How does speech timing work?

1. Timing capabilities
2. How to measure speech timing
3. Preview of things to come

Acknowledgements

- Stefanie Shattuck-Hufnagel
- Satsuki Nakai
- Mariko Sugahara

Course Overview

- 1. General timing capabilities. How can we measure timing in speech?
- 2. Systematicity in surface timing patterns. What is timing for? What is timed?
- 3. How do speakers distinguish the many uses of duration?
- 4. Is speech rhythmic?
- 5. Predictability--indirect influence via prosodic structure?

Timing Capabilities

- Precise timing capabilities are required for many human actions, e.g.
  - Catching
  - Rhythmic Tapping
  - Dancing
  - Singing
- Even non-humans must have (some of) these abilities
  - Animals catching prey

Timing Capabilities (cont.)

- Catching a ball
  - Perceive the ball is coming your way
  - Predict when it will be in a certain position
  - Control body, arms, hands, to be in the right place at the right time
    - Involves knowing when to start moving to reach the target on time
    - Involves taking account of physical conditions that may be continuously changing
- Tapping to an external rhythm
  - Perceive the rhythm
  - Predict when the beats will occur
  - Control tapping to occur at the right time to coincide with the beat

Rhythmic tapping without an external stimulus

- Plan and generate a single tap,
- Plan and generate a second tap after a planned inter-tap interval
- Remember the inter-tap interval duration
- Plan and generate a third tap after an interval of equal duration to the previous one
- Repeat…..

. . . . .

Rhythmic tapping (cont.)

. . . . .

- or
  - Generate an abstract rhythmic structure
  - Specify an inter-beat timing interval
  - Plan and generate taps to coincide with the beats

Abstract representations

- Allow us to create parametrized instructions for e.g.
  - Inter-beat timing interval
  - Tapping instrument (e.g. finger vs. pen vs. foot)
- Represent the equivalence of sets of actions
  - Tapping a rhythm at a fast rate = tapping the same rhythm at a slower rate
  - Rhythmic tapping with hand = rhythmic tapping with foot

Timing capabilities: Summary

- Internal mechanism for keeping track of time
- Perceive the timing of external events
- Predict the occurrence of future events
- Control the timing of our own actions
  - With respect to predicted external events
  - With respect to each other (inter-articulator coordination, e.g. hand-body, etc.)
  - In spite of changes in physical conditions
- Generate abstract representations and create parametrized instructions

What about speech?

So many thoughts, only one mouth

Must unfold in time

Timing in speech production

- Timing capabilities for other actions are assumed to be available for speech.
- Do we use them?
- Is speech timing systematic?
  - If yes, it is likely that we use at least some of the same capabilities that we use for other timed actions.
  - E.g., timekeeping, timing perception, prediction, motor control, abstract representations

Is speech timing systematic?

- To answer this question, we must have a way of measuring timing.

  1. Acoustic measures
  2. Articulatory measures

     Electropalatography

     Laryngography

     Fleshpoint motion tracking (e.g. electromagnetic articulometry)

- Preview: Answer is yes!

Acoustic timing measures

- Based on discontinuities in the acoustic signal (cf. Landmarks, Stevens 2002)
- Two types, with different articulatory origins
  - Supralaryngeal:
    - Onset and release of constrictions
  - Laryngeal: Voiced vs. voiceless intervals
    - Caveat: voiceless intervals can be due to
      - Open glottis
      - Tightly closed glottis

Intervals based on correlates of oral activity

- Constriction intervals for consonants
- Intervals between constrictions (often referred to as vowels)

Oral vs. laryngeal landmarks

- Correlates of laryngeal activity do not always coincide with oral activity,
- e.g. Voicing can extend into closure for a [-voice] stop.
- Oral landmarks have an advantage for duration measurements
  - Mouth opening and closing criteria are roughly comparable across segment types (e.g. mouth opening for [b] ~ mouth opening for [p])
- It may be important to additionally measure voicing-based intervals, e.g. VOT for voiceless stops
- Important to specify measurement criteria (laryngeal vs. oral correlates)

