/**nailon**/ – software for online analysis of prosody

Interspeech 2006 special session: *The prosody of turn-taking and dialog acts*
September 20, 2006

Jens Edlund & Mattias Heldner
KTH Speech, Music and Hearing, Stockholm, Sweden

How is our work situated with respect to the wider field?

How is our work situated with respect to the wider field?

- We 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

How is our work situated with respect to the wider field?

- We 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

- /**nailon**/ – our software for online analysis of prosody – is an important tool in these **interaction control** studies

How is our work situated with respect to the wider field?

- We 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

- /**nailon**/ – our software for online analysis of prosody – is an important tool in these **interaction control** studies

- **Basic research** about human behaviour, potentially giving insights into the social dynamics of human interaction

How is our work situated with respect to the wider field?

- We 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

- /**nailon**/ – our software for online analysis of prosody – is an important tool in these **interaction control** studies
  - **Basic research** about human behaviour, potentially giving insights into the social dynamics of human interaction
  - Motivated by a wish to improve the interaction control (including the timing) in **spoken dialogue systems**

How is our work situated with respect to the wider field?

- We 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

- /**nailon**/ – our software for online analysis of prosody – is an important tool in these **interaction control** studies

- **Basic research** about human behaviour, potentially giving insights into the social dynamics of human interaction

- Motivated by a wish to improve the interaction control (including the timing) in **spoken dialogue systems**

- Automatic **online** methods based on **acoustic information** only

How is our work situated with respect to the wider field?

- We 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

- /**nailon**/ – our software for online analysis of prosody – is an important tool in these **interaction control** studies

- **Basic research** about human behaviour, potentially giving insights into the social dynamics of human interaction

- Motivated by a wish to improve the interaction control (including the timing) in **spoken dialogue systems**

- Automatic **online** methods based on **acoustic information** only

- Progress is gauged relative to current dialogue systems

Interaction control

- 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

Features relevant for interaction control

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

Current dialogue systems...

Current dialogue systems...

...use silence.

Current dialogue systems...

...use silence.

VAD, **SAD**, **EOU**, **EOS**, **EPD**...

...are all basically **silence duration thresholds**

Typical threshold values: 500 ms – 2000 ms

Human speakers...

Human speakers...

...frequently pause before they are finished

Human speakers...

...frequently pause before they are finished

– for example when hesitating

Human speakers...

...frequently pause before they are finished

– for example when hesitating

– these pauses are often longer than 2000 ms

Humans talking to dialogue systems...

Humans talking to dialogue systems...

...will get long **response times** (500 ms - 2000 ms), but...

Humans talking to dialogue systems...

...will get long **response times** (500 ms - 2000 ms), but...

...will also run the risk of being **interrupted**!

Humans talking to dialogue systems...

...will get long **response times** (500 ms - 2000 ms), but...

...will also run the risk of being **interrupted**!

In addition, they will also be **misunderstood** more often, as the system's speech understanding is likely to be impaired by badly segmented input

How to improve the situation?

How to improve the situation?

Increase the silence duration thresholds even more to reduce the number of interruptions – at the cost of even longer response times???

How to improve the situation?

Increase the silence duration thresholds even more to reduce the number of interruptions – at the cost of even longer response times???

No! The proportion of unfinished utterances remains virtually unchanged with increased thresholds...

How to improve the situation?

Increase the silence duration thresholds even more to reduce the number of interruptions – at the cost of even longer response times???

No! The proportion of unfinished utterances remains virtually unchanged with increased thresholds...

Add more relevant features!

How to improve the situation?

Increase the silence duration thresholds even more to reduce the number of interruptions – at the cost of even longer response times???

No! The proportion of unfinished utterances remains virtually unchanged with increased thresholds...

Add more relevant features!

We have tried making prosodic features, and in particular intonation patterns before silences, available to spoken dialogue systems through /**`nailon`**/

/**nailon**/ – software for online analysis of prosody

/**nailon**/ – software for online analysis of prosody

- Segmentation: speech vs. silence, and pseudo syllables

/**nailon**/ – software for online analysis of prosody

- Segmentation: speech vs. silence, and pseudo syllables
- Captures silence durations; voicing; intensity and pitch; with online normalisation (pitch and intensity) and incremental analyses

/**nailon**/ – software for online analysis of prosody

- Segmentation: speech vs. silence, and pseudo syllables
- Captures silence durations; voicing; intensity and pitch; with online normalisation (pitch and intensity) and incremental analyses
- Classification of intonation patterns within psyllables before silences

/**nailon**/ – software for online analysis of prosody

- Segmentation: speech vs. silence, and pseudo syllables
- Captures silence durations; voicing; intensity and pitch; with online normalisation (pitch and intensity) and incremental analyses
- Classification of intonation patterns within psyllables before silences
- All processing is online in the sense of relying on past and present information only – /**nailon**/ is a phonetic anagram of online...

