KTH Speech, Music and Hearing

Centre for Speech Technology

Online detection of Vocal Listener Responses with Maximum Latency Constraints

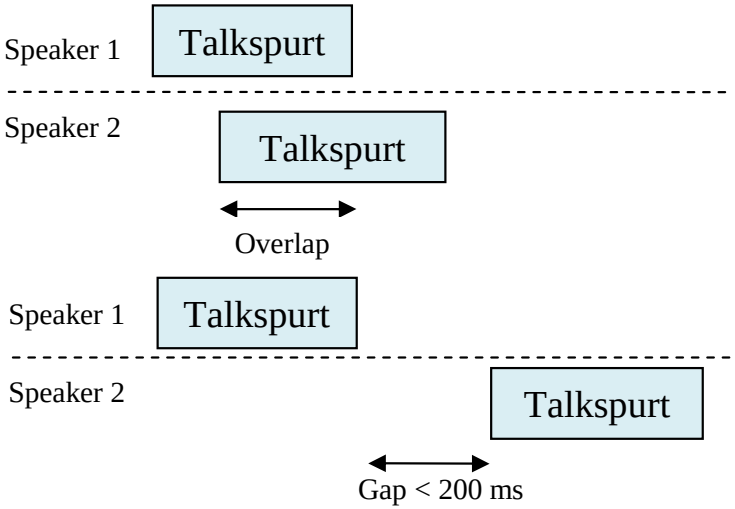
Daniel Neiberg (TMH/KTH)

Khiet P. Truong (Uni. of Twente)

Motivations

- Legacy conversational systems tend to be developed using the half duplex paradigm with long response times
- But 54-59% of all speaker shifts occur in overlap up to a 200 ms gap – the minimum reponse time to a pause (M. Heldner and J. Edlund, 2011)

Partial overlap



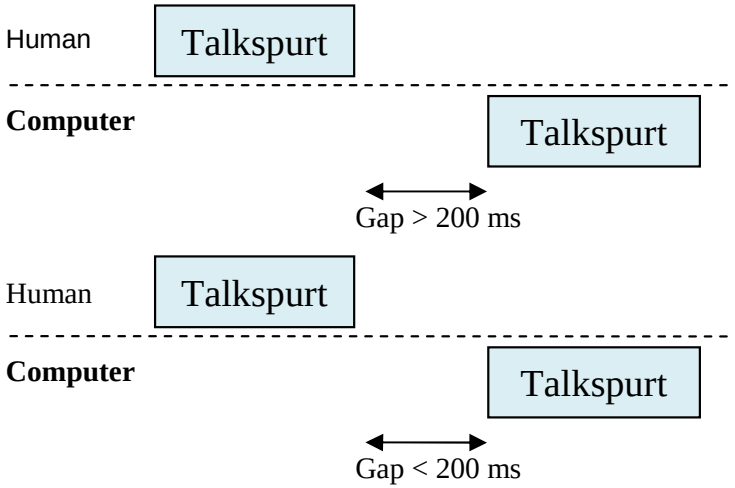
Gap

Scope and target

Two cases of turn-taking:

