



KTH Datavetenskap
och kommunikation

Exploring prosody in interaction control

Jens Edlund & Mattias Heldner

KTH Department of Speech, Music and Hearing



Progress in Experimental Phonology: From Communicative Function to Phonetic Substance and Vice Versa



KTH Datavetenskap
och kommunikation

Special Issue: *Phonetica* 2005,
Vol. 62, No. 2-4

Abstract

In this seminar we present an investigation of prosodic aspects of turn-taking in conversation with a view to improving the efficiency of identifying relevant places at which a machine can legitimately begin to talk to a human interlocutor.



KTH Datavetenskap
och kommunikation

We examine the relationship between interaction control, the communicative function of which is to regulate the flow of information between interlocutors, and its phonetic manifestation.

Specifically, the listener's perception of such interaction control phenomena is modelled. Algorithms for automatic online extraction of prosodic phenomena liable to be relevant for interaction control, such as silent pauses and intonation patterns, are presented and evaluated in experiments using Swedish Map Task data.

We show that the automatically extracted prosodic features can be used to avoid many of the places where current dialogue systems run the risk of interrupting their users, and also to identify suitable places to take the turn.

Outline



KTH Datavetenskap
och kommunikation

- Motivation
 - improving spoken dialogue systems
- The relationship between
 - *interaction control*
 - its manifestation in the conversation
- **/nailon/**
 - automatic online extraction of prosodic phenomena
- Evaluation using Swedish Map Task data
- Automatically extracted prosodic features helps
 - avoiding many places where current dialogue systems risk interrupting their users
 - identifying suitable places to take the turn