/**nailon**/ – software for online analysis of prosody

- Segmentation: speech vs. silence, and pseudo syllables
- Captures silence durations; voicing; intensity and pitch; with online normalisation (pitch and intensity) and incremental analyses
- Classification of intonation patterns within psyllables before silences
- All processing is online in the sense of relying on past and present information only – /**nailon**/ is a phonetic anagram of online...
- Adds online and real-time abilities to the ESPS get_f0 function in the Snack Sound Toolkit

/**nailon**/ – software for online analysis of prosody

- Segmentation: speech vs. silence, and pseudo syllables
- Captures silence durations; voicing; intensity and pitch; with online normalisation (pitch and intensity) and incremental analyses
- Classification of intonation patterns within psyllables before silences
- All processing is online in the sense of relying on past and present information only – /**nailon**/ is a phonetic anagram of online...
- Adds online and real-time abilities to the ESPS get_f0 function in the Snack Sound Toolkit
- Scripted in Tcl/Tk. Performs in real-time, with a small and constant latency, footprint and flexible processor usage, on a standard windows PC (ought to run on Mac OS X and Linux too, but not tested yet)

Can /**nailon**/ be used to improve the interaction control in spoken dialogue systems? Method

---

Can /**nailon**/ be used to improve the interaction control in spoken dialogue systems? Method

- Swedish Map task dialogues (1 hour)

Can /**nailon**/ be used to improve the interaction control in spoken dialogue systems? Method

- Swedish Map task dialogues (1 hour)
- Automatic segmentation into pause bounded units – **IPUs**

Can /**nailon**/ be used to improve the interaction control in spoken dialogue systems? Method

- Swedish Map task dialogues (1 hour)
- Automatic segmentation into pause bounded units – **IPUs**
- Automatic classification of transitions from one IPU to another into **speaker changes** and **speaker holds**

Can /**nailon**/ be used to improve the interaction control in spoken dialogue systems? Method

- Swedish Map task dialogues (1 hour)
- Automatic segmentation into pause bounded units – **IPUs**
- Automatic classification of transitions from one IPU to another into **speaker changes** and **speaker holds**
- Automatic extraction of intonation patterns from the region immediately before the IPU boundary (i.e. the silence) using /**nailon**/

Can /**nailon**/ be used to improve the interaction control in spoken dialogue systems? Method

- Swedish Map task dialogues (1 hour)
- Automatic segmentation into pause bounded units – **IPUs**
- Automatic classification of transitions from one IPU to another into **speaker changes** and **speaker holds**
- Automatic extraction of intonation patterns from the region immediately before the IPU boundary (i.e. the silence) using /**nailon**/
- Turn-taking decisions using intonation patterns

Can /**nailon**/ be used to improve the interaction control in spoken dialogue systems? Method

- Swedish Map task dialogues (1 hour)
- Automatic segmentation into pause bounded units – **IPUs**
- Automatic classification of transitions from one IPU to another into **speaker changes** and **speaker holds**
- Automatic extraction of intonation patterns from the region immediately before the IPU boundary (i.e. the silence) using /**nailon**/
- Turn-taking decisions using intonation patterns
- Evaluation of turn-taking decisions with respect to the speaker change vs. speaker hold classification

Detail: Turn-taking decisions

- Based on observations in the literature – not trained
- **Low** or **low and falling** intonation patterns (i.e. low relative to the online normalisation of the speaker's pitch range) were taken to indicate **turn-yielding**, i.e. suitable places for turn-taking
- **Mid and level** intonation patterns were taken to indicate **turn-keeping**, i.e. unsuitable places for turn-taking
- Other intonation patterns (including rises) may indicate turn-keeping as well as turn-yielding and were classified as **don't know**, i.e. as **garbage**

Does /**nailon**/ improve the situation?

Does /**nailon**/ improve the situation?

- **Prosody** can be used to improve the efficiency of identifying relevant places at which a machine can legitimately begin to talk to a human interlocutor

Does /**nailon**/ improve the situation?

- **Prosody** can be used to improve the efficiency of identifying relevant places at which a machine can legitimately begin to talk to a human interlocutor
- Turn-taking decisions based on **/nailon/** features
  - **avoids** 84% of the unsuitable places – the risk of interrupting the user considerably decreased
  - **recognises** 40% of the suitable places

Does /**nailon**/ improve the situation?