Constriction intervals

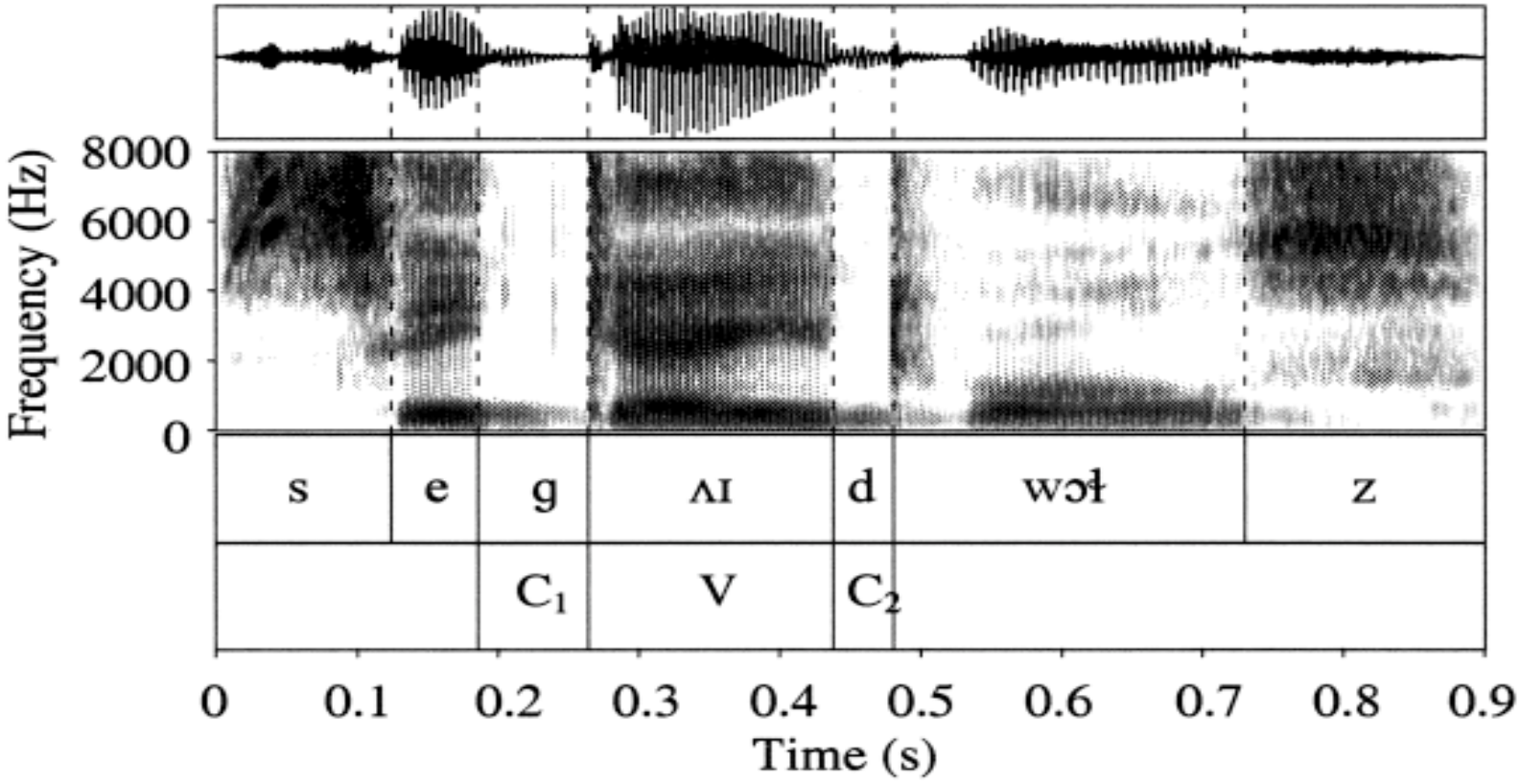


Figure 1: *Say guide walls,* spoken by a female Scottish English speaker

Figure from Turk, Nakai & Sugahara 2006

Oral vs. laryngeal criteria

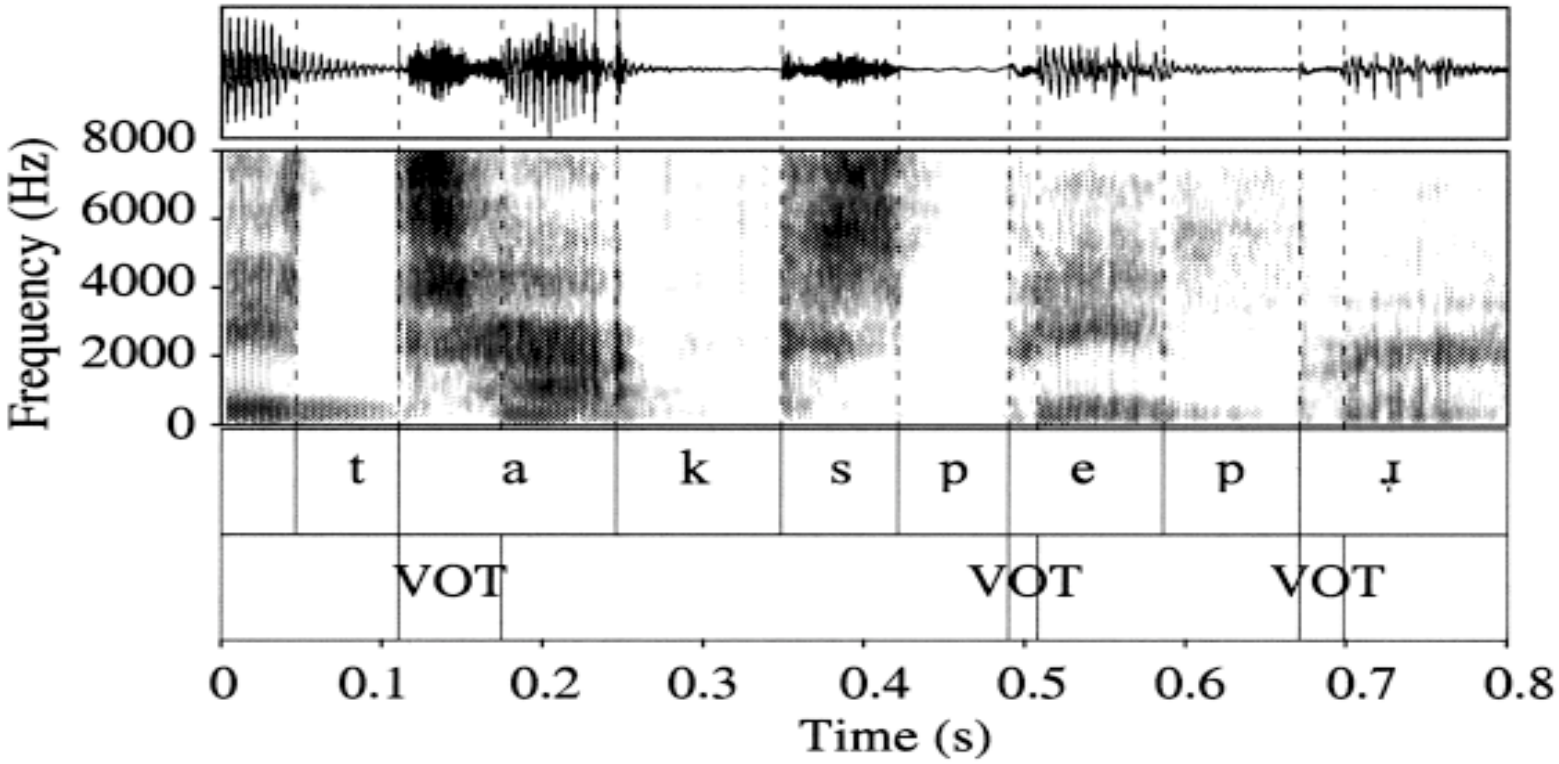


Figure 3: *Tax paper,* spoken by a female Scottish English speaker. The boundaries for the offsets of /a/ and /e/ are placed on the last glottal pulse peak in the intervals delimited by continuous F2.

Figure from Turk, Nakai & Sugahara 2006

Oral vs. laryngeal criteria

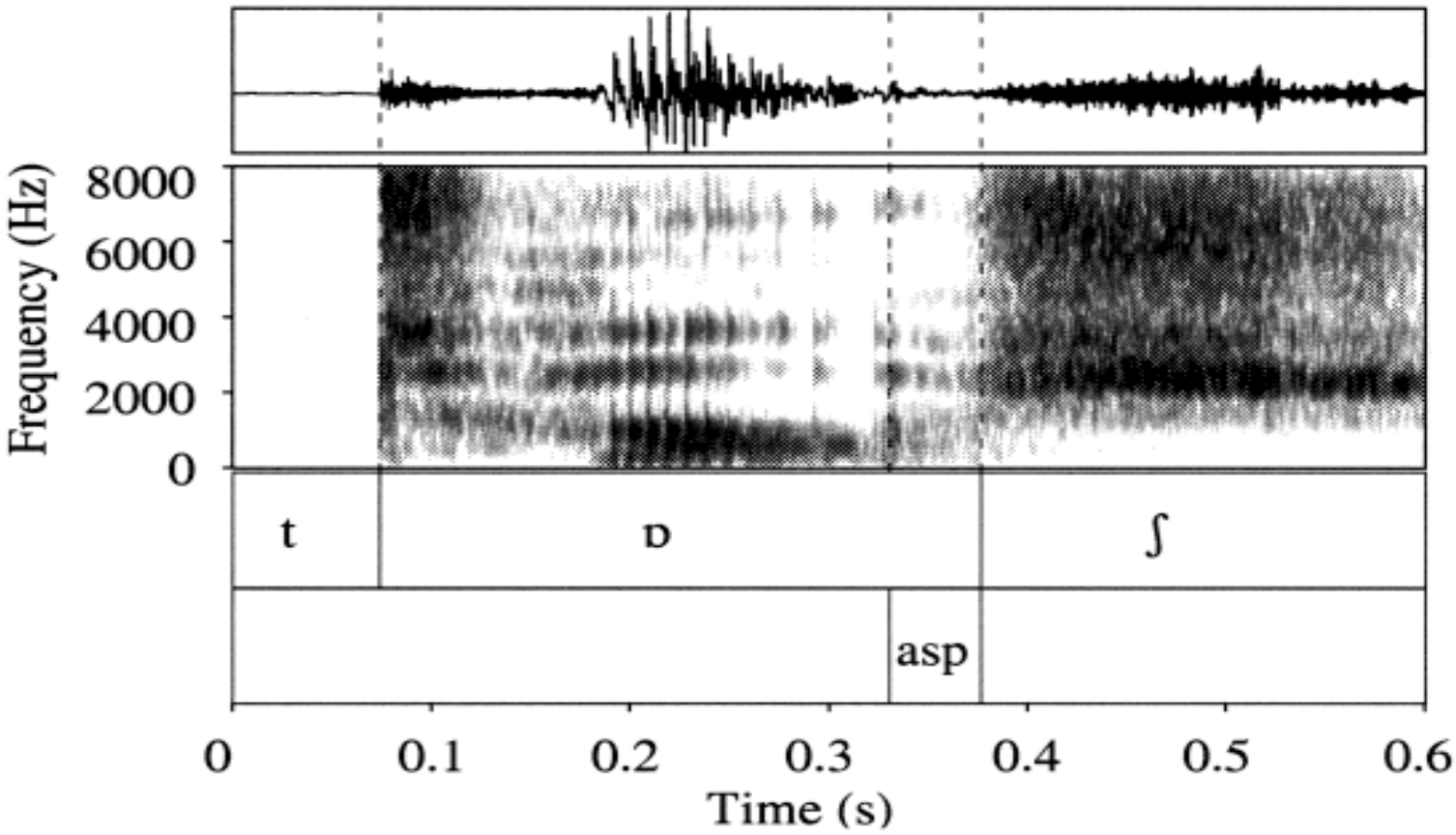


Figure 6: *Tosh,* spoken by a male Southern Standard British English speaker

Figure from Turk, Nakai & Sugahara 2006

Oral vs. laryngeal criteria

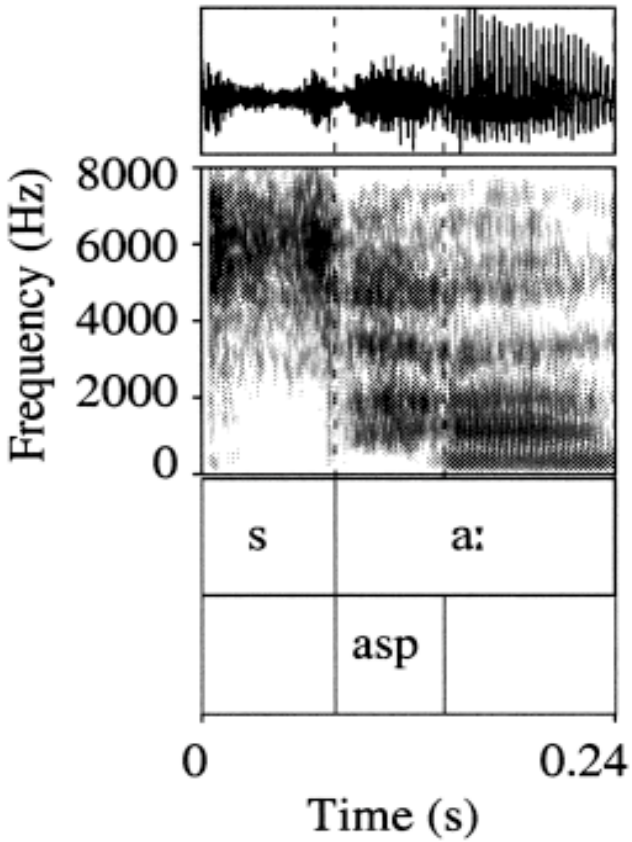


Figure 7: A fragment from *Buun-sensei ICHI-BAN-ga "saasaa"-tte ittakedo* 'Mr. Boone said NUMBER ONE is "saasaa"', spoken by a female Standard Japanese speaker. *Saasaa* is a nonsense word.