Projects



KTH Datavetenskap
och kommunikation

- GROG
- AdApt
- CHIL
- Higgins

Goal



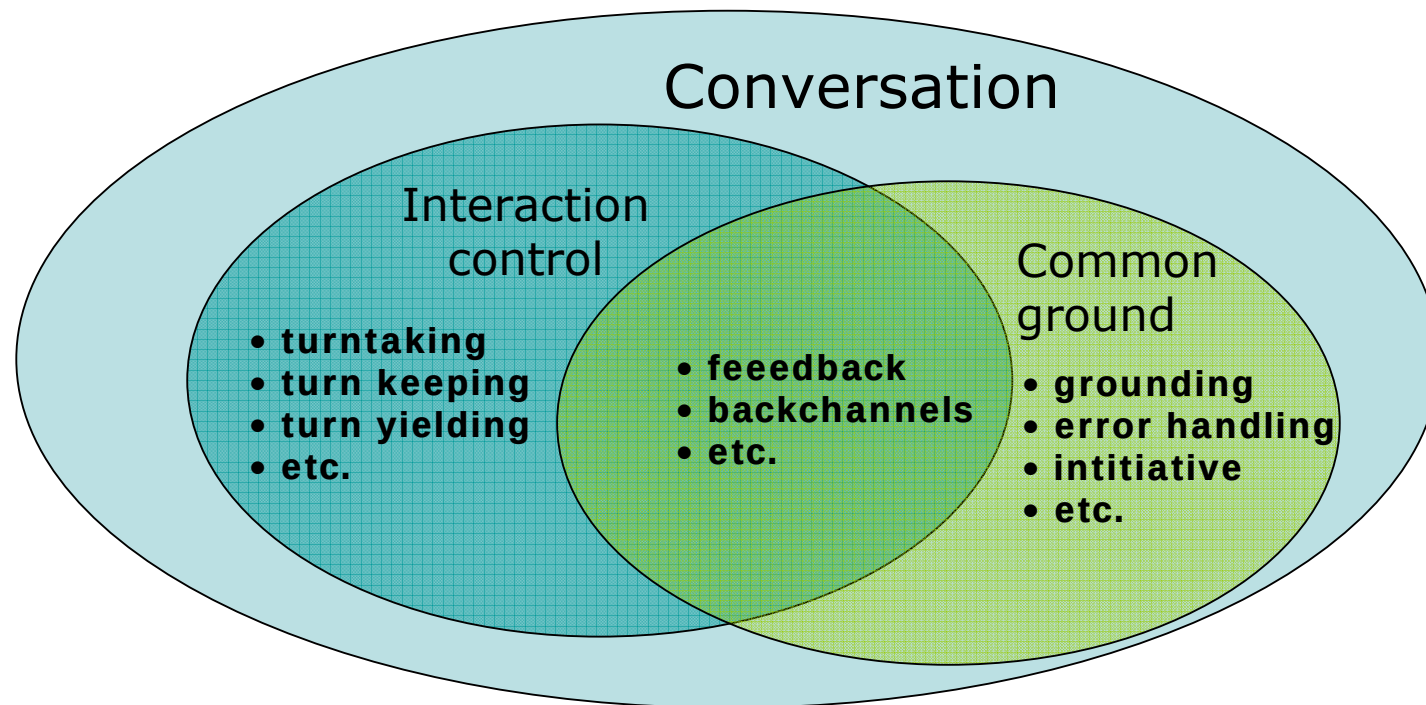
**KTH Datavetenskap
och kommunikation**

Investigate conversation with a view to improving the efficiency of identifying relevant places at which a machine can legitimately begin to talk to a human interlocutor.

Conversation & interaction control



KTH Datavetenskap
och kommunikation



Interaction control



KTH Datavetenskap
och kommunikation

- Regulate the flow of information between **interlocutors** (**speakers** and **listeners**) to make it proceed smoothly and efficiently
- Collaborative effort where interlocutors continuously monitor various aspects of each other's behaviour in order to make decisions about **turn-taking** and **feedback**
- Interaction control includes, for example,
 - what the **speaker** does to keep the floor, i.e. **turn-keeping**
 - or to hand over the floor, i.e. **turn-yielding**
 - how the **listener** finds suitable places to take the floor
 - or to give feedback to the speaker



Example

A: och då är jag väldigt nära den där kust<ehm>remsan och utanför min kustremsa så ligg finns det sjöhästar

B: mm dom har jag också

A: ja och precis <> ja strax söder om dom där sjöhästarna så når min bana nästan ända fram till kust<>linjen där

B: okej

A: och sen så fortsätter vi ned och gör en mjuk in<>buktning åt österut och rundar en stor förskräcklig fågel där

B: mm jag har en förskräcklig fågel också men jag undrar om det är samma för min fågel är en bit norrut

A: ja

B: min fågel är i nedre högra hörnet av en <> av den nordligaste bukten på västra sidan av ön

A: ja just det

B: okej

A: mm



KTH Datavetenskap
och kommunikation

Features relevant for interaction control



KTH Datavetenskap
och kommunikation

- Auditory
 - **Silent pauses**
 - Intonation patterns
 - Creaky voice
 - Vocal tract configuration (open/closed)
- Visual
 - Nods
 - Glances
 - Mimicry Gestures
- Structural (in)completeness
 - Semantic
 - Pragmatic
 - Syntactic

Current dialogue systems



KTH Datavetenskap
och kommunikation

- Use silence
- VAD, SAD, EOU, EOS, EPD...
 - ...are all basically silence duration thresholds
 - 500ms – 2000ms (sic!)

Human speakers



KTH Datavetenskap
och kommunikation

- Frequently pause before they are finished
- These pauses are often longer than 2000 ms
- For example when hesitating

Humans talking to dialogue systems



KTH Datavetenskap
och kommunikation

- Will get long response times, but...
- ...will also run the risk of being interrupted
- Limits to in-speech pause length???
 - Can we increase the 2000ms further to eliminate system interruptions?
 - Other methods?

