec D: 0.00000 L:436.71075 R:436.71075 (F:--------)
5000
0
-5000
samples -42
oru_yw_4sp.f0 (S.F.: 100.0) (left:up/down move mid:play between marks right:menu)
Time:436.71075sec D: 0.00000 L:436.71075 R:436.71075 (F:--------)
400
350
300
250
200
150
F0 0.0
431.0
431.5
432.0
432.5
433.0
433.5
434.0
434.5
435.0
435.5
436.0
436.5
oru_yw_4sp.fspec (S.F.: 500.0) (left:up/down move mid:modify intensity right:menu)
Time(f): 12.0027 Freq: 0.00 Value: 39 D: 0.00000 L:436.71075 R:436.71075 (F:--------)
6000
4000
2000
0
431
431.5
432
432.5
433
433.5
434
434.5
435
435.5
436
436.5
oru_yw_4sp
s
v
o
OCH
sASP
eASP
EH
PAUS
även lite gympa på
PAUS
vinterhalvåret när
EH
PAUS
fotbollen är slut

'AND EH PAUSE even a little erobics IN INHALE the wintertime WHEN EH PAUSE soccer is over'

OCH EH PAUSE även lite gympa PÅ INHALE vinterhalvåret NÄR EH PAUSE fotbollen är slut

Speech ’Chunking’ problem (Miller 1956)

Units for speech processing

(speech synthesis/speech recognition)

Timing restrictions on speech processing

•Memory research (Baddley 1997)

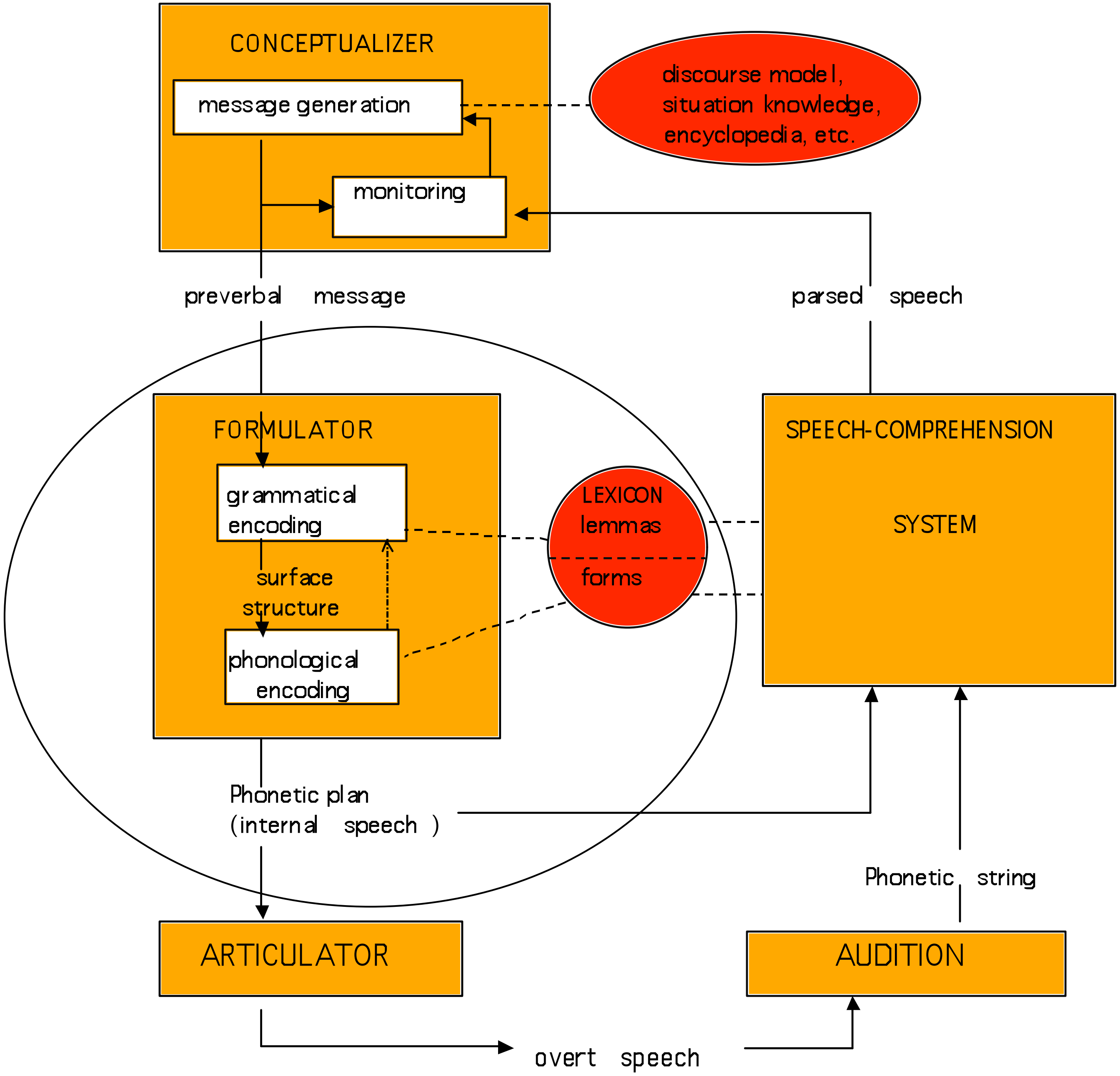
‘Phonological loop’ (2 sec.)