Gap > 200 ms:
handled by: End of utterance predictors

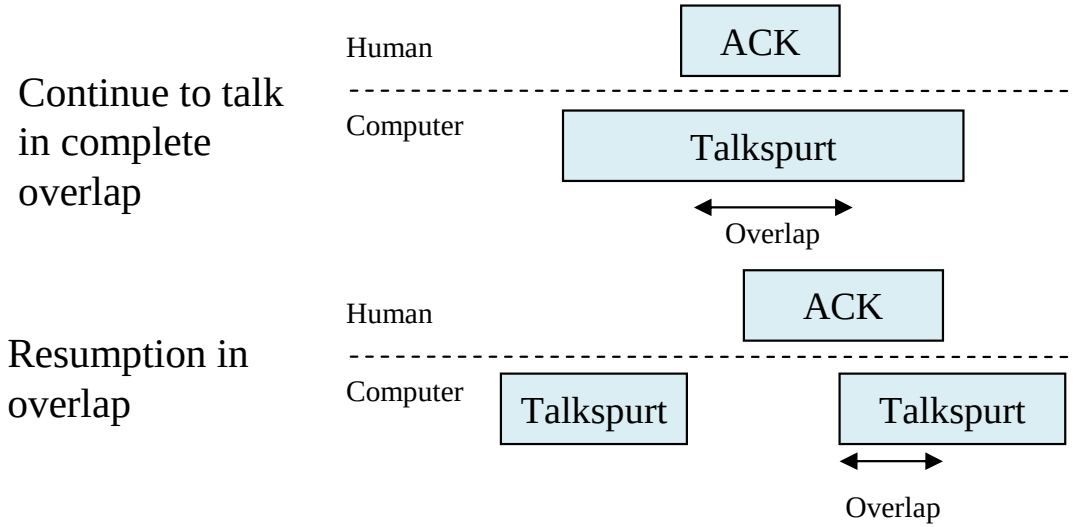
Gap < 200 ms:
Our target



A simplified approach

- ACK: acknowledgment move (backchannel type dialog act)
- These are non-intrusive, they don't compete with the floor

We want a dialog system to be able to do this:



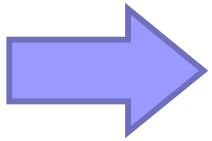
The HCRC Map task

- Scottish English
- 64 face-to-face dialogs
- One subject explains a route to the other
- HCRC Map Task Acknowledgment Move "**ACK**":
  - *"a verbal response that minimally shows that the speaker has heard the move to which it responds, and often also demonstrates that the move was understood and accepted"*
- The provided segmentation excluding extra-linguistic sounds is converted into talkspurts (Brady 1968)
  - Minimum voice activity duration threshold of 50 ms
  - Minimum inter-pause duration threshold of 200 ms (minimally perceivable pause)

Acks: Tokens

word	%	word	%	word	%	word	%
right	28.2	*oh*	2.7	*got*	0.9	*a*	0.7
okay	14.9	*the*	2.3	*it*	0.9	*to*	0.7
mmhmm	5.3	*that's*	1.6	*you*	0.9	*fine*	0.6
uh-huh	5.3	*no*	1.5	*that*	0.8	*I've*	0.6
yeah	3.9	*I*	1.4	*mm*	0.7	other	26.1

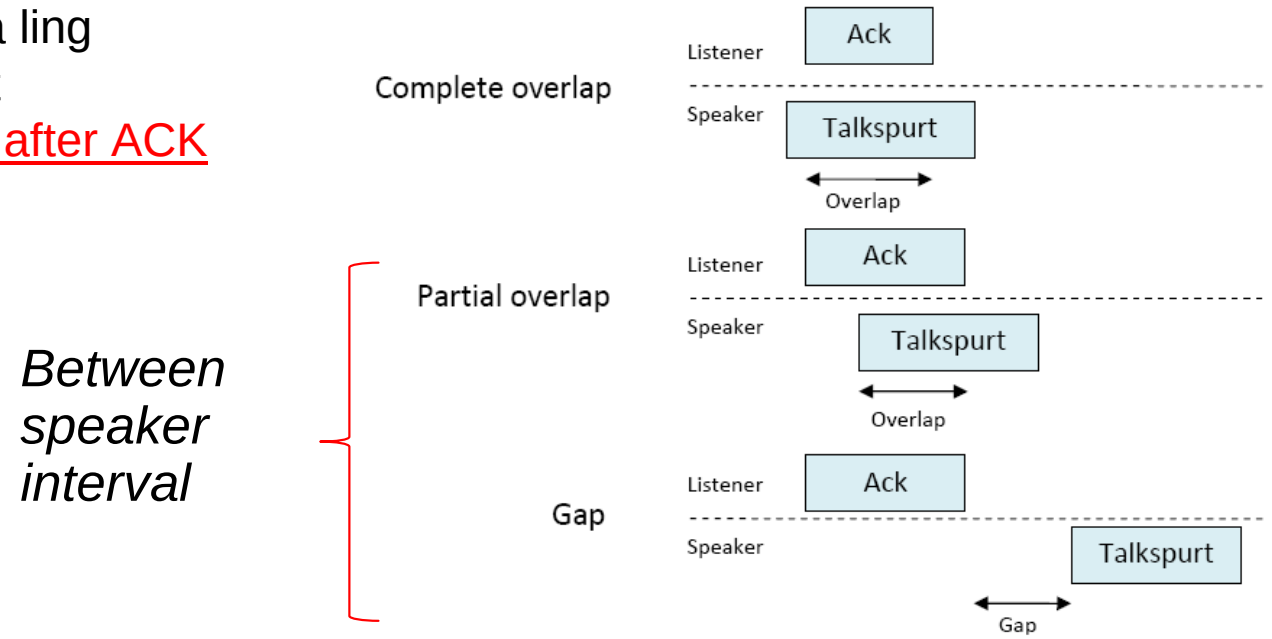
Table 1. Top 20 most frequently occurring words in all Acknowledgment Moves found in the Map Task corpus, as percentages of a total of 9823 words.