Features relevant for interaction control



KTH Datavetenskap
och kommunikation

- Auditory
 - Silent pauses
 - Intonation patterns
 - Creaky voice
 - Vocal tract configuration (open/closed)
- Visual
 - Nods
 - Glances
 - Mimicry Gestures
- Structural (in)completeness
 - Semantic
 - Pragmatic
 - Syntactic

Method: outline



KTH Datavetenskap
och kommunikation

- Map task dialogues
- Segmentation into pause bounded units – IPU
- Classification of IPU into speaker changes and speaker holds
- Prosodic features extracted from the region immediately before the IPU boundary using /nailon/
- Turn-taking decisions using prosodic features
- Evaluation of turn-taking decisions with respect to the speaker change vs. speaker hold classification
- Additional mark-up and analyses
 - Perceptual judgments as to whether the IPU were finished or not on a 5-point scale by three judges
 - Extraction of intercontribution intervals (ICIs)

Speech material



KTH Datavetenskap
och kommunikation

- Map task dialogues
 - Designed to elicit natural-sounding spontaneous dialogues
 - Instruction giver & instruction follower
 - One map each, but maps not identical
 - Instruction giver describe a route to the follower
- Swedish map task dialogues
 - 4 dialogues: two pairs of speakers (each speaker acted as giver once and as follower once)
 - 1100 dialogue contributions (incl. feedback/backchannels)
 - Near perfect separation of giver and follower channels
 - Total duration 1 hour
 - Many thanks to Pétur Helgason at Stockholm University!

IPU segmentation



KTH Datavetenskap
och kommunikation

- Interpausal units (**IPUs**)
- Giver and follower channels automatically segmented into **speech** vs. **silence** using a basic speech activity detector (SAD) with some smoothing
- **IPU** = Transition from speech to long enough silence (>300 ms) in giver channel with no overlapping speech in follower channel
- Inter contribution intervals (**ICIs**) = actual duration of silent pauses between IPUs extracted automatically

Speaker change vs. speaker hold classification



KTH Datavetenskap
och kommunikation

- Each IPU classified as either **speaker change** or **speaker hold** automatically
- **Speaker change** = speech in the giver channel followed by at least 300 ms silence in the same channel, and non-overlapping speech in the follower channel
 - Minimum inter contribution interval (**ICI**) in a speaker change is 10 ms
- **Speaker hold** = speech in the giver channel followed by at least 300 ms silence in the same channel, and then more speech in the giver channel
 - Minimum inter contribution interval (**ICI**) in a speaker hold is 300 ms

Speaker change illustration

Giver channel:	[...] Speech		Long enough silent pause [...]	
		ICI		
Follower channel:	[...] Long enough silent pause		Speech [...]	

Speaker change vs. speaker hold used as gold standard



KTH Datavetenskap
och kommunikation

- Shows the actual turn of events in the dialogue
 - Is a direct reflection of the interlocutors' behaviour
 - Ensures that speaker changes and speaker holds were perceived as such by the interlocutors
- Does not show how things must be by necessity!
 - A speaker hold may be a suitable place to give a contribution except one where the other simply refrained from saying something
 - A speaker change may be an unsuitable place to give a contribution if the speaker was interrupted
- Makes no distinction between 'turns' and 'backchannels'
 - An appropriate place for a backchannel may not be appropriate for any other contributions than backchannels

Slight detour...



**KTH Datavetenskap
och kommunikation**

Is *silence only* a problem?



KTH Datavetenskap
och kommunikation

- **52.3 %** of all IPUs were **speaker holds**
 - Some of these may have been potential places for speaker changes
- **36.7 %** of all IPUs were judged as **unfinished** by the human judges
 - These are the cases where a dialogue system using silence only runs the risk of interrupting its users
- Silence only is a substantial problem!

Would increasing the silence thresholds help?



KTH Datavetenskap
och kommunikation

Silence threshold	>300 ms	>500 ms	>1000 ms	>1500 ms	>2000 ms
Number of IPUs	634	441	168	69	22
Speaker holds	68%	71%	68%	67%	55%

- Less than 50% of the ICIs were longer than 500 ms
 - Waiting for 500 ms (or more) is not the way humans does it
- The percentage of speaker holds virtually unchanged with longer ICIs
- Increasing silence thresholds leads to sluggish behaviour, only!

Prosodic analysis



KTH Datavetenskap
och kommunikation

Investigate conversation with a view to improving the efficiency of identifying relevant places at which a **machine** can legitimately begin to talk to a human interlocutor.