•Cognitive linguistics (Chafe 1994)

‘Focus of consciousness’ (1-2 sec)

Speech planning unit

Ca. 2-2.5 sec. time limit



CONCEPTUALIZER

message generation

monitoring

discourse model, situation knowledge, encyclopedia, etc.

preverbal message

parsed speech

FORMULATOR

grammatical encoding

surface structure

phonological encoding

LEXICON

lemmas

forms

SPEECH-COMPREHENSION SYSTEM

Phonetic plan (internal speech)

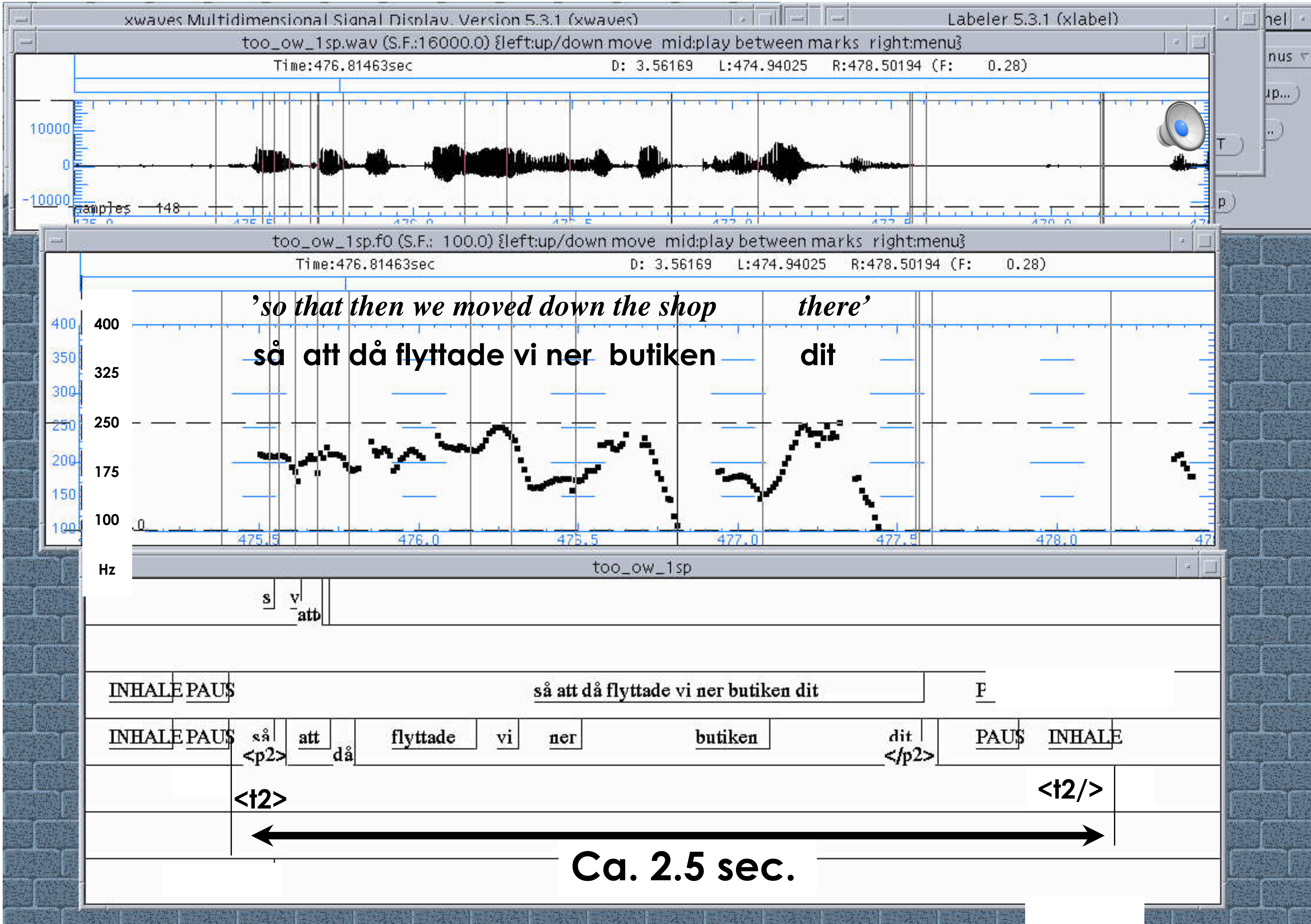
Phonetic string

ARTICULATOR

AUDITION

overt speech

The Speaker as Information Processor
From Levelt *(1995) Speaking*.