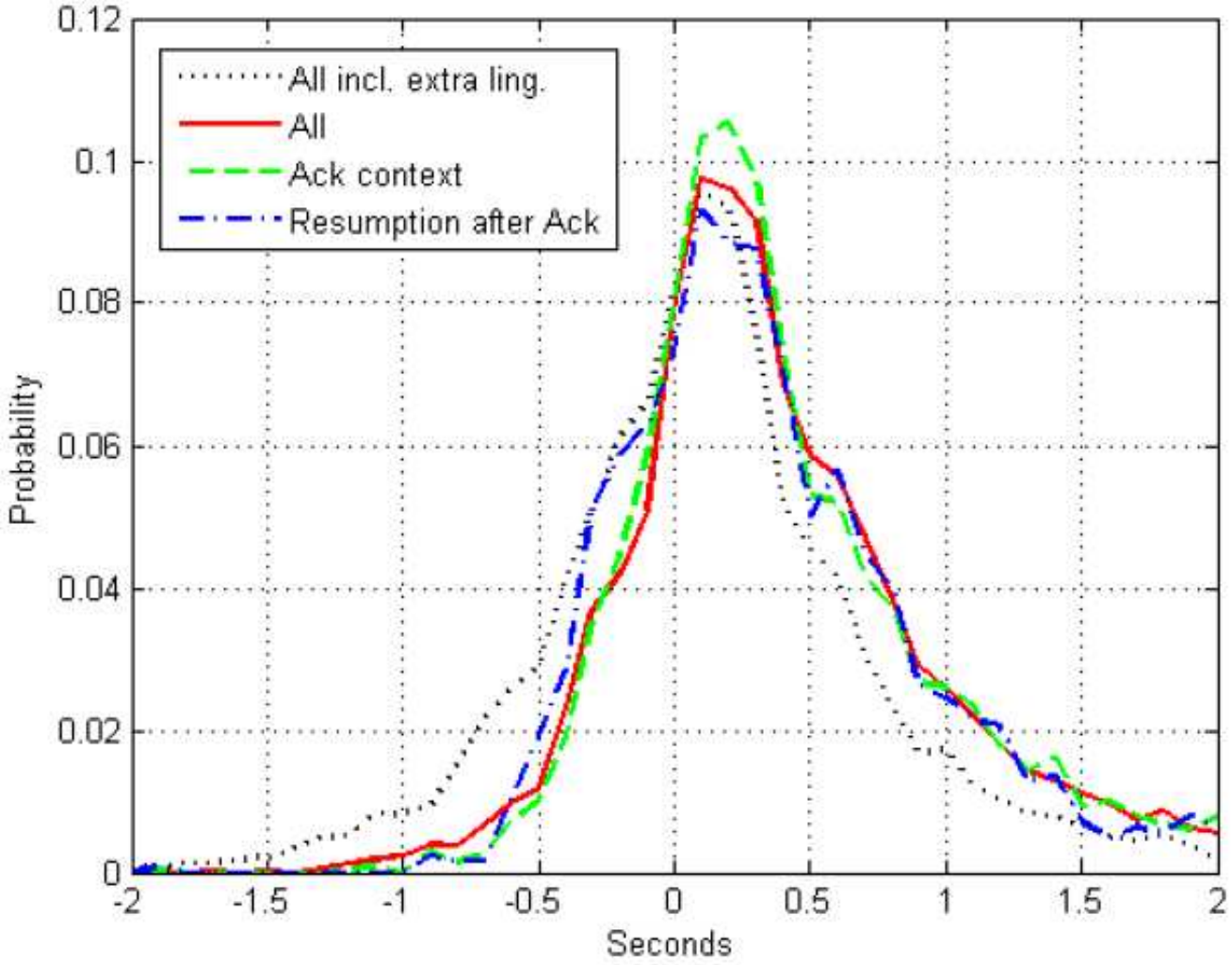
ACKs may be defined by their lexical content