Figure from Turk, Nakai & Sugahara 2006

Cluster division based on a voicing criterion

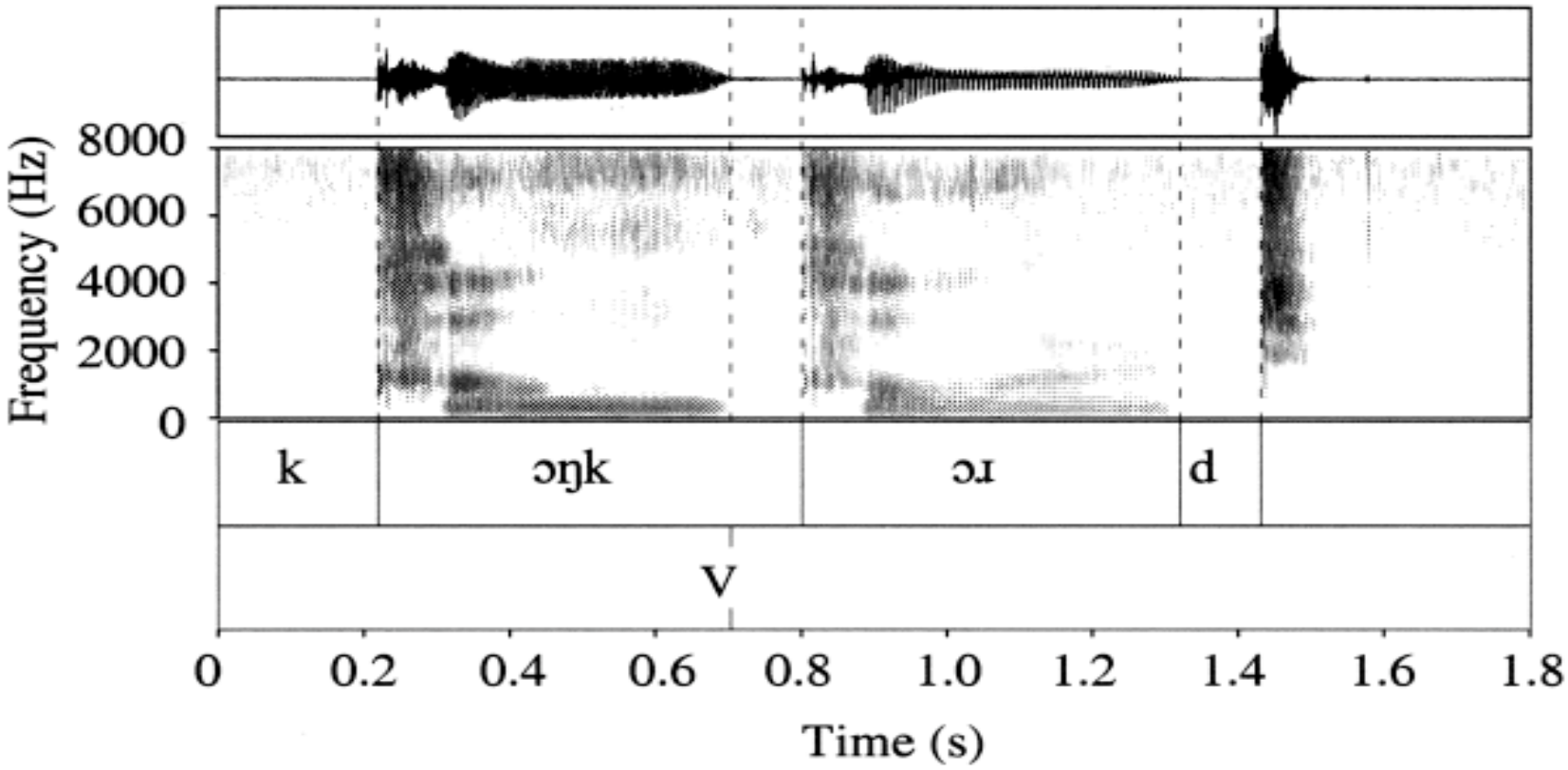


Figure 4: *Concord,* spoken by a female Scottish English speaker. *V* in the second label tier indicates the offset of voicing for [k].

Figure from Turk, Nakai & Sugahara 2006

Approximants are challenging—sugg. avoid these when possible

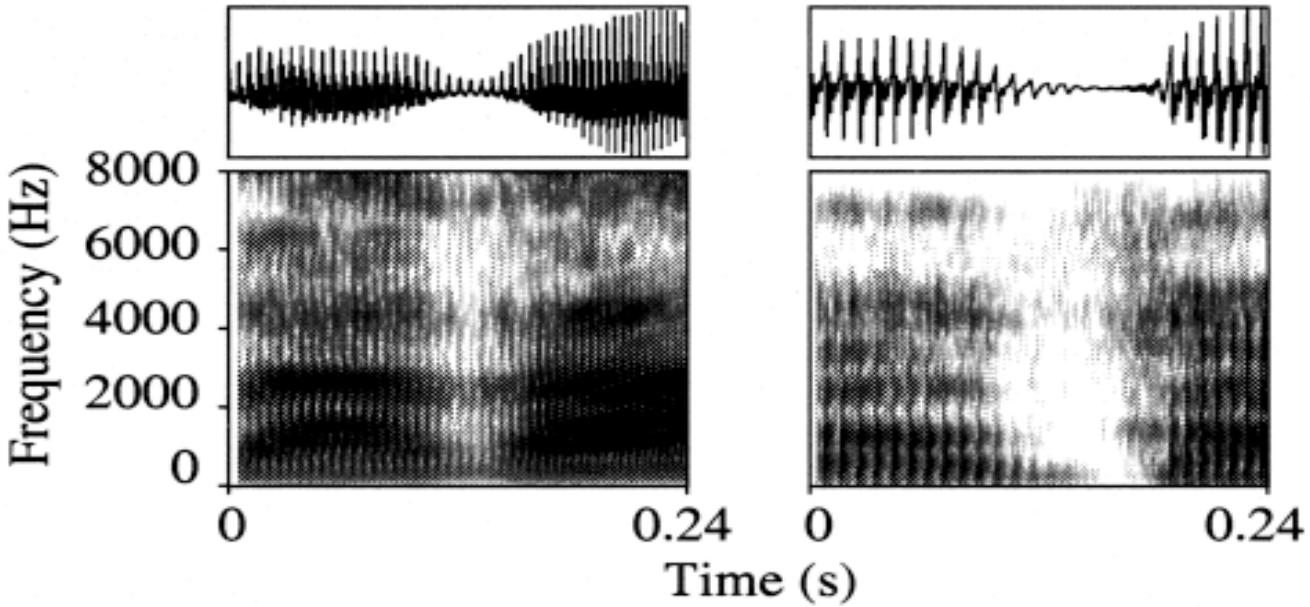


Figure 5: Japanese /aga/ where /g/ is realised as an approximant (left panel; spoken by a female Standard Japanese speaker), and as a fricative (right panel; spoken by a male Standard Japanese speaker). /aga/ is a fragment from *Sei-wa "gansani"-o totemo yorokobu* 'Sei is very pleased with "gansani"'. *Wa* is a topic marker; *gansani* is a nonsense word.

Cluster segmentation: A difficult case [st]

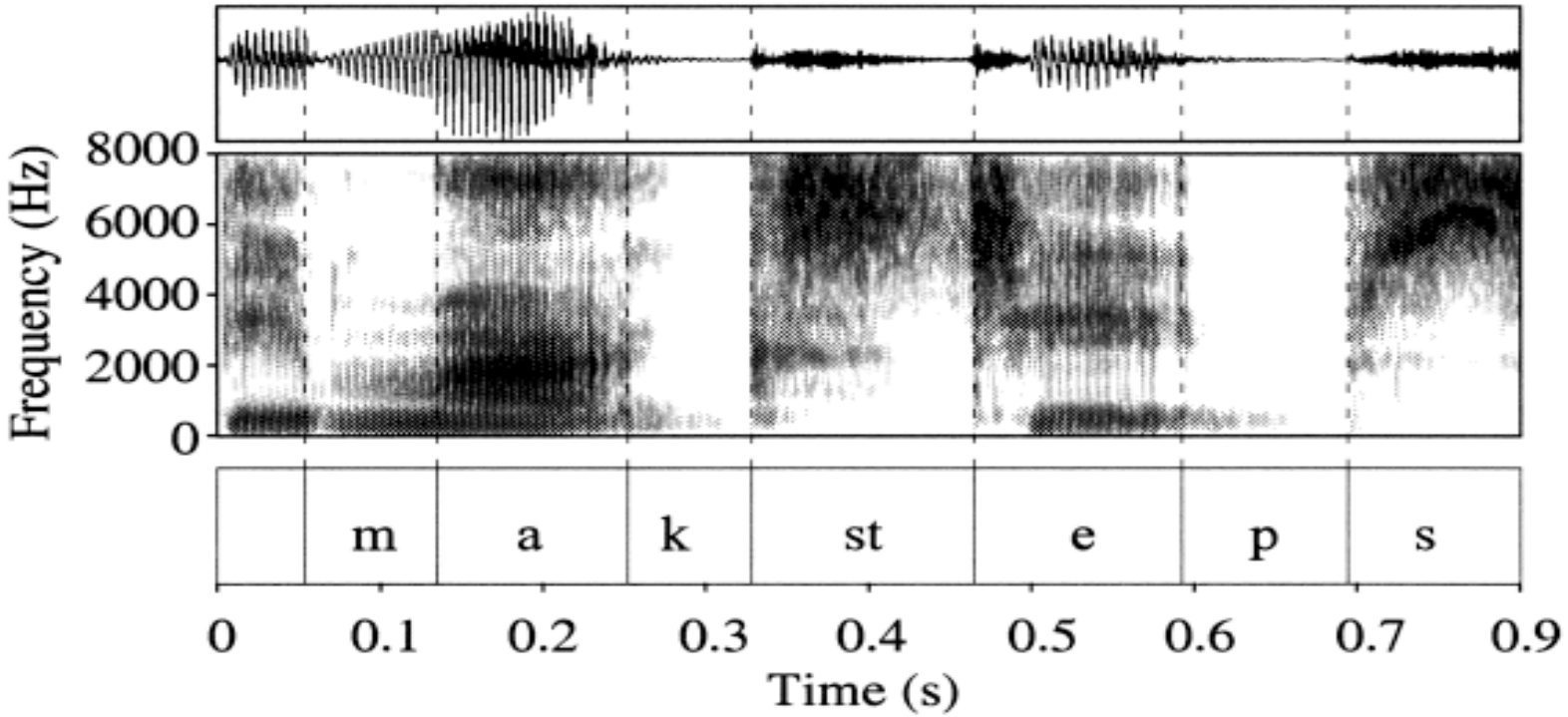


Figure 8: *Max tapes*, spoken by a female Scottish English speaker. The boundaries for the offsets of /a/ and /e/ are placed on the last glottal pulse peak in the intervals delimited by continuous F2.