xwaves Multidimensional Signal Display, Version 5.3.1 (xwaves)

Labeler 5.3.1 (xlabel)

too_ow_1sp.wav (S.F.:16000.0) {left:up/down move mid:play between marks right:menu}

Time:476.81463sec D: 3.56169 L:474.94025 R:478.50194 (F: 0.28)

too_ow_1sp.f0 (S.F.: 100.0) {left:up/down move mid:play between marks right:menu}

Time:476.81463sec D: 3.56169 L:474.94025 R:478.50194 (F: 0.28)

’so that then we moved down the shop there’

så att då flyttade vi ner butiken dit

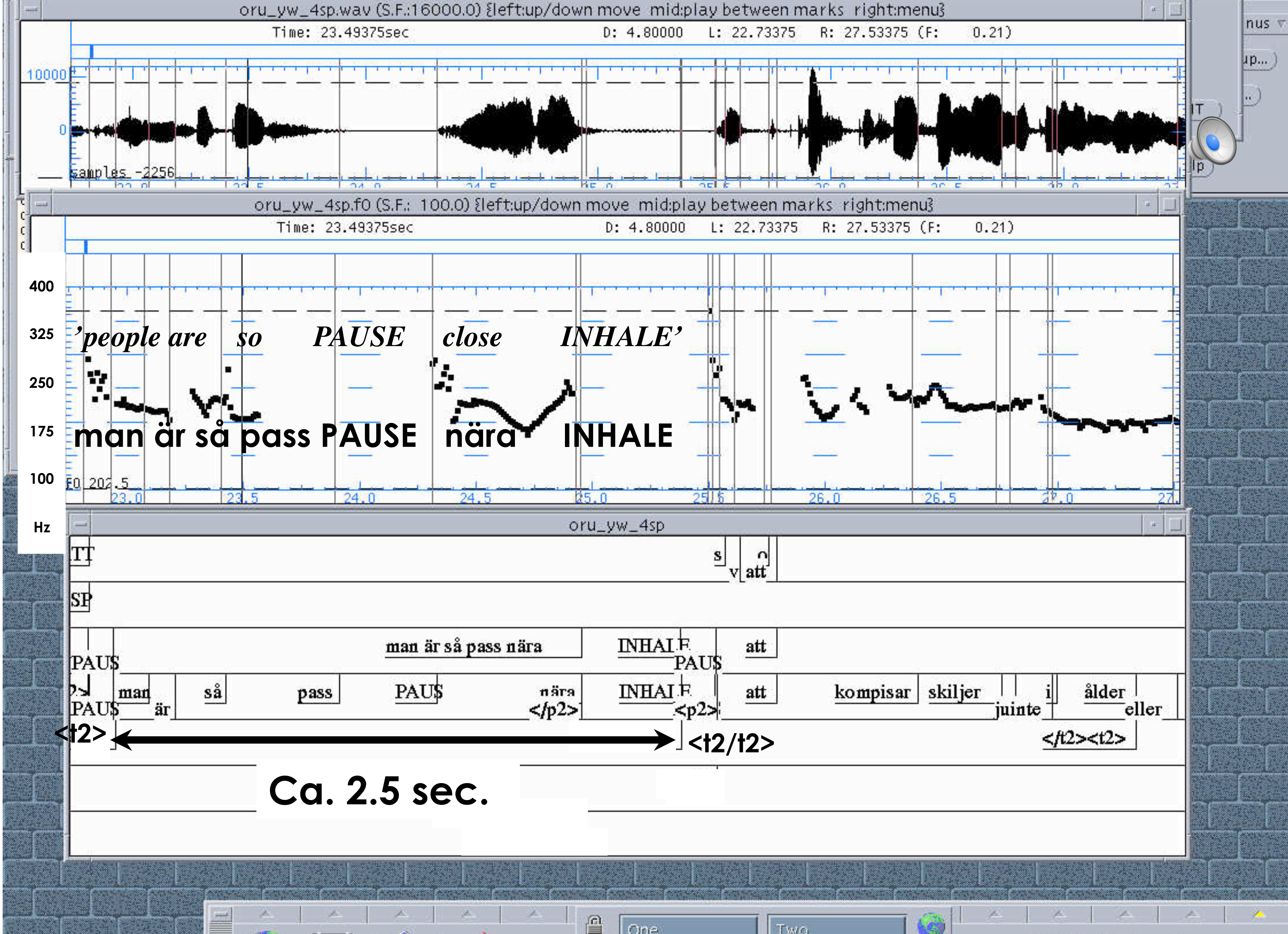
too_ow_1sp

INHALE PAUS så att då flyttade vi ner butiken dit P

INHALE PAUS så <p2> att då flyttade vi ner butiken dit </p2> PAUS INHALE

<t2> <t2/>

Ca. 2.5 sec.