How does ACKs occur in overlap?

- Durations measured by 10 ms frames
  - Given a frame in non-overlap, it is 5 % probability it is an ACK
  - Given a frame in overlap, it is 35 % probability it is an ACK
- ACK are more common in overlap
- Let's compute the between speaker interval for the talkspurts of type:
  - All incl. extra ling
  - All excl extra ling
  - ACK context
  - Resumption after ACK



Between speaker intervals



From: Reidsma, D., de Kok, I., Neiberg, D., Pammi, S., van Straalen, B., Truong, K., & van Welbergen, H. (in press). Continuous Interaction with a Virtual Human. *In Journal on Multimodal User Interfaces*.

Implications of overlap measurements

- The overrepresentation of ACK in overlap seems mostly be due to interjection into complete overlap
- For both interjection into complete overlap and resumption after ACK interjection into silence we need to classify incoming speech as ACK or not quite early

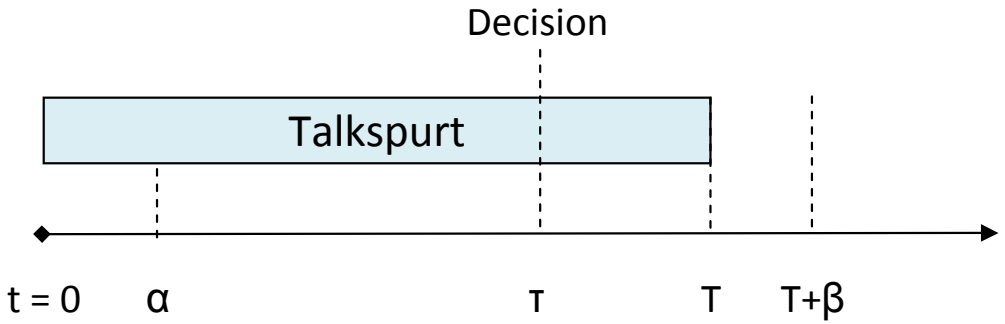
Maximum Latency Classification

How to guarantee a decision within a duration threshold:

Case 1

$\tau < T$

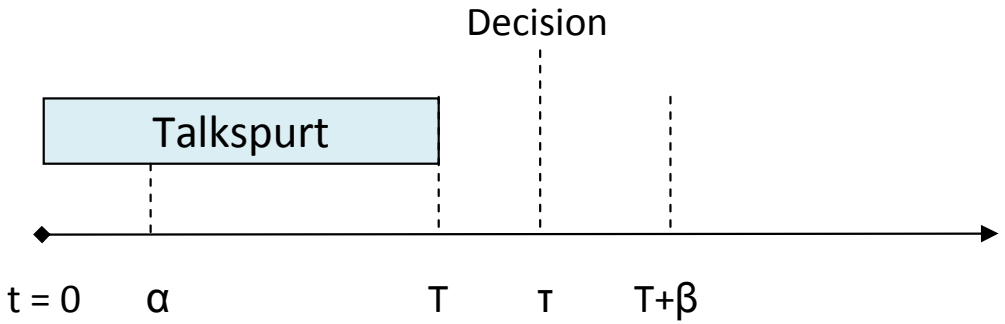
Latency: τ (max)