- online
 - no right context
- realtime
 - responsive
 - predictable
- general
 - any speaker
 - any domain

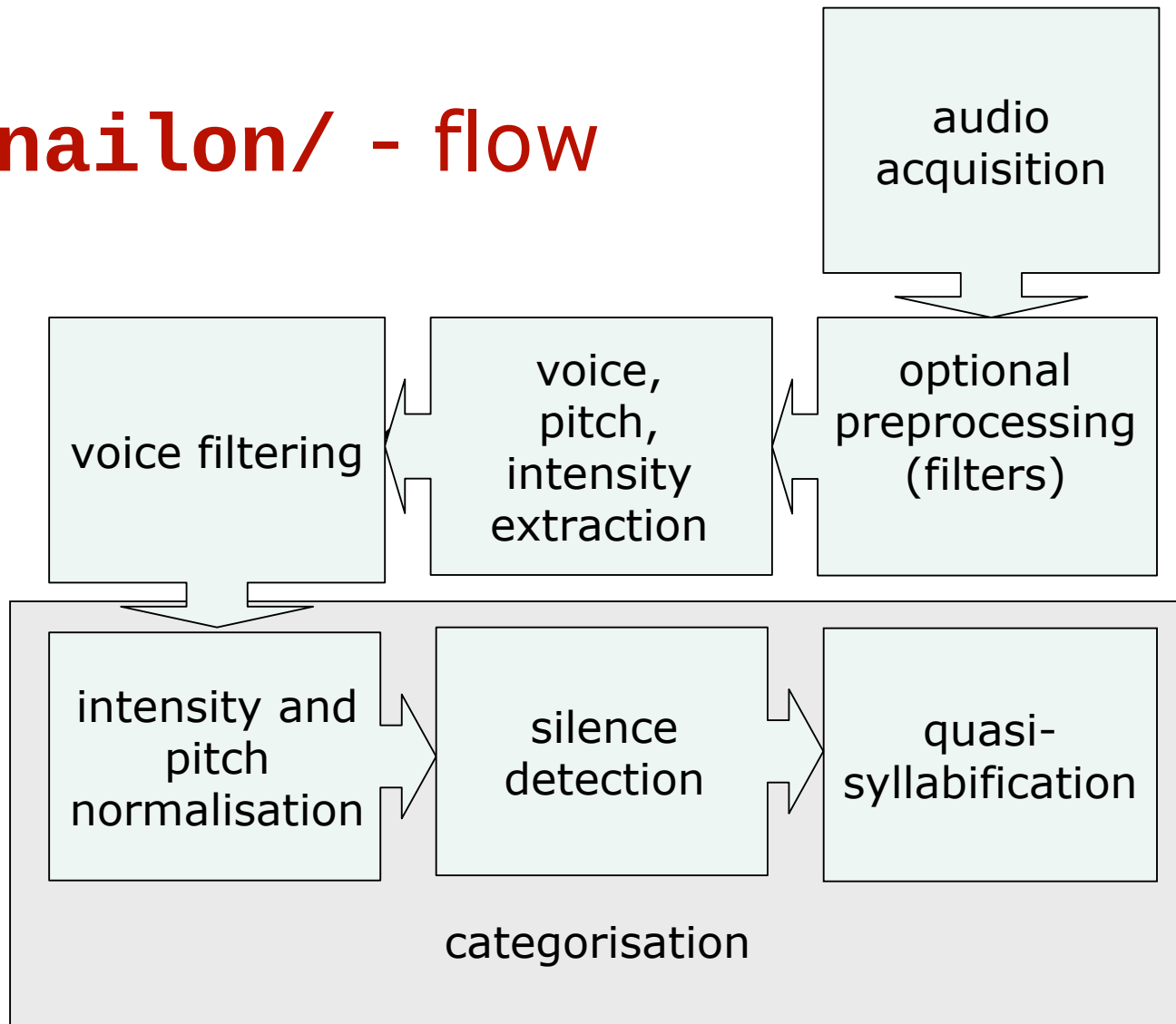
/nailon/



KTH Datavetenskap
och kommunikation

- Software for analysis of prosodic features
- Scripting in Tcl/Tk
- Based on Kåre Sjölander's Snack
- ESPS F_0 extraction

