NPFL099 - Statistical dialogue systems

Dialogue system evaluation

Filip Jurčíček

Institute of Formal and Applied Linguistics
Charles University in Prague
Czech Republic

Home page: http://ufal.mff.cuni.cz/~jurcicek

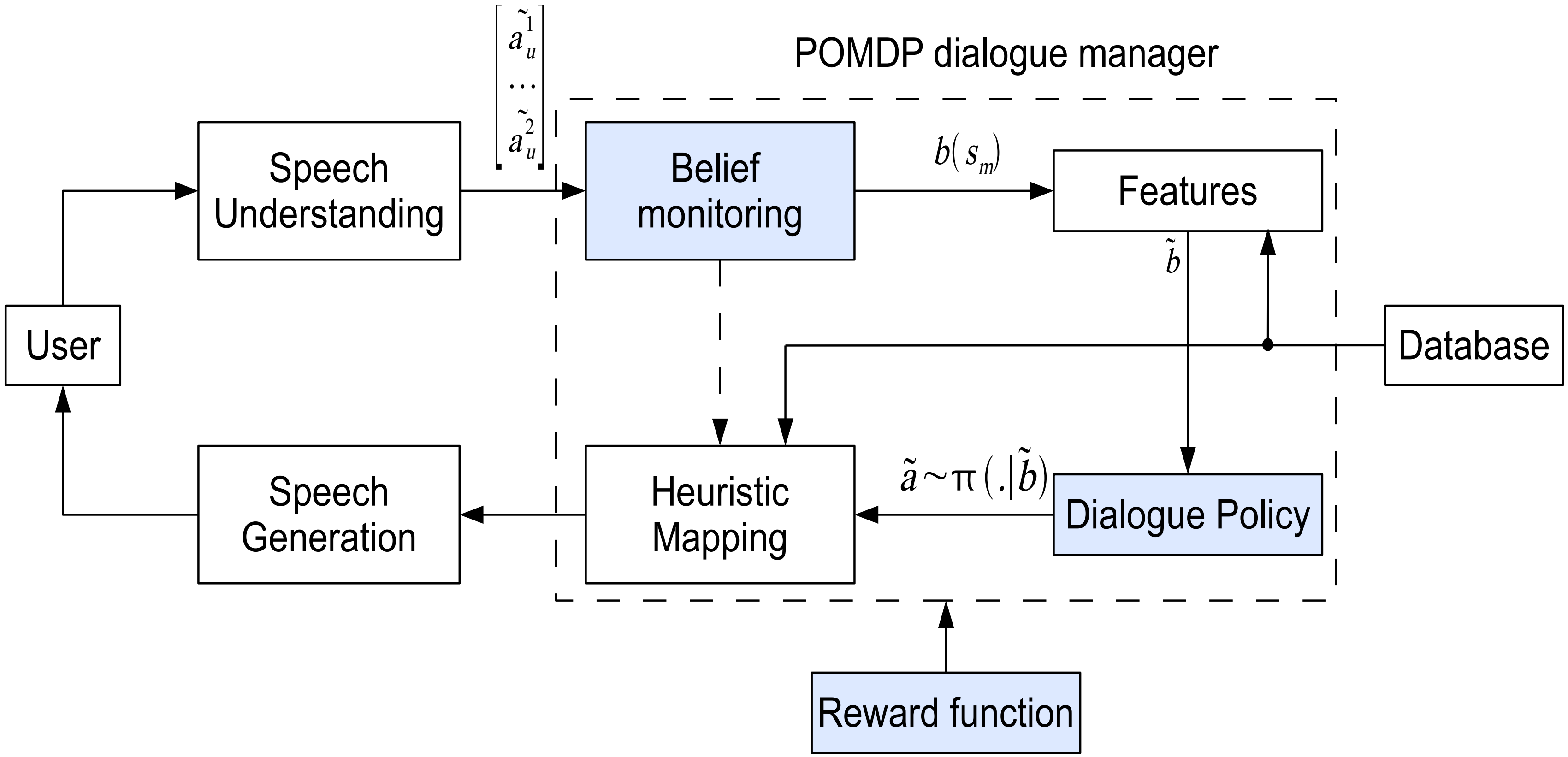
Version: 14/05/2013

Outline

- Evaluation of dialogue systems
- Laboratory evaluation
- Crowdsouricing
- Real user evaluation