Case 2

$T < \tau < T+\beta$

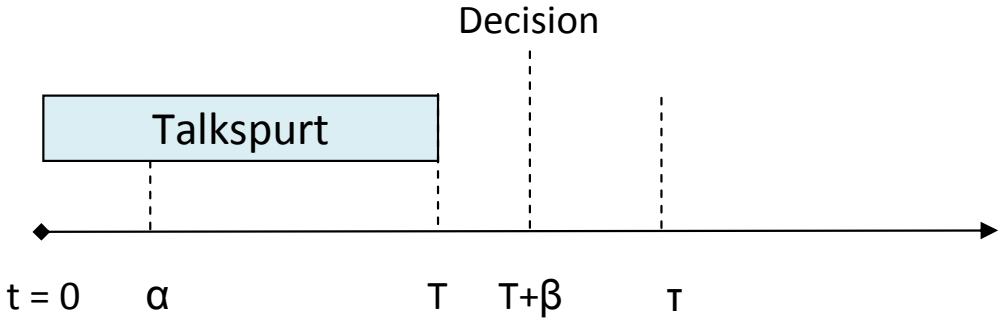
Latency: τ (max)



Case 3

$T+\beta < \tau$

Latency: $T+\beta$



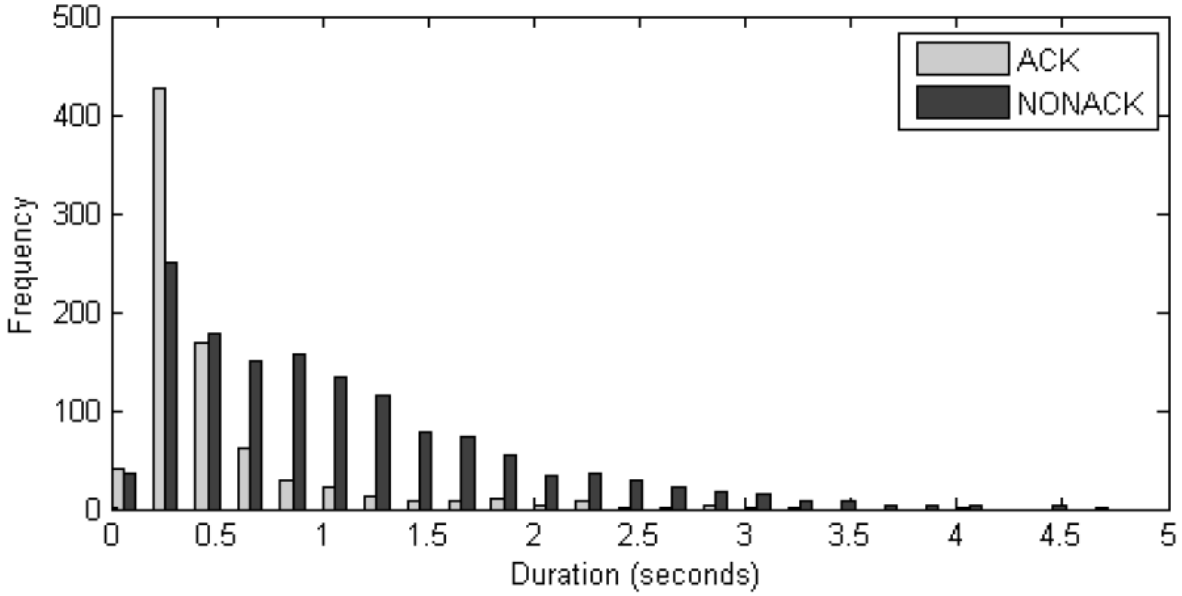
α= Minimum speech activity threshold

β = Minimum pause threshold

τ = maximum latency threshold

T = Talkspurt duration

Design implications



The longer maximum latency, the better is duration as a feature but the less overlap-talk opportunities

Let’s try 100 ms, 300 ms and 500 ms

Length-Invariant Parameterization

Maximum latency constraints gives variable segment lengths!