/nailon/ - flow



**KTH Datavetenskap
och kommunikation**

/nailon/ - audio acquisition



KTH Datavetenskap
och kommunikation

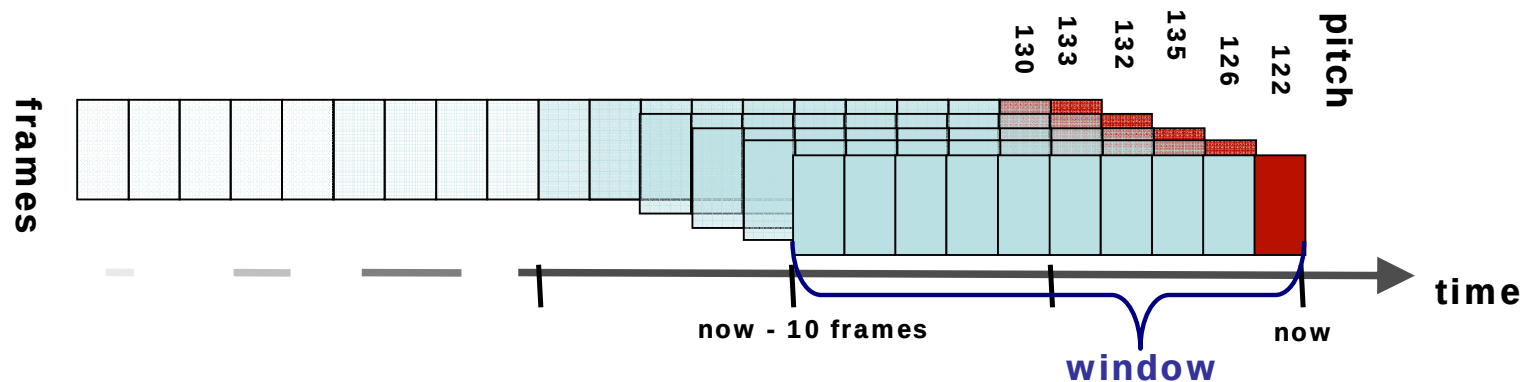
- Snack sound object
- Fixed size moving window
- Frees memory continuously
- Small footprint
- Optional pre-processing
 - e.g. Snack filters

/nailon/ - voice, pitch, intensity extraction

- Basic extraction by Snack/ESPS but
- Incremental:
 - repeated over fixed size moving window
 - last frame only
 - outputs realtime data stream (latency $\approx$ frame length)



KTH Datavetenskap
och kommunikation



/nailon/ - voice, pitch, intensity



KTH Datavetenskap
och kommunikation

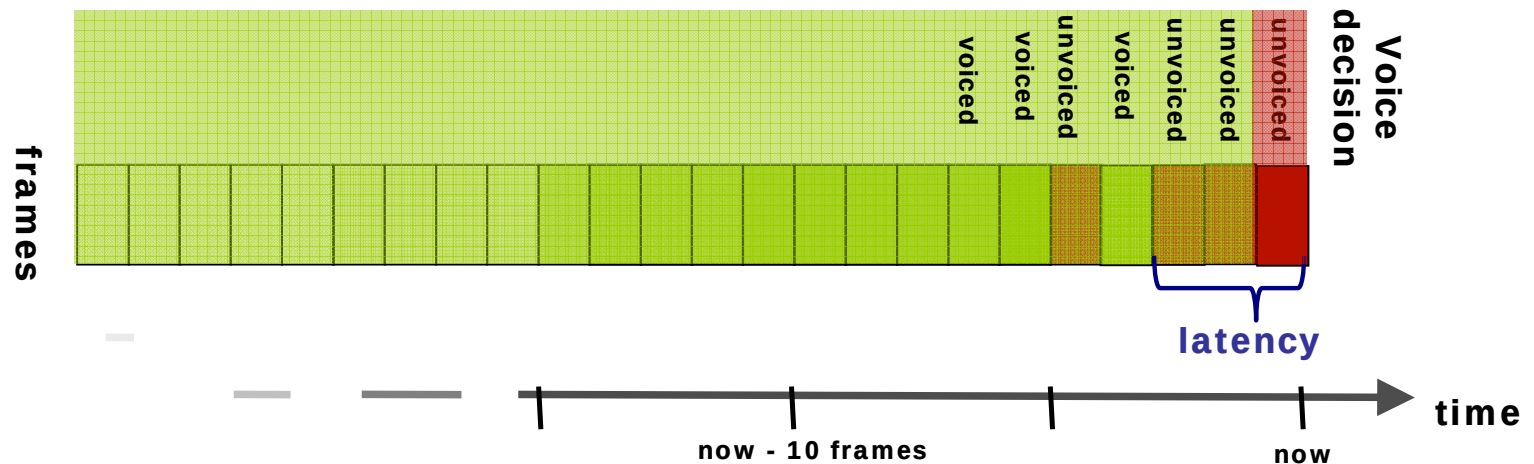
- Pitch semitone transform (optional)
- Intensity dB transform (optional)
- Sanity checks
 - Pitch octave errors
 - Intensity spikes