oru_yw_4sp.wav (S.F.:16000.0) {left:up/down move mid:play between marks right:menu}

Time: 23.49375sec D: 4.80000 L: 22.73375 R: 27.53375 (F: 0.21)

oru_yw_4sp.f0 (S.F.: 100.0) {left:up/down move mid:play between marks right:menu}

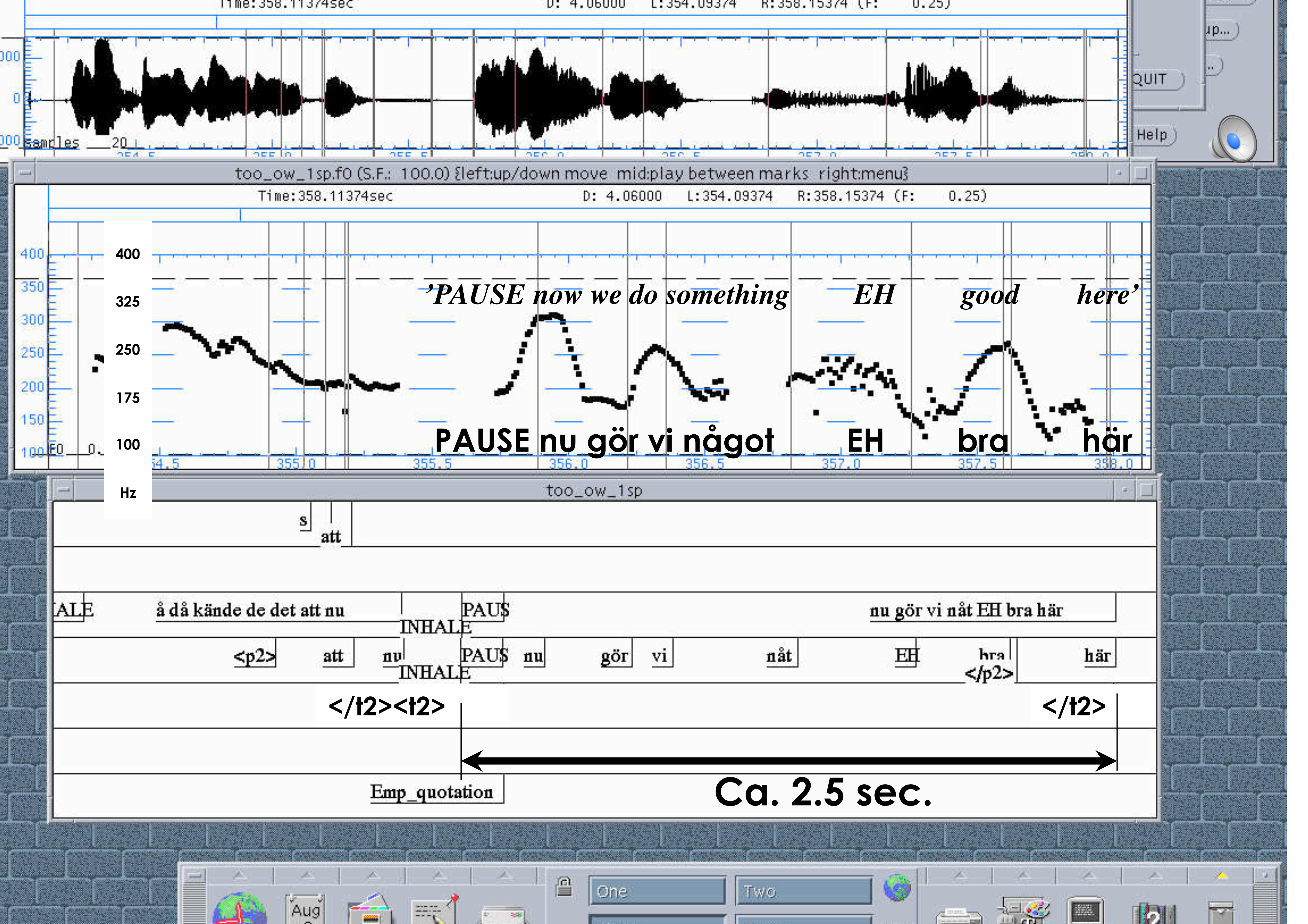
Time: 23.49375sec D: 4.80000 L: 22.73375 R: 27.53375 (F: 0.21)

'people are so PAUSE close INHALE'

man är så pass PAUSE nära INHALE

oru_yw_4sp

Ca. 2.5 sec.