Figure from Turk, Nakai & Sugahara 2006

Another difficult cluster: No attempt at segmentation

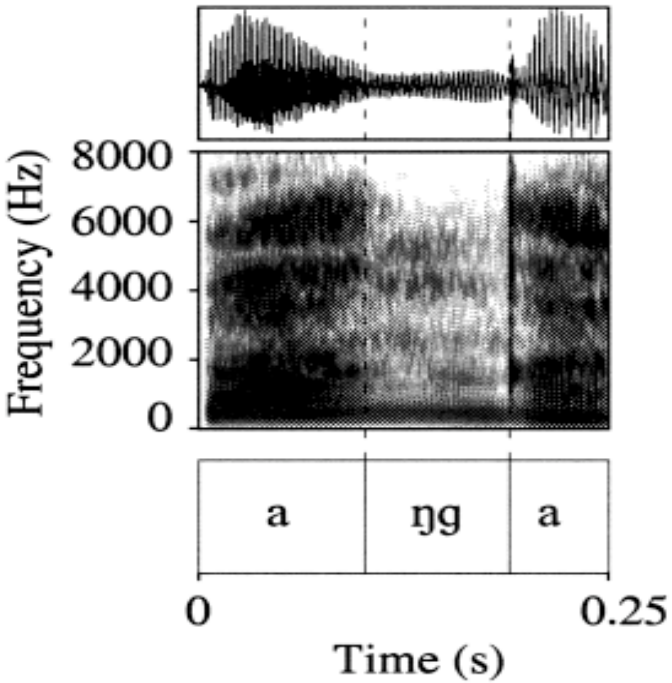


Figure 11: A fragment from *Buun-sensei ICHI-BAN-ga* "*saasaa*"*-tte ittakedo* 'Mr. Boone said NUMBER ONE is "saasaa"', spoken by a female Standard Japanese speaker. *Ban* 'number' is a suffix; *ga* is a nominative particle.

V-Nasal coda sequences can sometimes be difficult

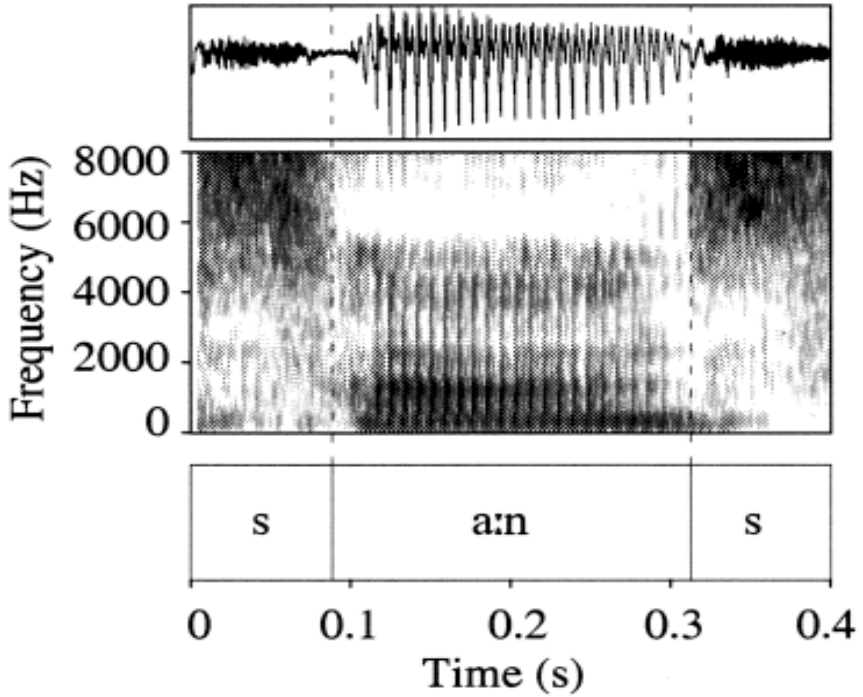


Figure 10: A fragment from *Toujou-sensei-ni kii-tara ICHI-BAN-ga* "*saansa*" 'According to Mr. Tojo, NUMBER ONE is "saansa"', spoken by a male Standard Japanese speaker. *Saansa* is a nonsense word.

Figure from Turk, Nakai & Sugahara 2006

Prosodic interval durations on the basis of segmental durations

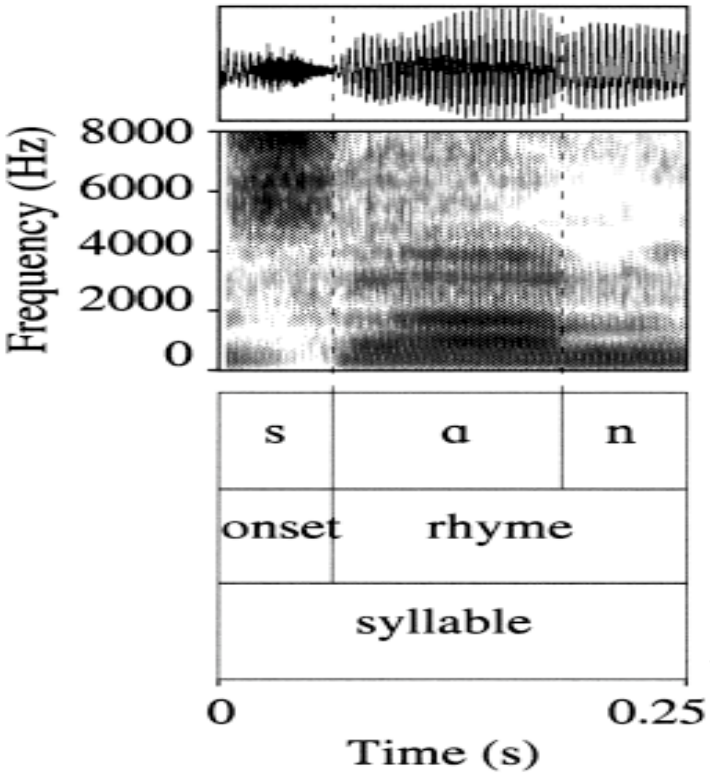


Figure 9: A fragment from *MINUSTA* "*san*" *sopii kohtaan tuhatkaksisataa* 'I THINK "san" fits [#] 1200', spoken by a female Northern Finnish speaker. *San* is a nonsense word.

Figure from Turk, Nakai & Sugahara 2006

Where does the pause start?