Evaluating of SDS

- evaluate each component separately
- or in the context of all others



Metrics - subjective

- Typically a feedback form
  - Did you find all the information you were looking for?
    - to evaluate the dialogue manager
  - The system understood me well
    - to evaluate the spoken language understanding component
  - The phrasing of the system's responses was good
    - to evaluate the language generation component
  - The system's voice was of good quality
    - to evaluate the speech synthesizer

Metrics - subjective

- You do not want to ask many questions
  - It is boring
  - The answers typically correlate
- Ideally yes / no questions
- or selection from N options
  - Likert scale
  - ``strongly disagree'', ``disagree'', ``lightly disagree'', ``slightly agree'', ``agree'', and ``strongly agree''.

Metrics - objective

- So far, we talked about subjective metrics

- Some times
  - subjective metrics are not available
  - objective (automatic) metrics can be better ???

- Objective metrics
  - PERSEVAL – Walker et. al
    - trainable model from a corpus of human ratings
    - most explanatory features are accuracy of ASR, the length of the dialogue
  - BETTER: was the call routed to a human operator?
    - not always applicable

Evaluation in a laboratory

- In controlled environment
  - noiseless or with generated back ground noise

- Typically
  - each user gets training
  - is supervised by an assistant
  - users rating is controlled

- When interacting with the SDS
  - user is given a goal
  - the assistant could point out
    - errors in rating
    - missing constraints in the goal

Evaluation in a laboratory

- It is time consuming
- Search for subjects among colleagues or students
- You have to make appointments
  - some people does not show up
- Expensive – in CAM, we paid £15 for an hour
- Still, we could not get enough subjects

Evaluation using crowdsourcing

- Similar setup but hiring users differently
- Amazon Mechanical Turk users
  - toll-free phone number in USA
  - mostly native English speakers from USA
  - some Canadians
  - many Indians
  - some non native speakers of English from USA