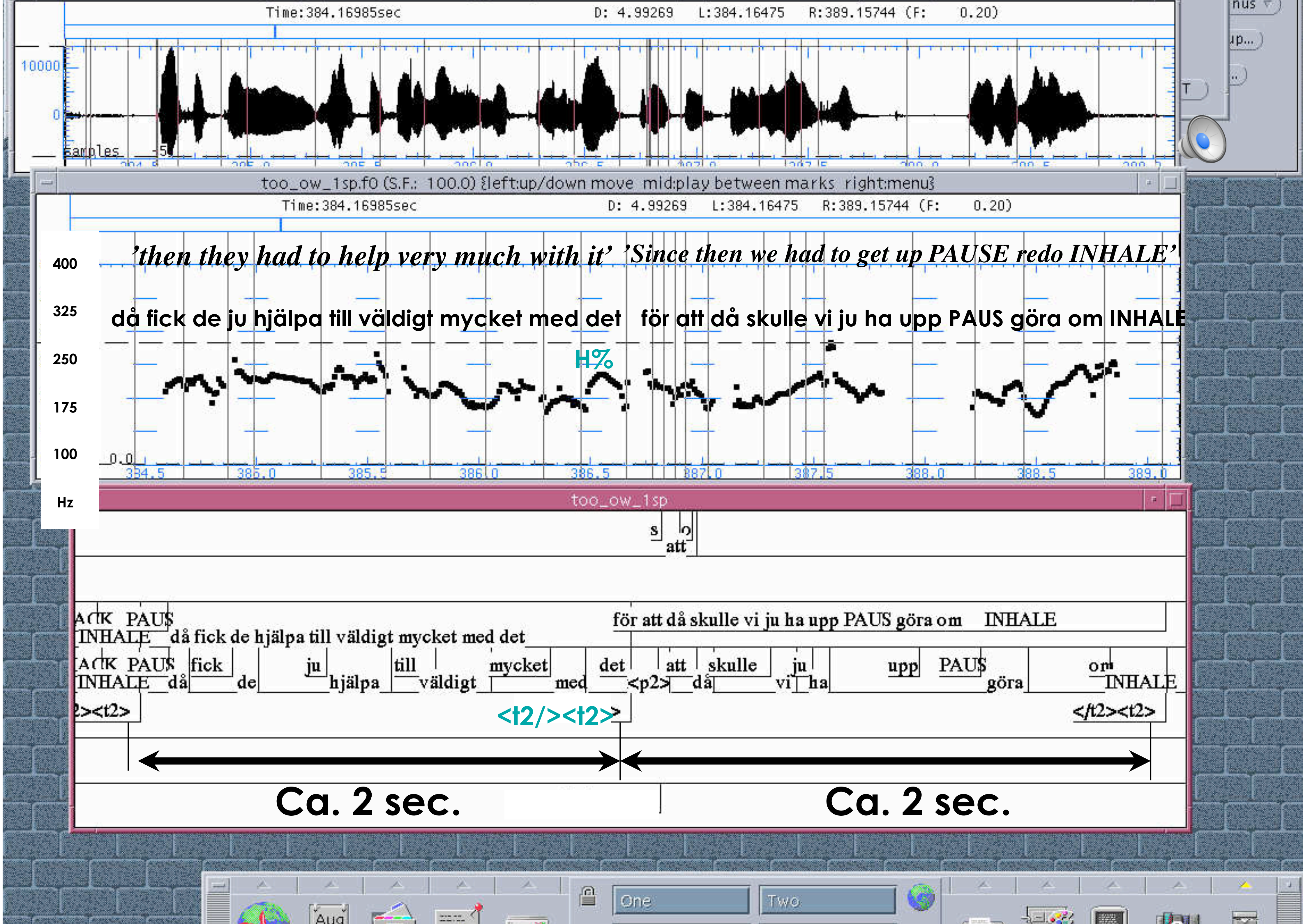


'PAUSE now we do something EH good here'

PAUSE nu gör vi något EH bra här

</t2><t2> </t2>

Ca. 2.5 sec.