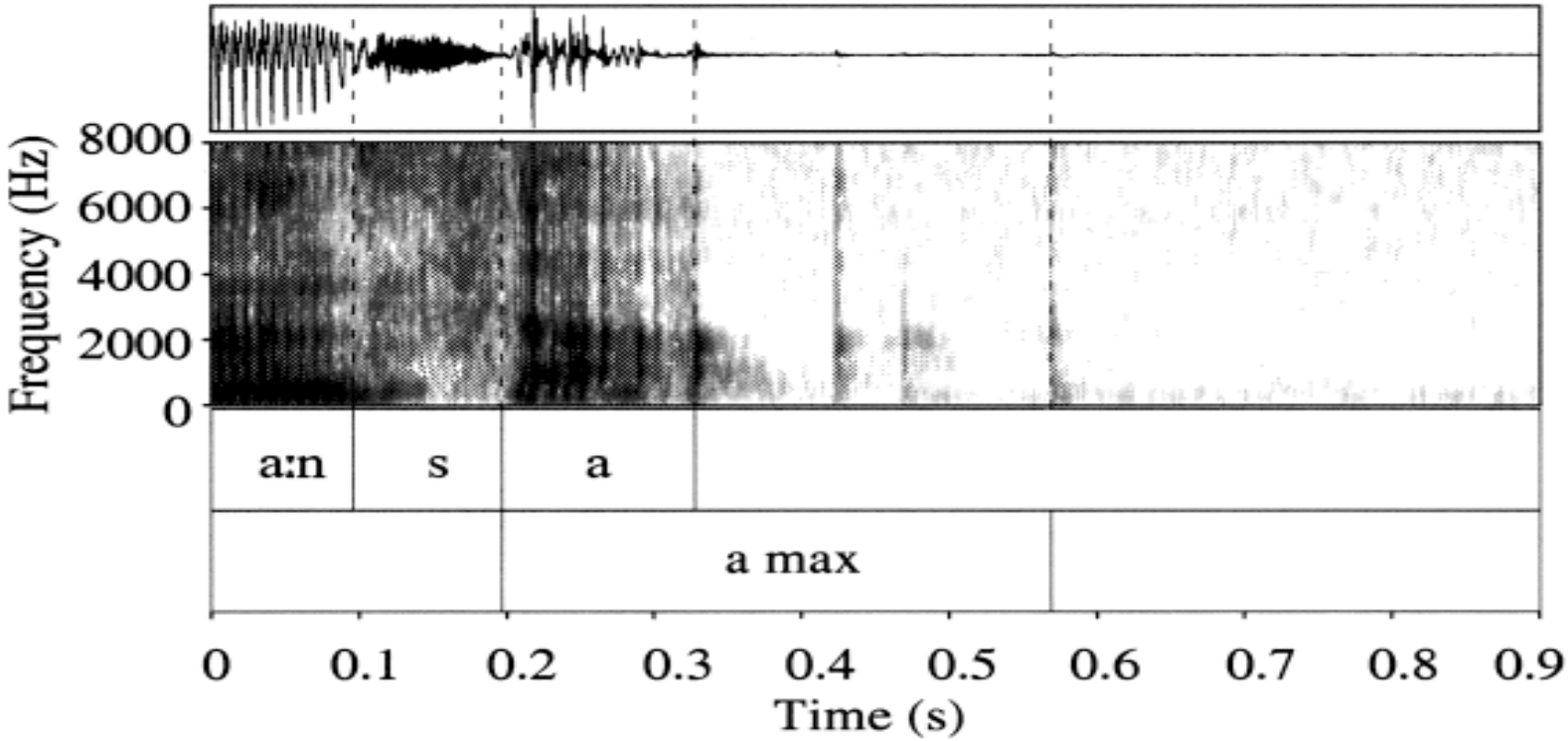


Figure 12: A fragment from *Toujou-sensei-ni kii-tara ICHI-BAN-ga* "*saansa*" 'According to Mr. Tojo, NUMBER ONE is "saansa"', spoken by a male Standard Japanese speaker. *Saansa* is a nonsense word. The boundary for the offset of /a/ is placed on the last glottal pulse peak in the interval delimited by continuous F2.

Pre-pausal durations: Final breathiness

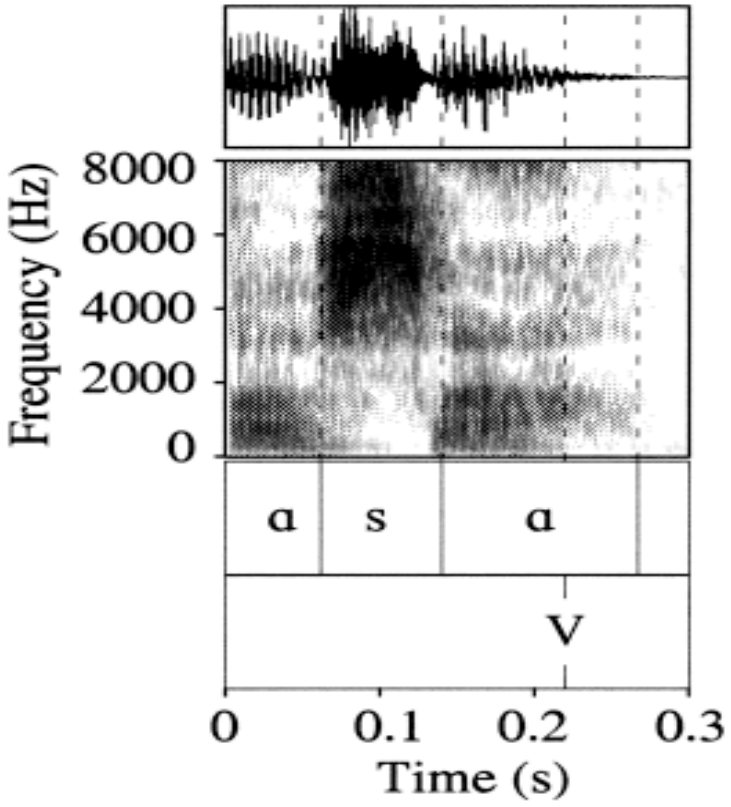


Figure 13: A fragment from *Kohtaan SEITSEMÄNSATAA voisi vastata* "*sasa*" 'For [#] SEVEN-HUNDRED [you] could answer "sasa"', spoken by a female Northern Finnish speaker. *Sasa* is a nonsense word. *V* in the second label tier indicates the offset of voicing for [ɑ].

Acoustic measurements: Summary

- Reliable landmarks for
  - stop, fricative, affricate & some oral constrictions for nasal stops
    - In intervocalic contexts
- Segments that present measurement difficulties
    - Approximants
    - Some clusters
    - Phrase-final, pre-pausal segments
- Experimental design should take these issues into account
- Data reliability depends on segmentation reliability

Measuring durations from articulatory records

- Electro-palatography

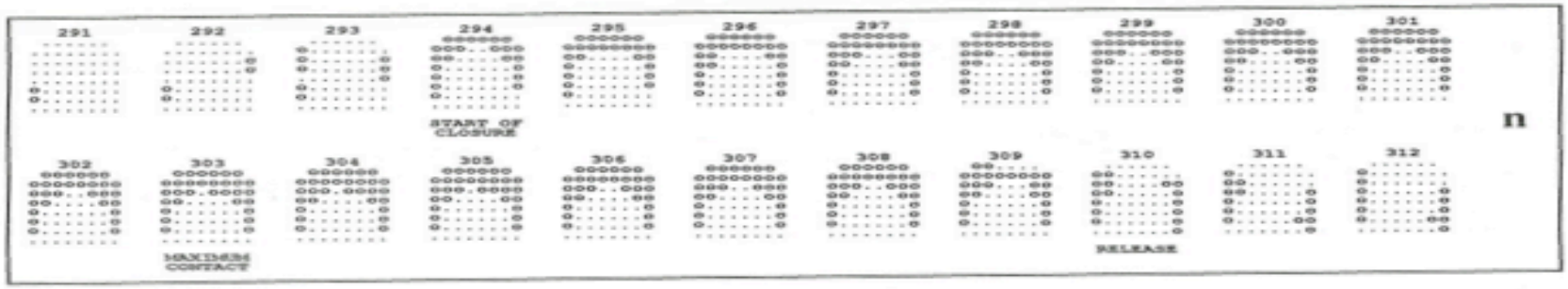
Measures tongue contact with roof of the mouth for each time frame (typical resolution 1 frame each 5 ms).

- Easy to identify closure/constriction intervals
- No confusion of oral vs. laryngeal activity
  - Only measures oral activity, i.e. contact of tongue on hard palate

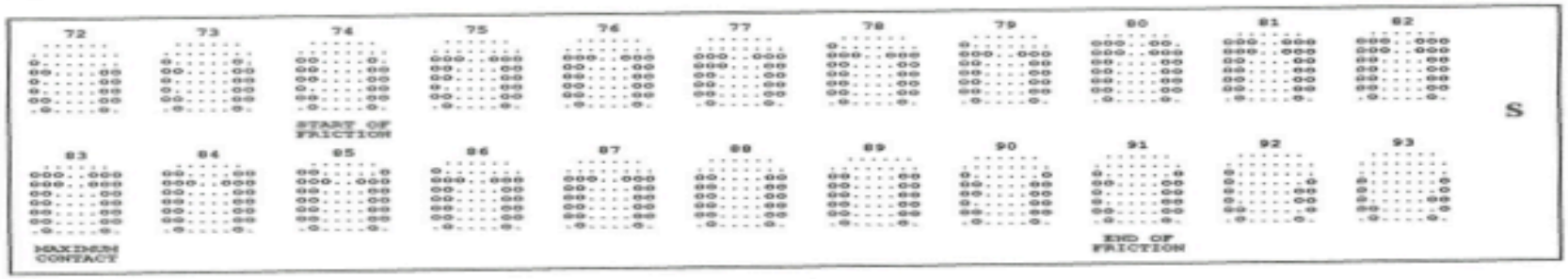
Figure from Gibbon, Stewart, Hardcastle & Crampin 1999

FIGURE 5. Examples of Robbie's normal EPG patterns pretreatment. Note the horseshoe shape configuration at maximum con for /n/ (frame 303), and the lateral bracing evident at maximum contact for /n/ (frame 303) and also for /s/ (frame 83). EPG patter for /s/ show evidence of a normal anterior groove configuration (frame 83). Also note production of /k/, which at maximum cont (frame 191) shows appropriate contact in the velar region, and less lateral and alveolar contact than the alveolar targets in Figu