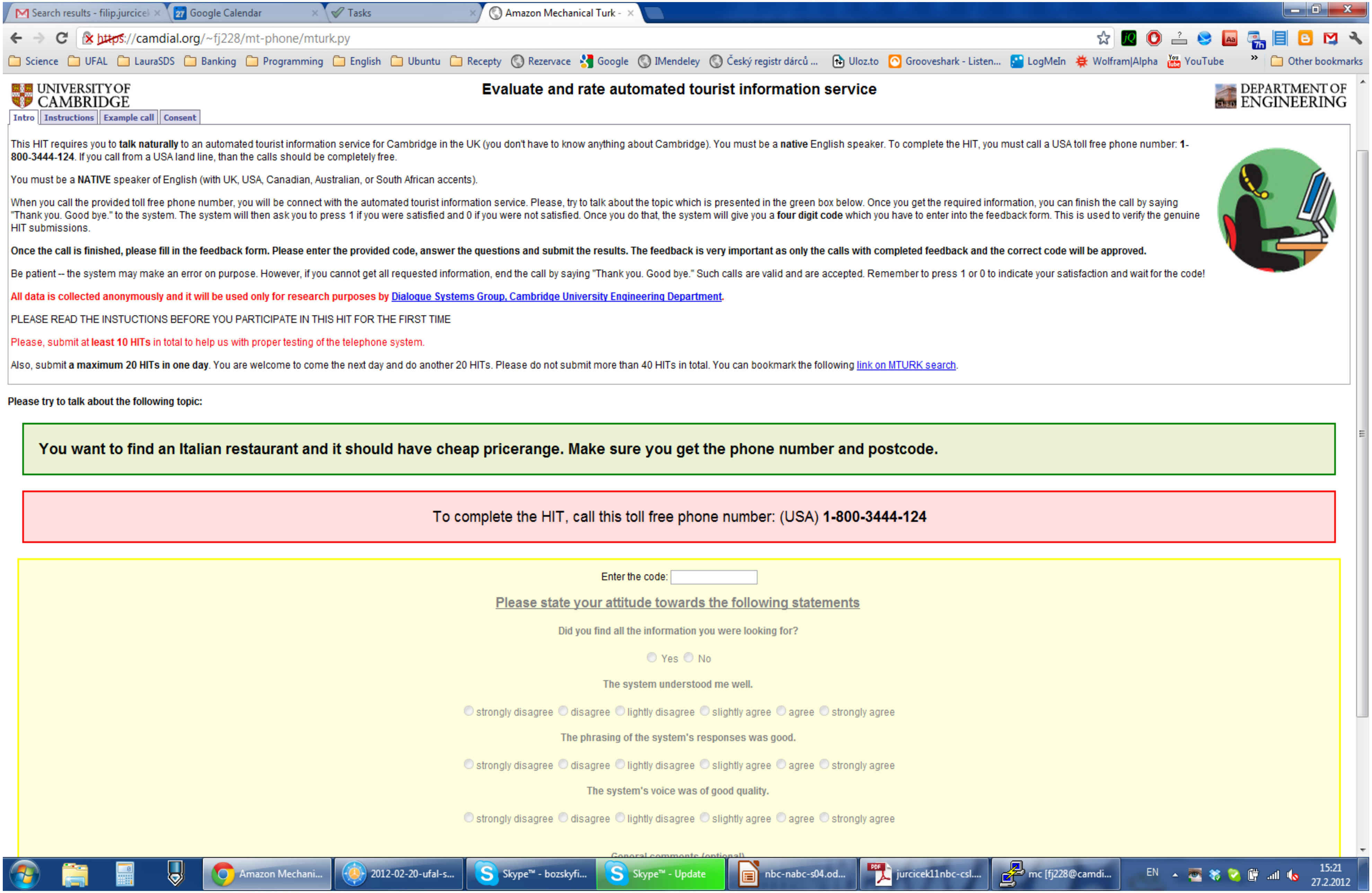
Evaluation using crowdsourcing

- Instead of coming to a lab, subjects were presented with a web page

- Web interface
  - To instruct users
  - To give tasks
  - To collect feedback

- Phones used to deliver voice
  - Calls routed using SIP to Cambridge, UK

Web interface



Evaluate and rate automated tourist information service

This HIT requires you to **talk naturally** to an automated tourist information service for Cambridge in the UK (you don't have to know anything about Cambridge). You must be a **native** English speaker. To complete the HIT, you must call a USA toll free phone number: **1-800-3444-124**. If you call from a USA land line, than the calls should be completely free.

You must be a **NATIVE** speaker of English (with UK, USA, Canadian, Australian, or South African accents).

When you call the provided toll free phone number, you will be connect with the automated tourist information service. Please, try to talk about the topic which is presented in the green box below. Once you get the required information, you can finish the call by saying "Thank you. Good bye." to the system. The system will then ask you to press 1 if you were satisfied and 0 if you were not satisfied. Once you do that, the system will give you a **four digit code** which you have to enter into the feedback form. This is used to verify the genuine HIT submissions.

Once the call is finished, please fill in the feedback form. Please enter the provided code, answer the questions and submit the results. The feedback is very important as only the calls with completed feedback and the correct code will be approved.

Be patient -- the system may make an error on purpose. However, if you cannot get all requested information, end the call by saying "Thank you. Good bye." Such calls are valid and are accepted. Remember to press 1 or 0 to indicate your satisfaction and wait for the code!

All data is collected anonymously and it will be used only for research purposes by Dialogue Systems Group, Cambridge University Engineering Department.

PLEASE READ THE INSTUCTIONS BEFORE YOU PARTICIPATE IN THIS HIT FOR THE FIRST TIME

Please, submit at **least 10 HITs** in total to help us with proper testing of the telephone system.

Also, submit **a maximum 20 HITs in one day**. You are welcome to come the next day and do another 20 HITs. Please do not submit more than 40 HITs in total. You can bookmark the following link on MTURK search.

Please try to talk about the following topic:

You want to find an Italian restaurant and it should have cheap pricerange. Make sure you get the phone number and postcode.

To complete the HIT, call this toll free phone number: (USA) **1-800-3444-124**

Enter the code: ____

Please state your attitude towards the following statements

Did you find all the information you were looking for?

○ Yes ○ No

The system understood me well.