/nailon/ - voice decision

- Currently only interested in voiced segments
- Robust "non-flimsy" voice decision
 - Cost: a small latency
 - Requires a certain number of frames to be judged the same for a change to take place
 - 3 frames, for example, introduces 3/framerate seconds of latency



KTH Datavetenskap
och kommunikation



/nailon/ - normalisation

- Absolute pitch and intensity numbers fluctuate too much
- Profiles introduce prerequisites:
 - speaker identity
 - acoustics
 - channel
- Online incremental calculation of mean and stddev
 - stabilises surprisingly quickly (< 30s speech needed)
 - degeneration can be used to ensure flexibility (not currently implemented)

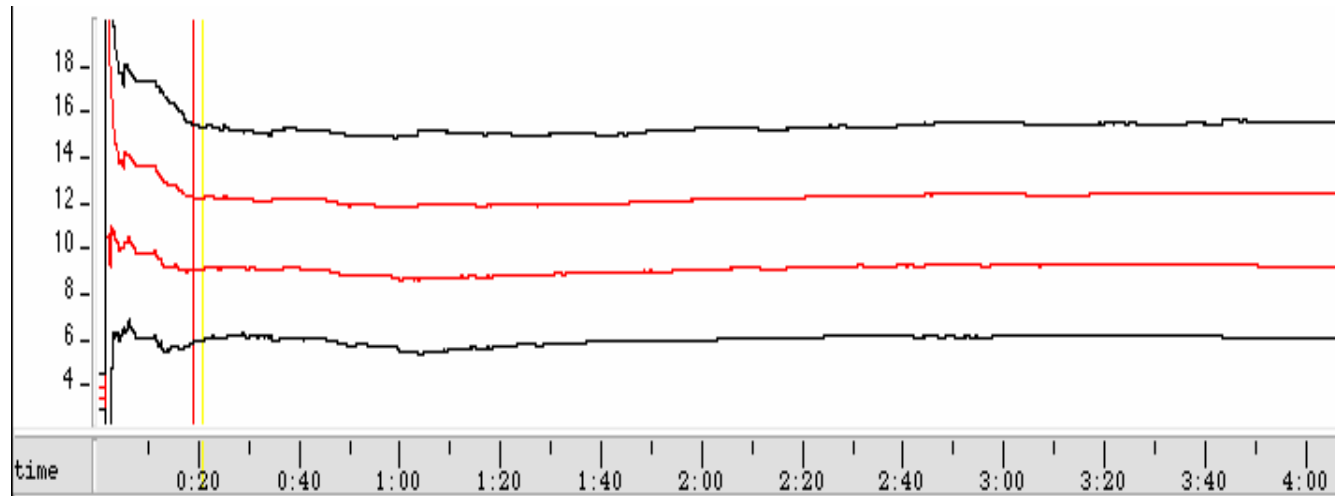


KTH Datavetenskap
och kommunikation

/nailon/ - F0 example



KTH Datavetenskap
och kommunikation



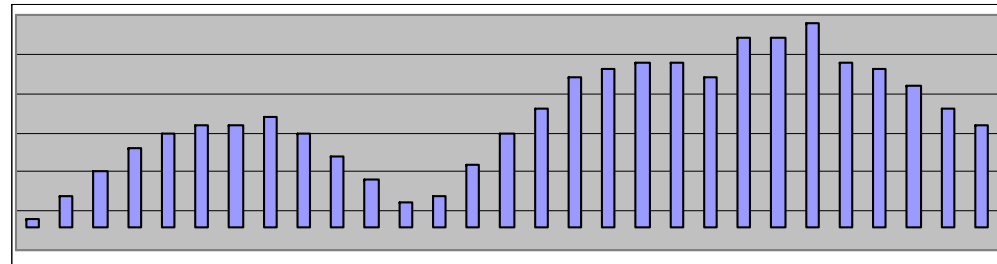