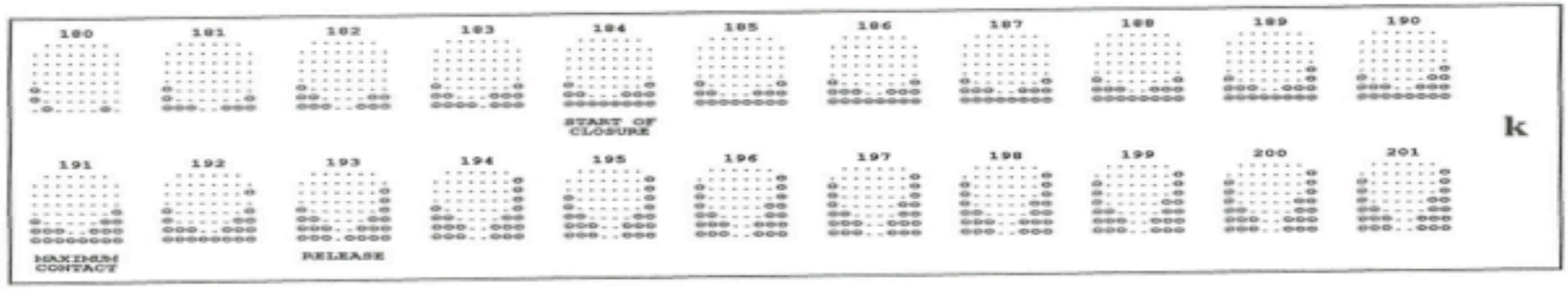
(a) Target /n/ (word-initial /n/ in a nest, transcribed as [n])



(b) Target /s/ (word-initial /s/ in sob, transcribed as [s])



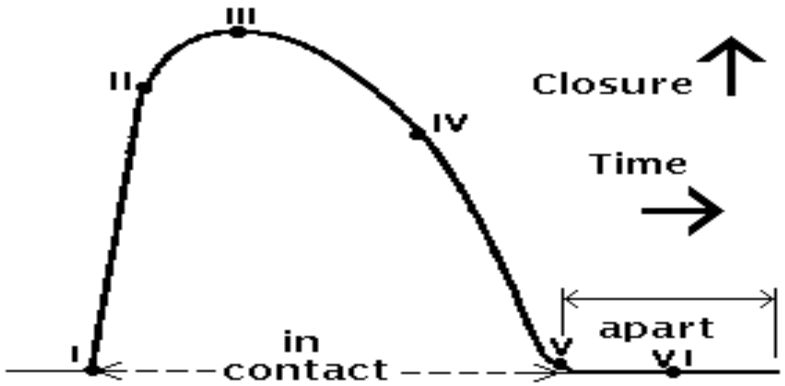
(c) Target /k/ (word-initial /k/ in Karen, transcribed as [k])



Measuring durations from articulatory records

- Laryngography
  - Can measure time intervals when vocal folds are closed/partially closed vs. open. Gives a signal proportional to vertical vocal fold contact area.. Figure from http://www.reading.ac.uk/AcaDepts/ll/speechlab/multichannel/lx/

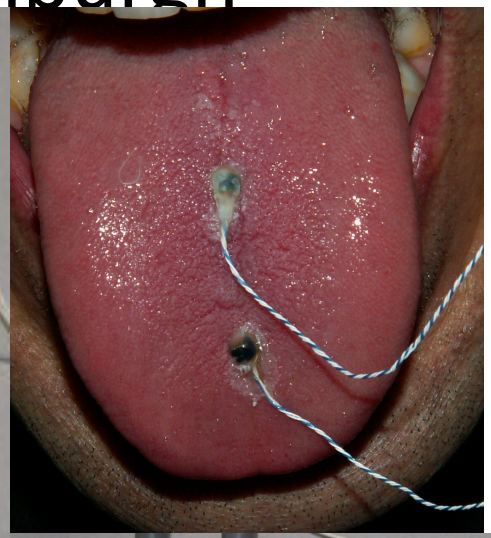
Vocal fold closure sequence.

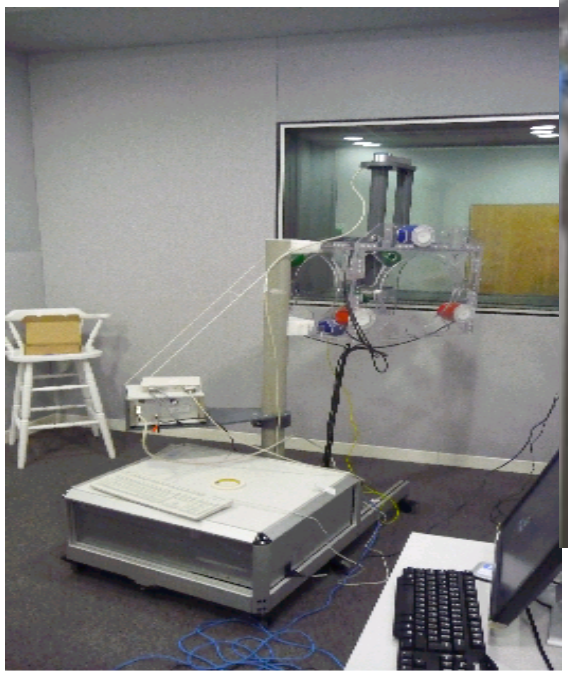


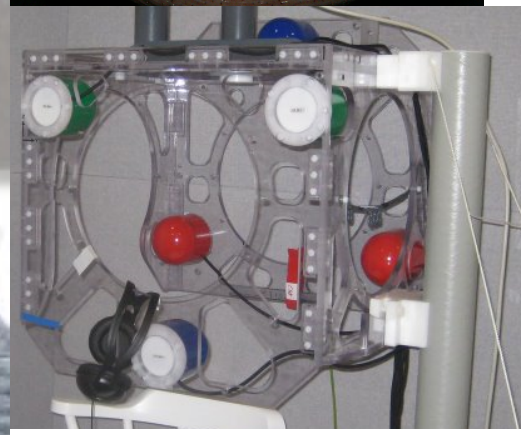
Fleshpoint tracking

- E.g. electromagnetic articulometry tracks the movements of sensors glued to the lips, tongue, jaw, head.

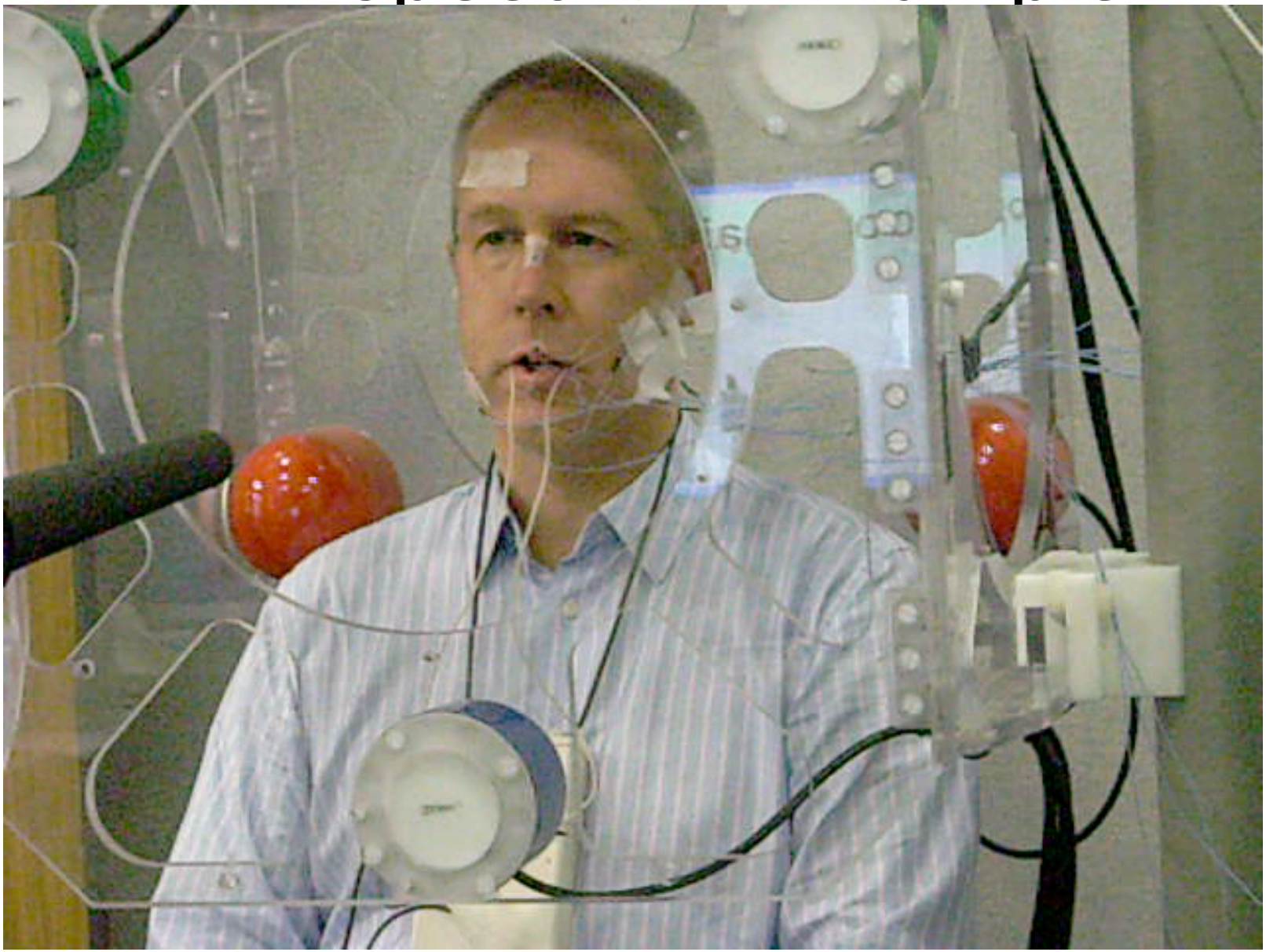
Electromagnetic Articulography: Carstens' AG500 EMA system, U. of Edinburgh