○ strongly disagree ○ disagree ○ lightly disagree ○ slightly agree ○ agree ○ strongly agree

The phrasing of the system's responses was good.

○ strongly disagree ○ disagree ○ lightly disagree ○ slightly agree ○ agree ○ strongly agree

The system's voice was of good quality.

○ strongly disagree ○ disagree ○ lightly disagree ○ slightly agree ○ agree ○ strongly agree

General comments (optional)

Evaluation using crowdsourcing

- Relatively easy to get users
  - between 100 – 200 calls a day
  - better to ask TURKs to test a system than to ask colleagues ;-)
- Cheap – minimum wage
  - we paid $6 for an hour
- Toll-free phone number cheap
  - $0.02 per minute

Evaluation with real user

- TURKS are not real user
  - they are still paid
  - their rating is in some extent random
    - though this is true for all humans
    - unless you go to the recommended restaurant, it is hard to rate usefulness of the SDS recommendation
- Would be better to have real users interested in using the SDS
  - only some have such applications
  - e.g. Speech Cycle, Nuance, France Telecom
    - have tens of thousand calls a month

Evaluation with real user

- Still, the rating does not have to be consistent

- The reward can be delayed

- I will know that the appointment booking was successful only when the technician comes on the date I wanted

  - FT: appointment booking application

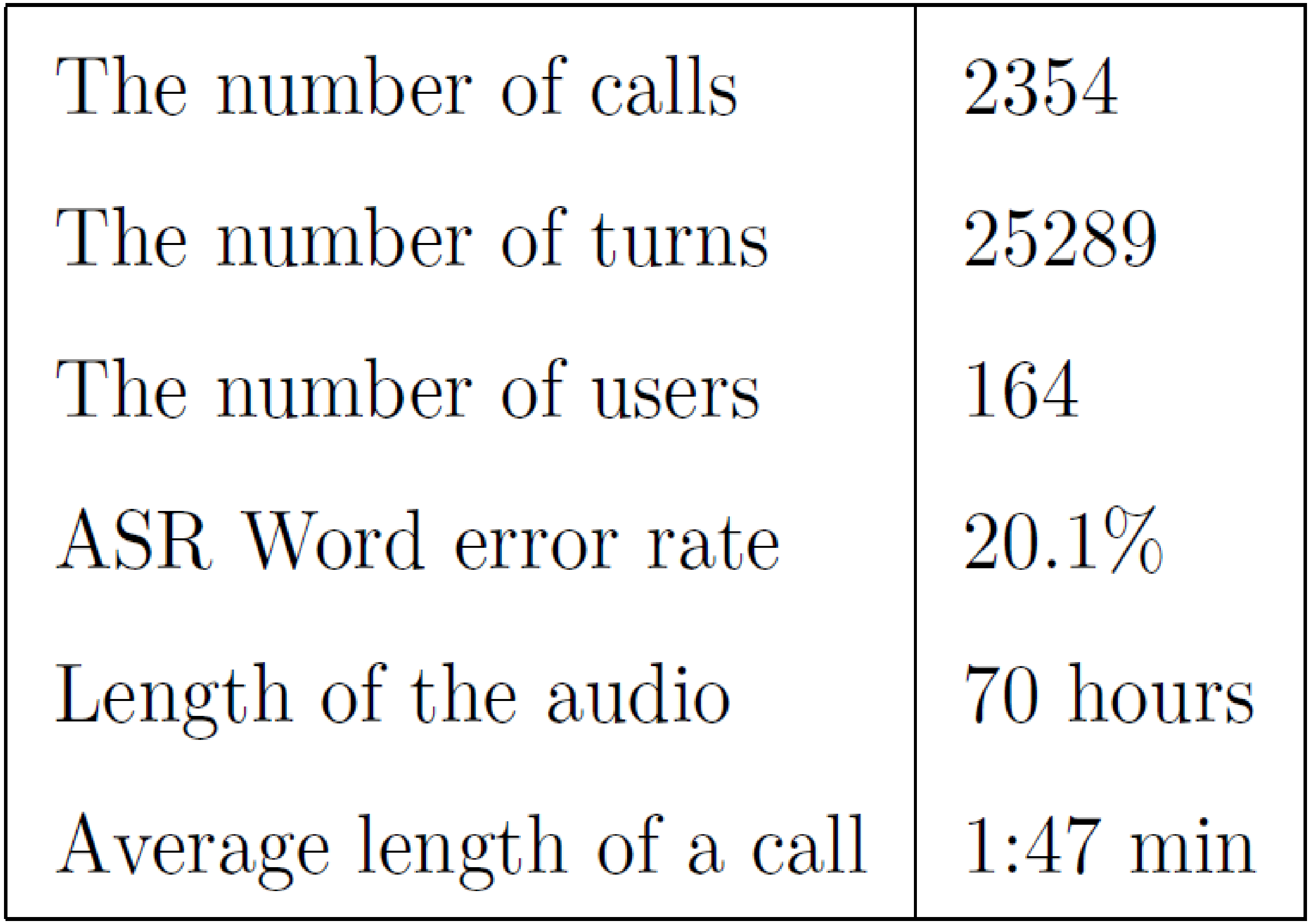
- You do not want ask all users
  - therefore automatic metrics are preferred