’then they had to help very much with it’ ’Since then we had to get up PAUSE redo INHALE’

då fick de ju hjälpa till väldigt mycket med det för att då skulle vi ju ha upp PAUS göra om INHALE

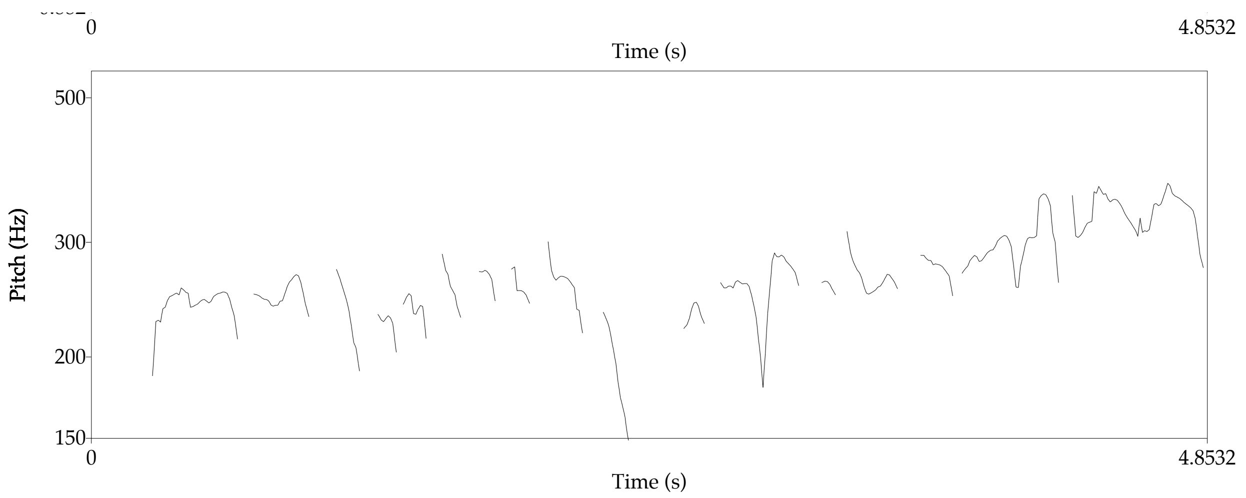
H%

<t2/><t2>

Ca. 2 sec.

Ca. 2 sec.

Time:384.16985sec D: 4.99269 L:384.16475 R:389.15744 (F: 0.20)
10000
0
samples
too_ow_1sp.f0 (S.F.: 100.0) {left:up/down move mid:play between marks right:menu}
Time:384.16985sec D: 4.99269 L:384.16475 R:389.15744 (F: 0.20)
400
325
250
175
100
Hz
0.0
384.5
385.0
385.5
386.0
386.5
387.0
387.5
388.0
388.5
389.0
too_ow_1sp
s
ɔ
att
ACK PAUS
INHALE
då fick de hjälpa till väldigt mycket med det
för att då skulle vi ju ha upp PAUS göra om INHALE
ACK PAUS fick ju till mycket det att skulle ju upp PAUS om
INHALE då de hjälpa väldigt med <p2> då vi ha göra INHALE
2><t2>
</t2><t2>
One
Two
Aug



[Jenny just nu nästsisist] [hon kan klippa Tchechenko]

[Å så är det hårt där framme] [å så kommer Johanna Hayes och ramlar]

Ca.2.1 sec

Ca. 2.3 sec

Inhalations as anchors for speech segmentation

- Can one automatically recognize inhalations?
  - Characteristics: noise, lack of F0, can have formant structure if oral
- Possible method: Template matching
  - Uses distance measures between a reference sound and signal being processed