- **Prosody** can be used to improve the efficiency of identifying relevant places at which a machine can legitimately begin to talk to a human interlocutor
- Turn-taking decisions based on **/nailon/** features
  - **avoids** 84% of the unsuitable places – the risk of interrupting the user considerably decreased
  - **recognises** 40% of the suitable places
- Considerable **responsivity gains**
  - turn-taking decisions made with a latency of 300 ms
  - to be compared with the typical response times in dialogue systems ranging from 500 ms – 2000 ms

Does /**nailon**/ improve the situation?

- **Prosody** can be used to improve the efficiency of identifying relevant places at which a machine can legitimately begin to talk to a human interlocutor
- Turn-taking decisions based on **/nailon/** features
  - **avoids** 84% of the unsuitable places – the risk of interrupting the user considerably decreased
  - **recognises** 40% of the suitable places
- Considerable **responsivity gains**
  - turn-taking decisions made with a latency of 300 ms
  - to be compared with the typical response times in dialogue systems ranging from 500 ms – 2000 ms
- No speech recognition, no pre-trained speaker models

Current work

Current work

Better internal models of interaction control

Current work

Better internal models of interaction control

- Further development of **`/nailon/`**
  - Adding features, e.g. distinguishing open vs. closed vocal tract, lengthening effects...
  - Machine learning of categorisation

Current work

Better internal models of interaction control

- Further development of **`/nailon/`**
    - Adding features, e.g. distinguishing open vs. closed vocal tract, lengthening effects...
    - Machine learning of categorisation

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

Current work

Better internal models of interaction control

- Further development of **`/nailon/`**
    - Adding features, e.g. distinguishing open vs. closed vocal tract, lengthening effects...
    - Machine learning of categorisation

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

- Relation to grounding, error handling

Current work

Better internal models of interaction control

- Further development of **`/nailon/`**
  - Adding features, e.g. distinguishing open vs. closed vocal tract, lengthening effects...
  - Machine learning of categorisation

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

- Relation to grounding, error handling

Integration in spoken dialogue systems

Current work

Better internal models of interaction control

- Further development of **`/nailon/`**
  - Adding features, e.g. distinguishing open vs. closed vocal tract, lengthening effects...
  - Machine learning of categorisation

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

- Relation to grounding, error handling

Integration in spoken dialogue systems

- Combination with other sources of knowledge such as semantic completeness

Prosody and turn-taking in 5 to 10 years

Prosody and turn-taking in 5 to 10 years

- In a long-term perspective, our goal is to design a spoken dialogue system that has **good-enough** conversational abilities for the users to consider it worthwhile having a **conversation** with – a system that is coherent with and can be understood using a **human metaphor.**

Prosody and turn-taking in 5 to 10 years

- In a long-term perspective, our goal is to design a spoken dialogue system that has **good-enough** conversational abilities for the users to consider it worthwhile having a **conversation** with – a system that is coherent with and can be understood using a **human metaphor.**
- In this line of work, we want to assess the importance of interaction control phenomena, such as **appropriate turn-taking behaviour**, **fast responses to greetings or channel checks**, and **well-timed verbal feedback** during the user's speech, for the perceived conversational ability of such dialogue systems.

Prosody and turn-taking in 5 to 10 years

- In a long-term perspective, our goal is to design a spoken dialogue system that has **good-enough** conversational abilities for the users to consider it worthwhile having a **conversation** with – a system that is coherent with and can be understood using a **human metaphor.**
- In this line of work, we want to assess the importance of interaction control phenomena, such as **appropriate turn-taking behaviour**, **fast responses to greetings or channel checks**, and **well-timed verbal feedback** during the user's speech, for the perceived conversational ability of such dialogue systems.
- It is our prediction that interaction is just as important as the content of the conversation. We may even attempt to build a dialogue system that masters interaction control, but that does not understand the meaning of words.