Metrics – how many user do we need?

- Many!!!
- Imagine testing a system
  - the success rate is about 50 – 60 percent
  - when you collect 500 dialogues then the 95% confidence interval +- 5%
- Using parametric tests, a difference of less than 5% is not statistically significant

Example: MTURK trial

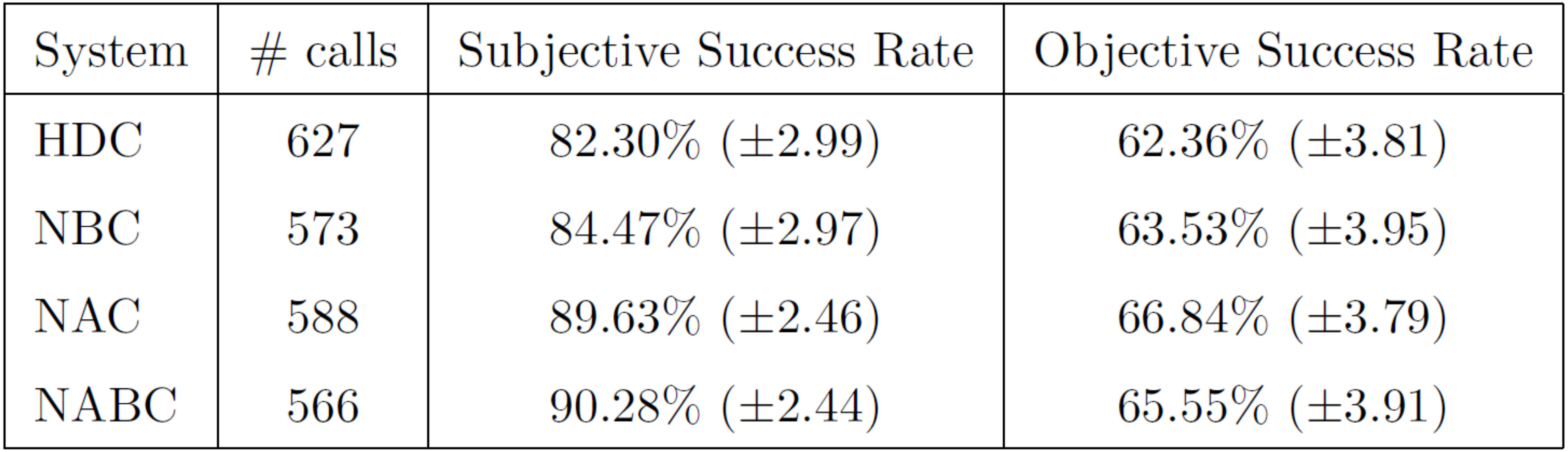
- Amazon Mechanical Turk users



The number of calls	2354
The number of turns	25289
The number of users	164
ASR Word error rate	20.1%
Length of the audio	70 hours
Average length of a call	1:47 min