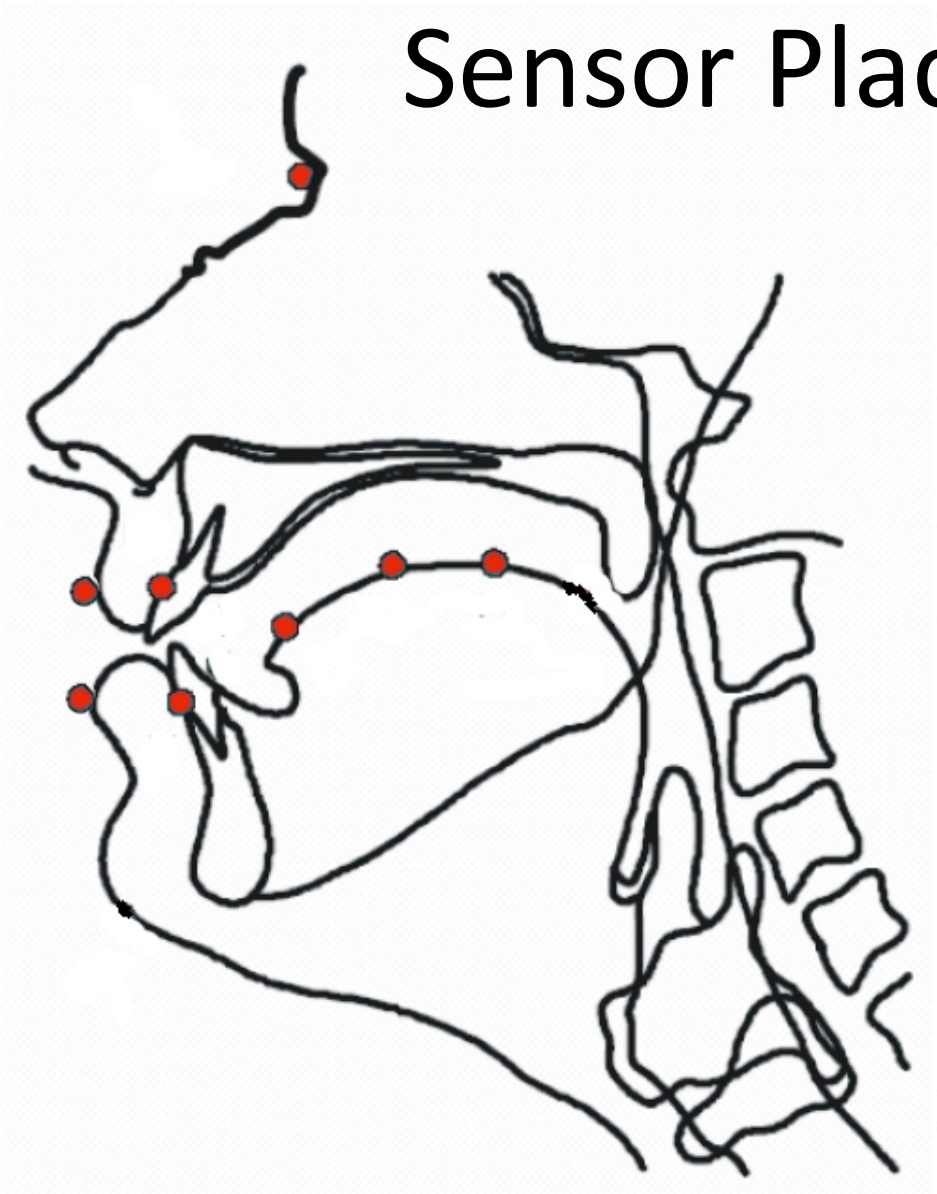




EMA Speech: An Example



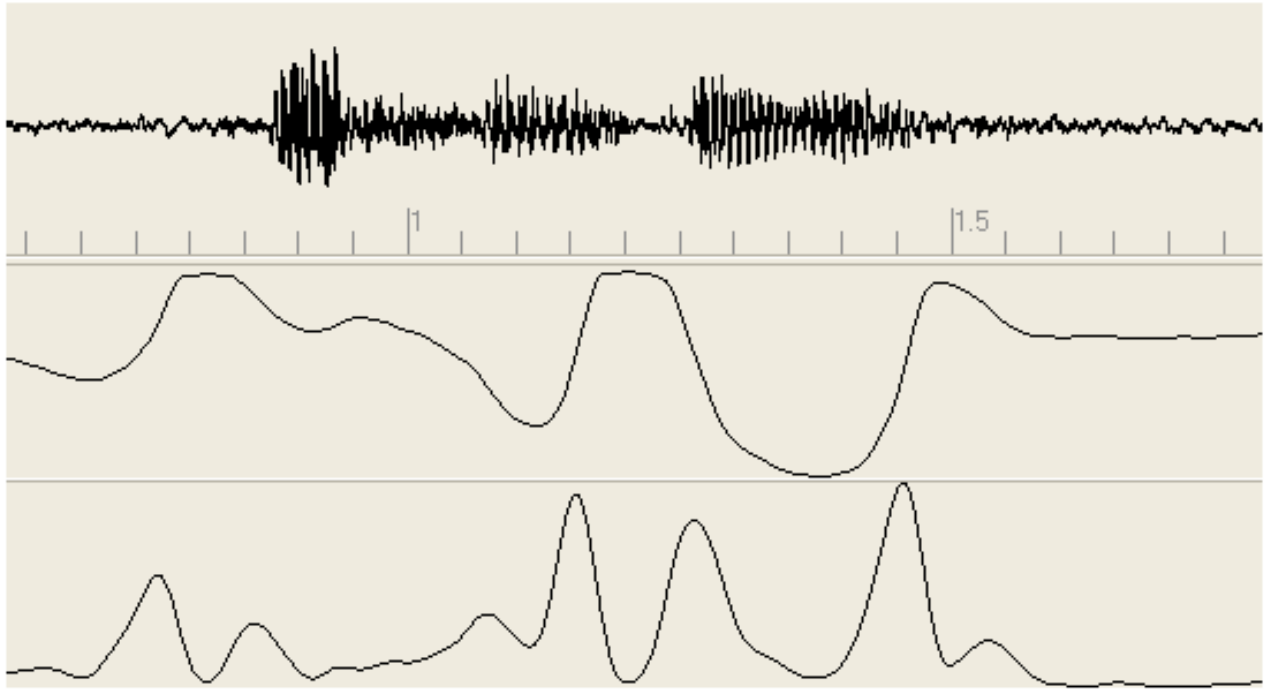
Sensor Placement



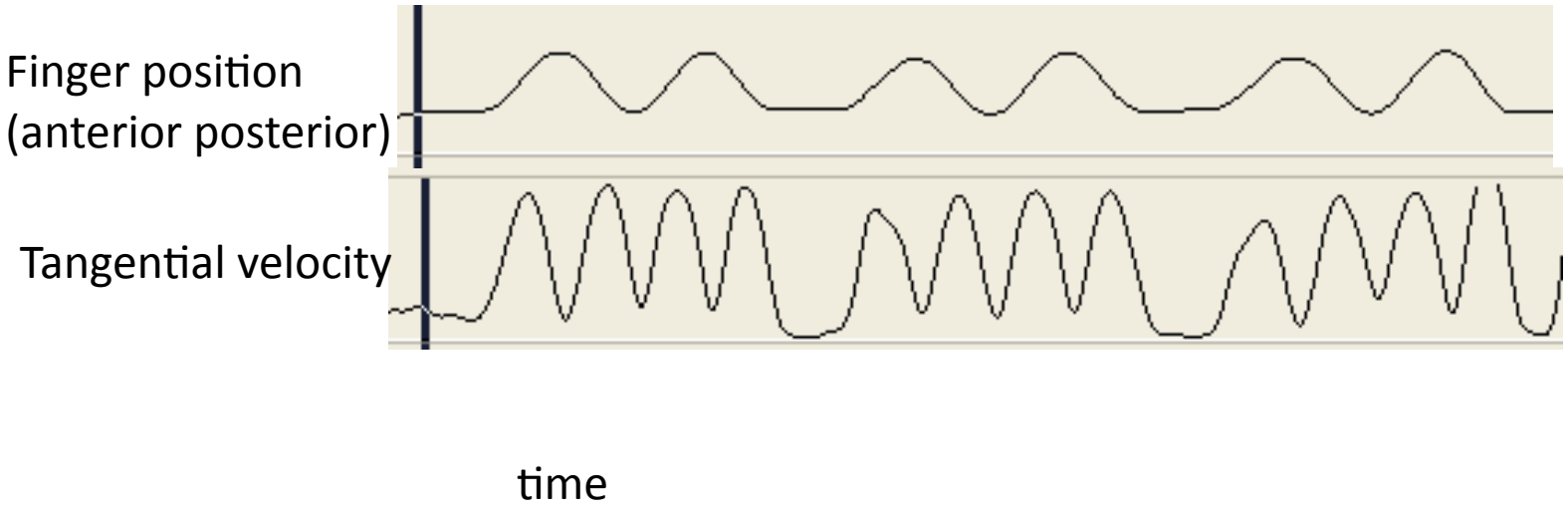
Additional sensors are placed behind both ears

‘Please say dad.’