To parameterize the trajectories of each type of acoustic feature throughout a talkspurt, we use DCT coefficients invariant to segment length:

$$X_k = \frac{1}{N} \sum_{n=0}^{N-1} x_n \cos\left(\frac{\pi}{N}(n+\frac{1}{2})k\right) \qquad k = 0, \ldots, K \leq N-1$$

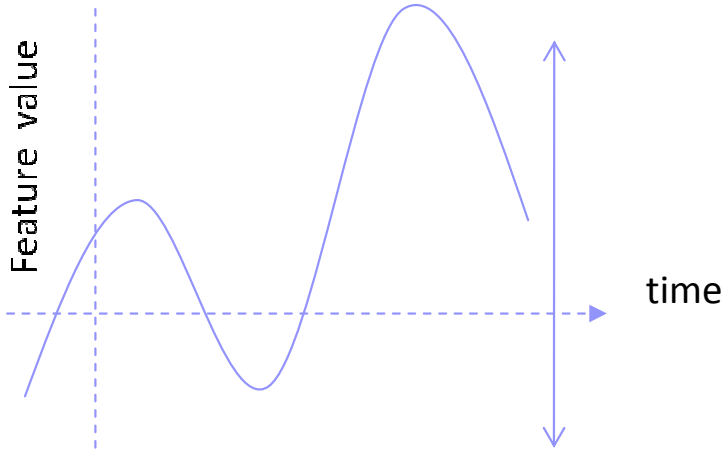
1. The basis functions are periodic which allows good interpolation of syllabic rhythm
2. The length-invariance gives a normalization for duration or speaking rate.
   - Speaking rate / duration can be separated in the machine learner/analysis
3. The 0'th coefficient is equal to the arithmetic average

Omitting the 0th is useful for

F0: Removes speaker dependent bias

Intensity: Removes bias caused by distance to the microphone

MFCC: Removes channel bias (applied in time dimension, forms a length-invariant cepstrum modulation spectrum space)