Results: MTURK trial

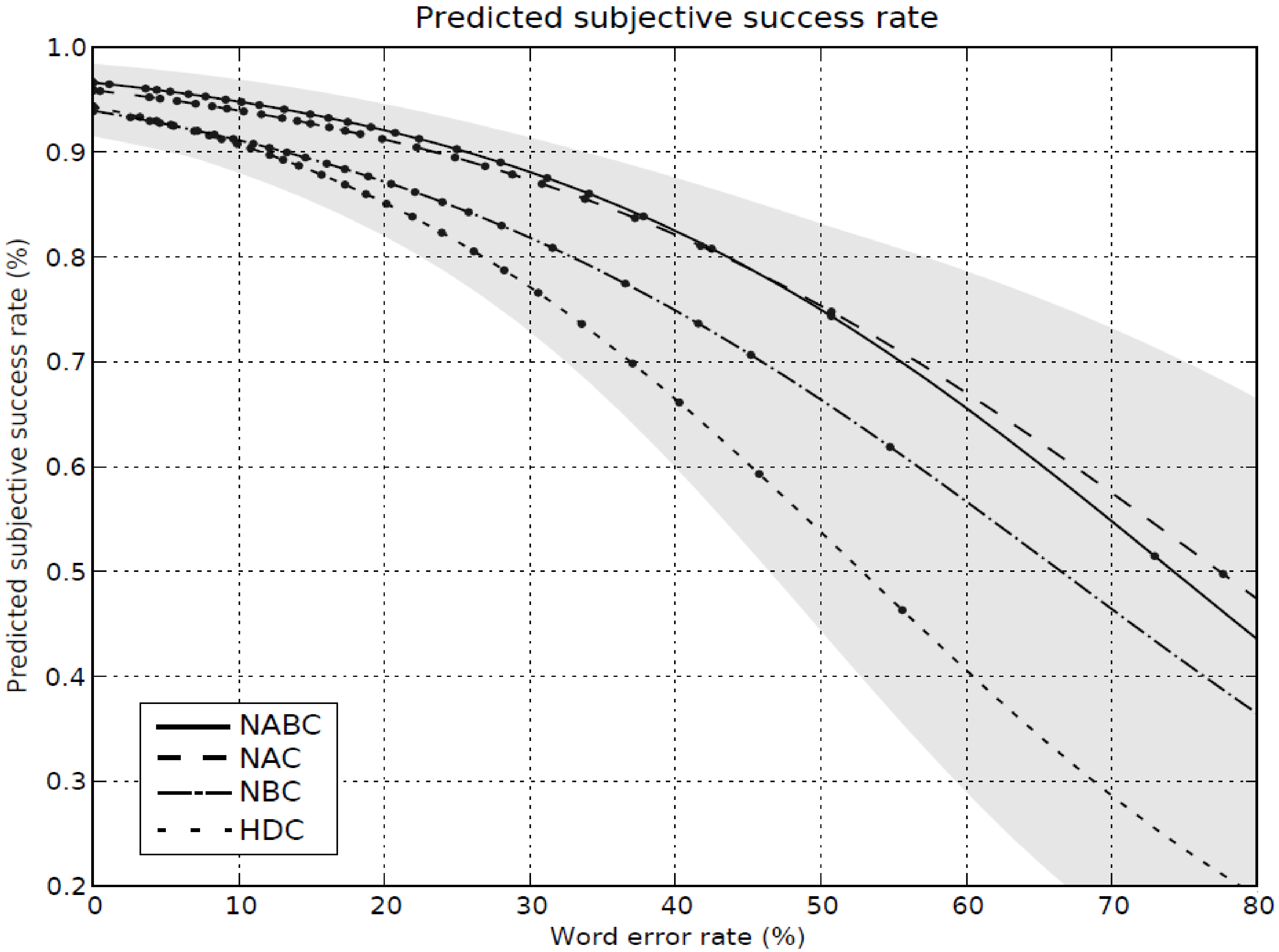
- Metrics
  - Subjective success rate – user ratings
  - Objective success rate – automatically derived