- Cumulative mean ± 2 standard deviations based on semitone transformed F0 data
- Voiced sequences only
- Stabilises after about 20 seconds
- High, mid and low registers

/nailon/ - silence detection

- Based on threshold of sequential silence
 - Intensity threshold re-calculated continuously
 - Intensity threshold currently the valley following the first peak in an intensity histogram



KTH Datavetenskap
och kommunikation



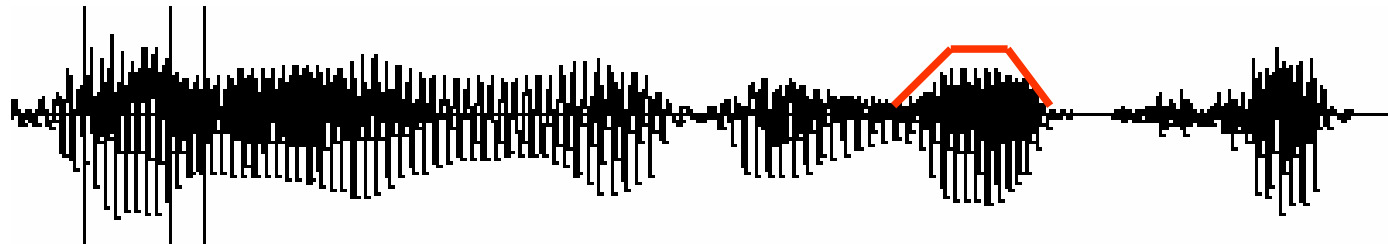
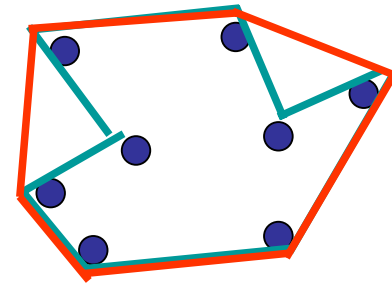
- Current duration threshold at 300ms of silence
- Reports what frame the silence *begun*

/nailon/ - quasi-syllabification



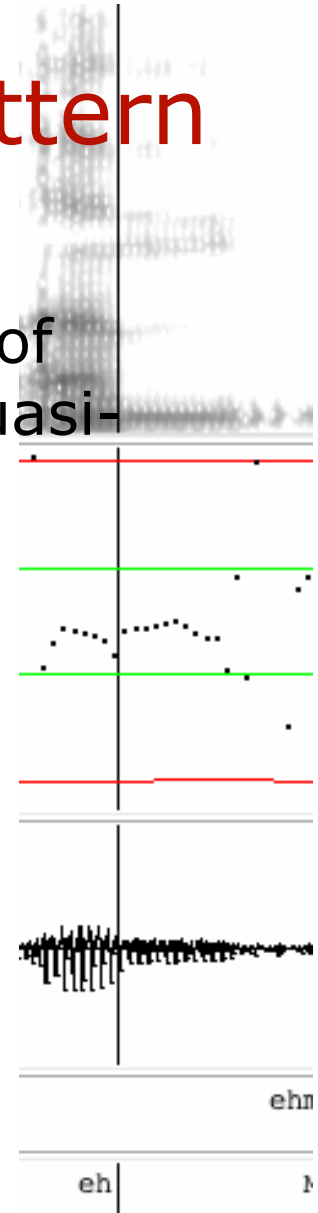
KTH Datavetenskap
och kommunikation

- Convex hulls (Mermelstein)
- Based on normalised intensity
 - Voiced sequences only
 - Incremental search for next complete hull
 - Remembers last seen hull only