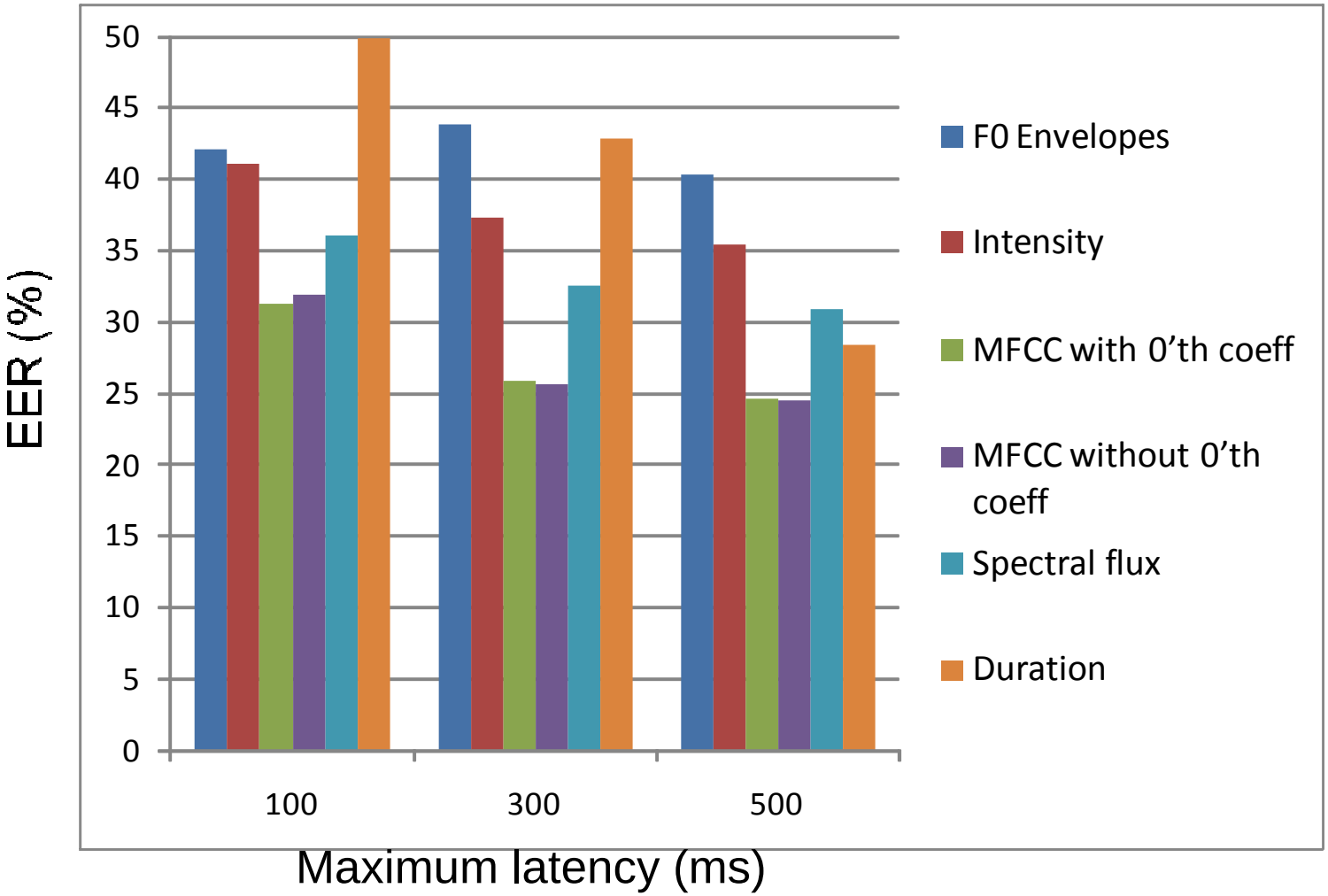


Classifier setup

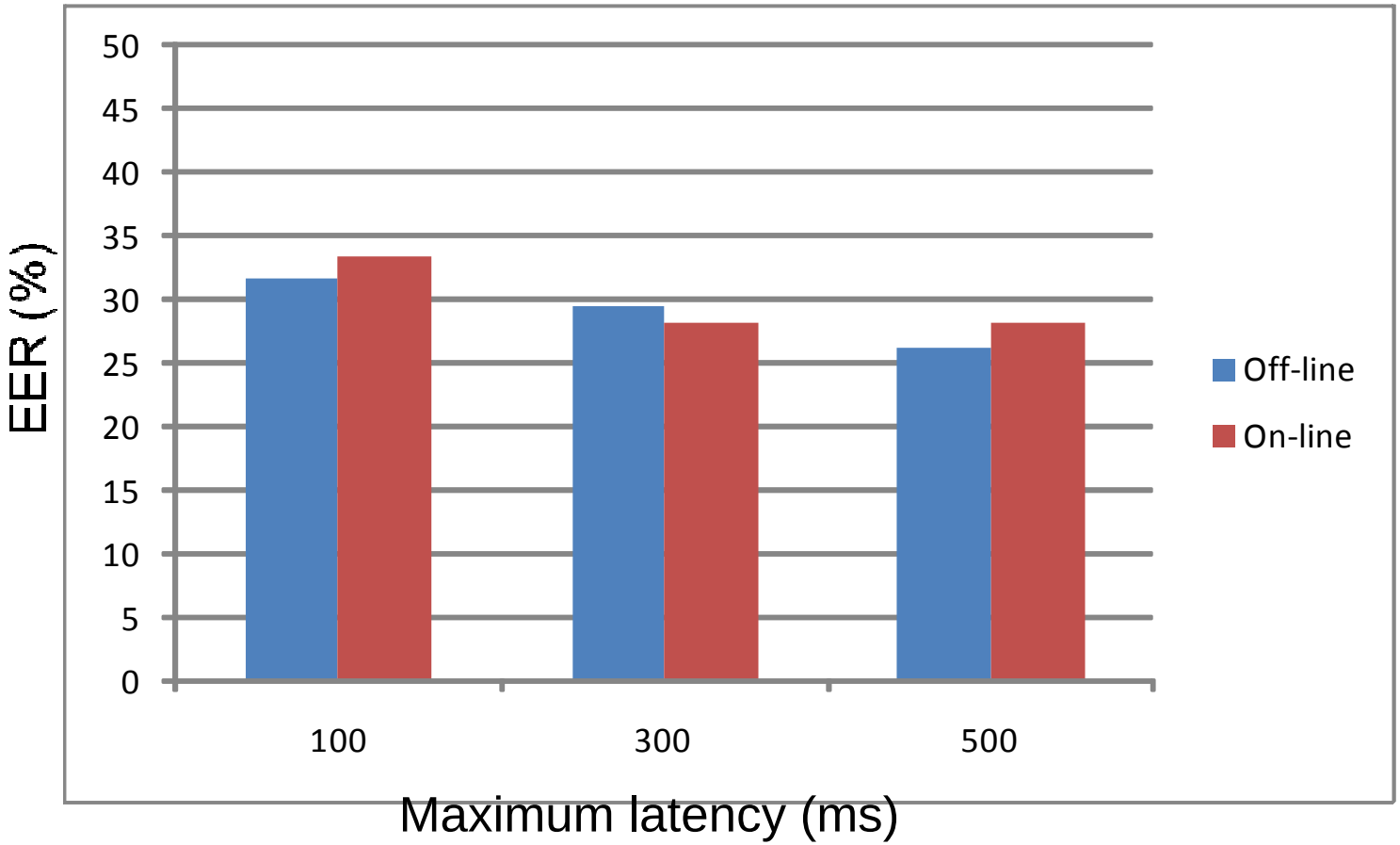
- **F0 Envelopes DCT 1-6** (rel. shape):
  - Back-channels has been shown to have a rise or drop in F0 (Benus et.al, 2007)(Gustavson et.al, 2010)
- **Intensity DCT 1-6** (rel. shape):
  - Back-channels has been shown to have distinct intensity contours (Benus et.al, 2007)(Gustavson et.al, 2010)
- **MFCC DCT 0/1-6** (abs or rel. shape):
  - Similar phonetic content may be captured by MFCCs.
- **Duration:**
  - Back-channels are short in duration (Edlund, 2008)
  - For training, the full talk-spurt duration was used
  - For testing, the duration up to the maximum latency threshold was used.
- **Spectral Flux DCT 0-5 (abs. shape):**
  - Common listener responses such as ``mmhmm'' and ``uh-huh'' are relatively homogeneous throughout their realization, and spectral flux should capture this property
- **Classifier type:** Support Vector Machines using Radial Base Kernel (libSVM)