Using cepstral coefficients for inhalation pause detection in spontaneous speech

Anders Johansson, Johan Frid, Merle Horne

SPECOM 2005, Patras,
Greece

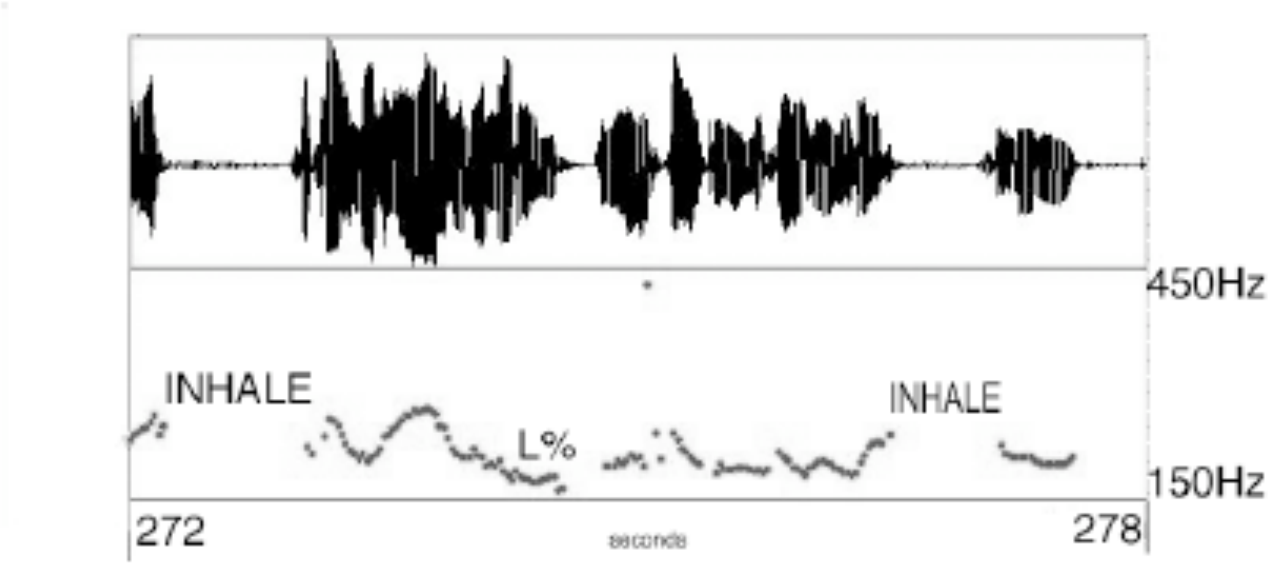


Figure 1: *An example of spontaneous speech illustrating how inhalations (labelled INHALE) can be used as anchors in the segmentation of speech into processing units. The speech between the inhalations consists of two clauses:* Så tränar man med jämna mellanrum *‘So you train regularly’ and* det gick pågick ju under flera månader *’it last lasted for several months’. The clauses constitute two prosodic phrases separated by a pause.*

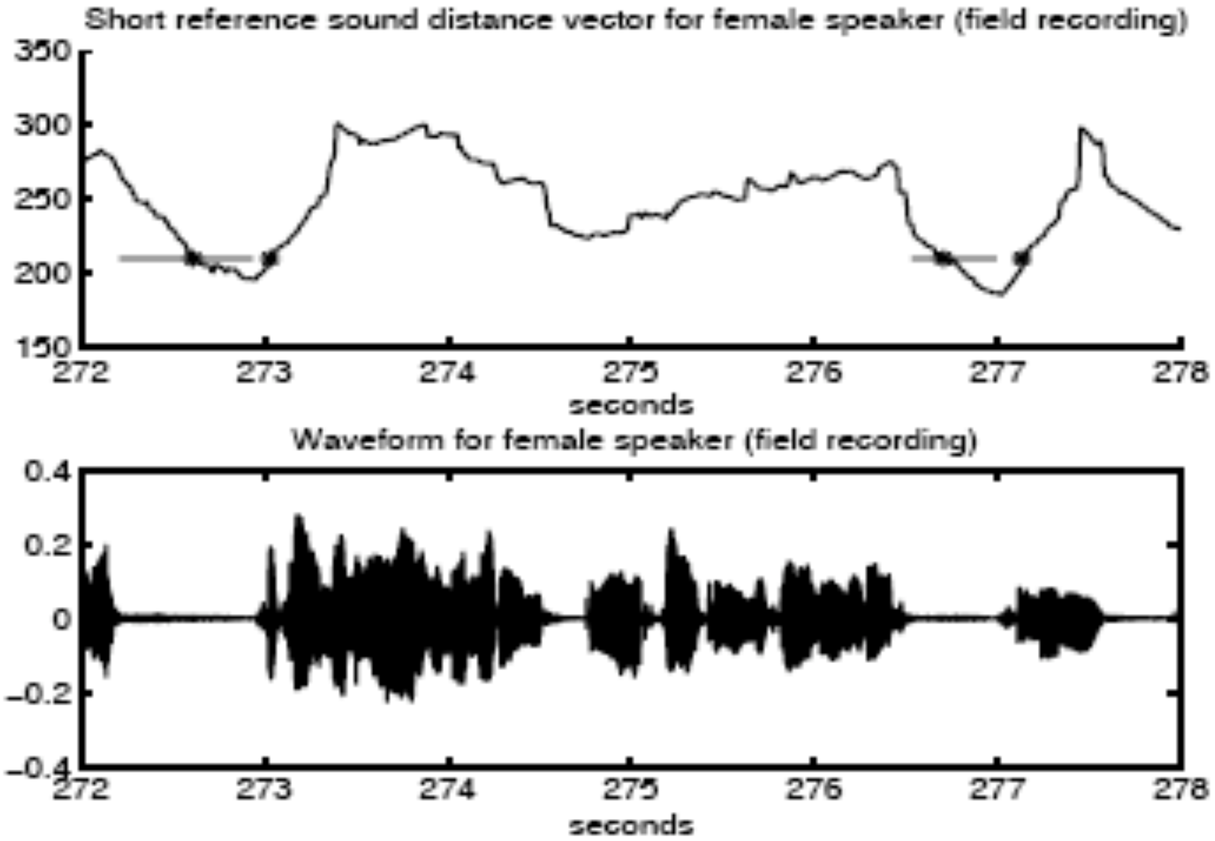


Figure 5: *Distance vector of field recording of female speaker as generated from short reference sound. Stars indicate detected inhalations, lines indicate tagged inhalations. Recording from the SweDia material.*