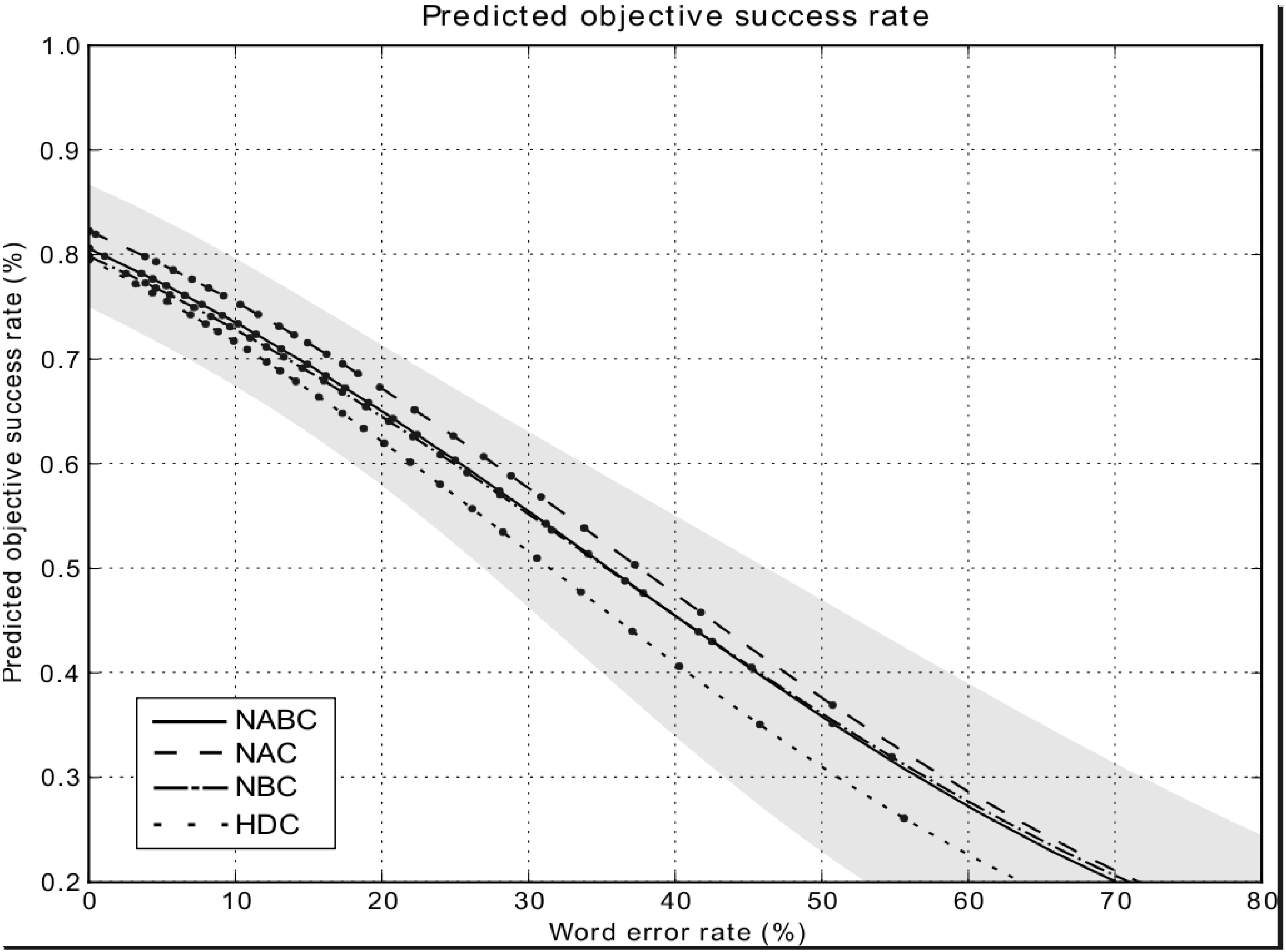
System	# calls	Subjective Success Rate	Objective Success Rate
HDC	627	82.30% (±2.99)	62.36% (±3.81)
NBC	573	84.47% (±2.97)	63.53% (±3.95)
NAC	588	89.63% (±2.46)	66.84% (±3.79)
NABC	566	90.28% (±2.44)	65.55% (±3.91)

- This does not say much about the performance at different error rates

Results: subjective scores



Results: objective scores



Thank you!

Filip Jurčíček

Institute of Formal and Applied Linguistics
Charles University in Prague
Czech Republic

Home page: http://ufal.mff.cuni.cz/~jurcicek