d a d



Same types of movements for non-speech: Example finger movements, Tracing zigzags on paper



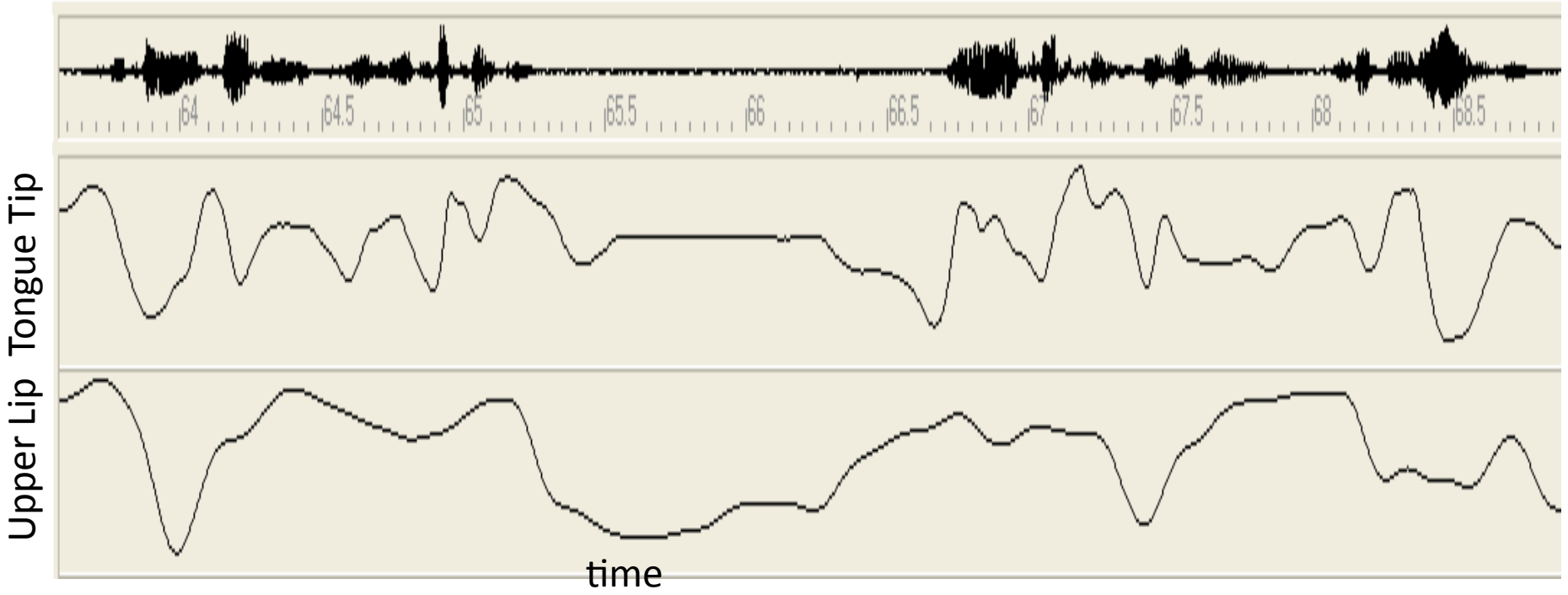
Fleshpoint tracking

- Allows us to simultaneously measure movements of different articulators
  - Inter-articulator timing/coordination
    - Articulatory overlap

Fleshpoint tracking

Different articulators move at different rates (e.g. lips slower than tongue tip)

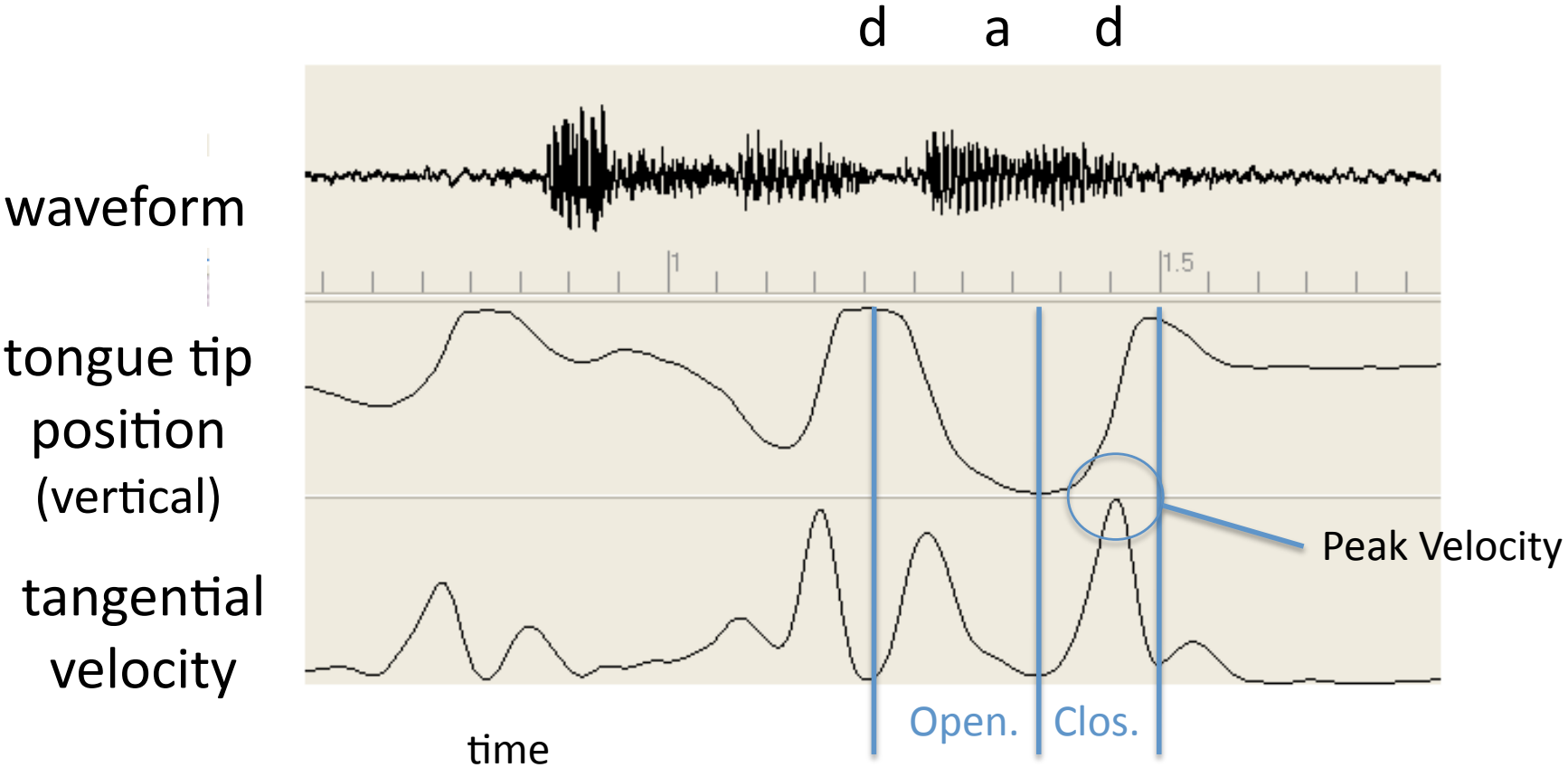
She came across a little cottage…and in the cottage were living … seven dwarves



Movements towards and away from target positions

- Caveat: Difficult/impossible? to identify spatial *targets*
  - Potential difference between what we are aiming to reach and the point we do reach
  - We can measure the points we do reach
- Movement intervals can be defined by
  - Points where articulators slow down before speeding up again
    - Valleys in the tangential velocity signal
- Can also define intervals where articulators are relatively stationary

'Please say dad.'



- For opening and closing intervals, we can measure
  - Durations
  - Distance an articulator moves
  - Peak velocity
  - Minimum velocity
  - How velocity changes over time (velocity profiles)
  - Also acceleration

Relationship between intervals durations (as measured on acoustic records) and properties of movements

- Fleshpoint tracking methods suggest that acoustic interval durational differences can be due to one or more articulatory control strategies, e.g.:
  - Change movement speed
  - Spending more time in target regions
  - Differences in articulatory overlap

What are we timing? How does timing control work?

- Articulatory speeds?
- Inter-articulatory overlap?
- Interval durations?
  - May find that the articulatory strategies don't matter

- CAVEAT: Measurement methods may bias our answer to this question

Preview

- Is speech timing systematic?
- If so, how is speech timing controlled?
  - What are we timing?
  - Which factors affect speech timing?
  - Which representations are involved?
  - To what extent does speech timing involve the use of general, non-speech specific control mechanisms?
- Matters of controversy
  - What is speech rhythm? Is speech rhythmic?
  - What types of constituents do we signal with duration?
  - Is predictability yet another factor that affects speech timing or does it affect duration via prosodic structure?

References

- Stevens, K.N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *Journal of the Acoustical Society of America 111 (4), 1872-1891.*
- Turk, A., Nakai, S. & Sugahara, M. (2006). Acoustic segment durations in prosodic research: a practical guide. In Sudhoff, Stefan, Denisa Lenertová, Roland Meyer, Sandra Pappert, Petra Augurzky, Ina Mleinek, Nicole Richter & Johannes Schliesser (eds): Methods in Empirical Prosody Research. Berlin, New York: De Gruyter (= Language, Context, and Cognition, 3), 1-28.