Prosody and turn-taking in 5 to 10 years

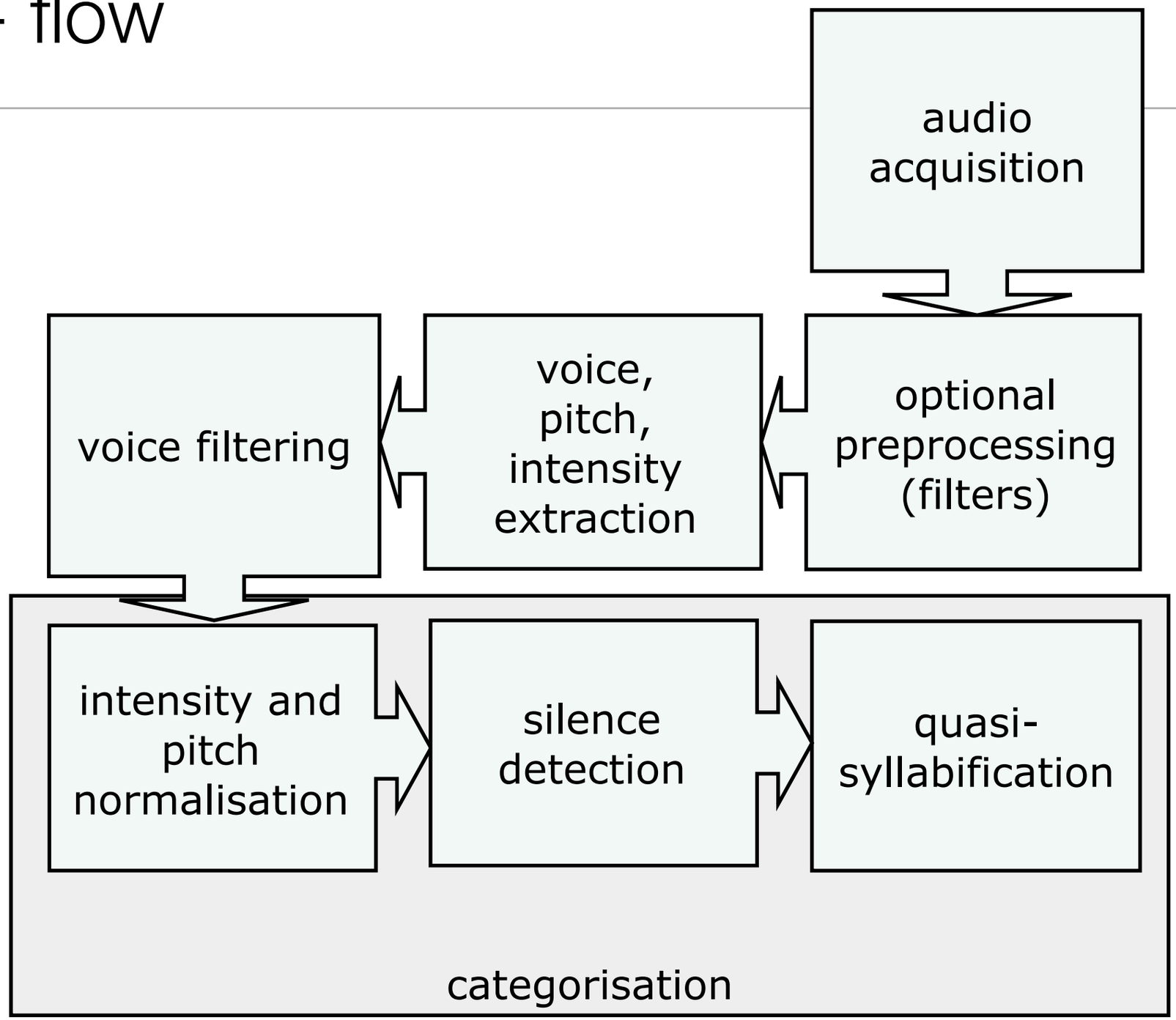
- In a long-term perspective, our goal is to design a spoken dialogue system that has **good-enough** conversational abilities for the users to consider it worthwhile having a **conversation** with – a system that is coherent with and can be understood using a **human metaphor.**
- In this line of work, we want to assess the importance of interaction control phenomena, such as **appropriate turn-taking behaviour**, **fast responses to greetings or channel checks**, and **well-timed verbal feedback** during the user's speech, for the perceived conversational ability of such dialogue systems.
- It is our prediction that interaction is just as important as the content of the conversation. We may even attempt to build a dialogue system that masters interaction control, but that does not understand the meaning of words.
- Ultimately, we need to combine the abilities of current dialogue systems with interaction control and other conversational phenomena such as grounding, error handling, pragmatic meaning conveyed by prosody, in order to create truly conversational systems.

/**nailon**/ will be made freely available.

Send an e-mail to edlund@speech.kth.se if you want to be notified.

Thank you for your attention!

/**nailon**/ - flow


audio acquisition
optional preprocessing (filters)
voice, pitch, intensity extraction
voice filtering
intensity and pitch normalisation
silence detection
quasi-syllabification
categorisation

Detail: Speaker change vs. speaker hold classification

- 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

Detail: A speaker change

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

Detail: Speaker change vs. speaker hold used as gold standard

- 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