/nailon/ - intonation pattern classification

- (Currently simplistic) classification of intonation over one convex hull (quasi-syllable) into either
 - Low or low and falling
 - Mid and level
 - Other



KTH Datavetenskap
och kommunikation

Turntaking decisions



KTH Datavetenskap
och kommunikation

- Low or low and falling intonation patterns taken to indicate suitable places for turn-taking **turn-yielding**
- Mid and level intonation patterns indicate unsuitable places **turn-keeping**
- Other intonation patterns may indicate turn-keeping as well as turn-yielding and were therefore classified as garbage **don't know** here

Does prosody help compared to silence (>300 ms) only?



KTH Datavetenskap
och kommunikation

	Turn-keeping	Don't know	Turn-yielding	Total
Change	23	212	158	393
Hold	105	255	71	431
Total	128	467	229	824

- 28% turn-yielding, 16% turn-keeping, 56% don't know
- Speaker changes were in the majority for turn-yielding (69%)
- Speaker holds were in the majority for turn-keeping (82%)
- Unobtrusive system (pooling turn-keeping & don't know): identifies 41% of the suitable places for turn-taking; avoids 84% of the impossible ones
- Responsive system (pooling turn-yielding & don't know): identifies 94% of the suitable places for turn-taking; avoids 24% of the unsuitable ones

Responsivity vs. unobtrusiveness



KTH Datavetenskap
och kommunikation

- All dialogue systems need some kind of interaction control capabilities
- More human-like systems will require more human-like interaction control
- Different dialogue situations put different demands on the interaction control capabilities

KTH Connector example



KTH Datavetenskap
och kommunikation

S This is the KTH Connector, how may I help you?

C Hi, this is Mattias.

S Hello Mattias

C I was supposed to be at the meeting, but I'm stuck on a train.

S Mhm

C Could you check with the others if it's ok to patch me in as a listener?

S Ok, please hold for a minute.

C Ok.

S [The KTH Connector waits for a suitable place to notify the meeting]

S Mattias is on the phone and will not be able to make it to the meeting. Is it ok to patch him in as a listener?

P Sure, but give him a speech channel as well.

S Ok, I'll let you know when it's done.

S Ok, Mattias, I'll patch you in. They suggested I'd give you a speech channel as well. Do you want one?

C No, it's too noisy on the train. Listening will do.

Responsivity vs. Unobtrusiveness (cont.)



KTH Datavetenskap
och kommunikation

- In dialogues with one listener only, responsiveness may be a key concern for the system, as sluggishness is likely to annoy the user
- Sometimes unobtrusiveness is the most important concern, for example when establishing a new connection for notifying the participant in a meeting

Future work

Better internal models of interaction control

- Further development of /nailon/
 - Adding features, e.g. distinguishing open vs closed vocal tract...
 - Machine learning of categorisation
- Combination with other sources of knowledge
 - Semantic completeness

Better models that relate interaction control to the bigger picture – conversation

- Relation to grounding, error handling
 - Utterance concept particularly difficult as regards e.g. backchannels
- Initiative



KTH Datavetenskap
och kommunikation

Perceptual judgments vs speaker change/speaker hold



KTH Datavetenskap
och kommunikation

	1 definitely unfinished	2 probably unfinished	3 could be finished or unfinished	4 probably finished	5 definitely finished	Total
Speaker change	169	66	40	200	705	1180
Speaker hold	806	91	25	112	240	1274
Total	975	157	65	312	945	2454