Dev-set Results

- Train-set: 32 first dialogs, Dev-set: next 16 dialogs, Eval-set: last 16 dialogs



MTAck vs NonMTAck: Eval Results



Conclusions on Analysis and Classifiers

- Duration and MFCCs appear to be strong discriminative features;
  - ACKs may be defined by spectral change, or by lexical content
  - ACKs may be defined by duration
- To integrate this classifier within an incremental dialog processing framework which is able to handle multiple ongoing plans, we suggest to run three classifiers in parallel. This would let the dialog manager to
  - prepare decisions at 100 ms, and then
  - execute decisions at 300 ms or 500 ms.
- The actual online implementation is done in OpenSmile

That’s all!

Our other usages of time varying length-invariant DCT at KTH:

- Gustafson, J., & Neiberg, D. (2010). Prosodic cues to engagement in non-lexical response tokens in Swedish. In DiSS-LPSS.
- Neiberg, D., Laukka, P., & Ananthakrishnan, G. (2010). Classification of Affective Speech using Normalized Time-Frequency Cepstra. In Prosody 2010.
- Neiberg, D., & Gustafson, J. (2010). The Prosody of Swedish Conversational Grunts. In Interspeech, Special Session on Social Signals in Speech.
- Ananthakrishnan, G., & Engwall, O. (2011). Mapping between Acoustic and Articulatory Gestures. Speech Communication, 53(4), 567-589.
- Picard, S., Ananthakrishnan, G., Wik, P., Engwall, O., & Abdou, S. (2010). Detection of Specific Mispronunciations using Audiovisual Features. To be published in International Conference on Auditory-Visual Speech Processing. Kanagawa, Japan.
- Neiberg, D. (2011). Prosodic Densities and Contours: Forming One from Many. To be published in Fonetik 2011
- Reidsma, D., de Kok, I., Neiberg, D., Pammi, S., van Straalen, B., Truong, K., & van Welbergen, H. (in press). Continuous Interaction with a Virtual Human. Journal on Multimodal User Interfaces.