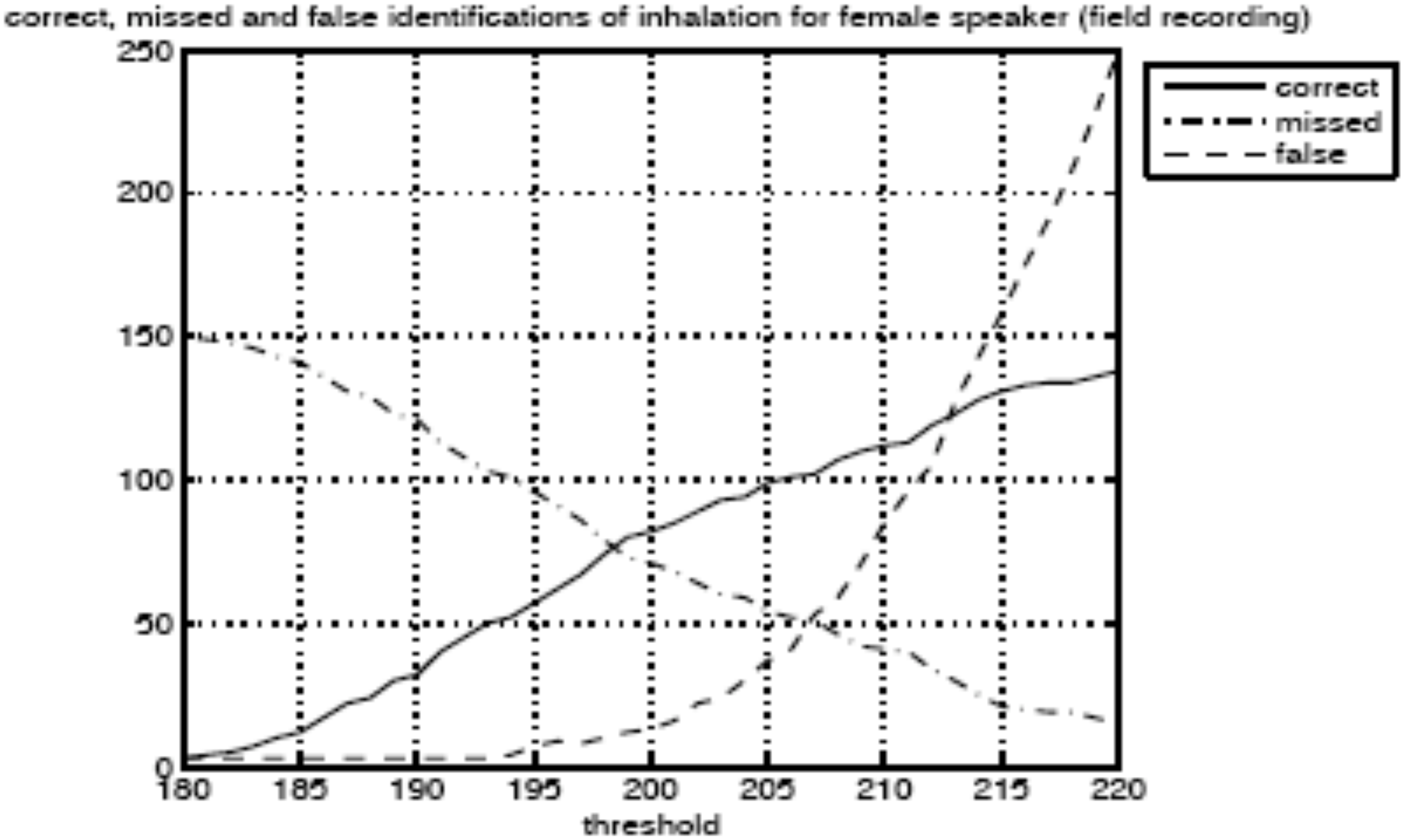


Figure 9: *Number of correct, false and missed identifications of inhalations in field recording of female speaker as generated from short reference sound.*

Observations:

False identifications associated with sounds of the following types: exhalations, word-final aspirated sounds, whispers and voiceless fricatives

Properties of inhalations quite stable, even over different speakers