Penalty Functions for Evaluation Measures of Unsegmented Speech Retrieval

Petra Galuščáková, Pavel Pecina, Jan Hajič

Institute of Formal and Applied Linguistics
Charles University in Prague
{galuscakova,pecina,hajic}@ufal.mff.cuni.cz

Motivation

- Speech Retrieval
  - Retrieving information from a collection of audio data in response to a given query
    – modality of the query could be arbitrary, either text or speech
  - Usually solved as text retrieval on transcriptions of the audio obtained by ASR
  - But: speech transcriptions are not 100% accurate, vocabulary is different, speech contains additional elements speech is usually not segmented into topically coherent passages

→ special evaluation methods for speech retrieval are needed

Evaluation of speech retrieval I

- Known Segments Boundaries
  - Speech collection is segmented to passages which can play the role of documents
  - **Precision/Recall**
  - **Average Precision**
    - arithmetic mean of the values of precision for the set of first most relevant retrieved documents
  - **Mean Average Precision**
    - arithmetic mean of the AP values for the set of the queries

Evaluation of speech retrieval II

- Unknown Boudaries
  - No topical segmentation, the system is expected to retrieve exact starting points for each query
  - **Mean Average Segment Precision**
    - recently introduced, used in MediaEval
    - designed for evaluation of retrieval of relevant document parts
  - **Mean Generalized Average Precision**
    - designed to allow certain tolerance in matching search results against a gold standard relevance assessment
    - tolerance is determined by the Penalty Function

Evaluation of speech retrieval - mGAP score

$$GAP = \frac{\sum_{R_k \neq 0} p_k}{N}$$

N = number of assessed starting points

R_k= reward calculated according to the Penalty Function

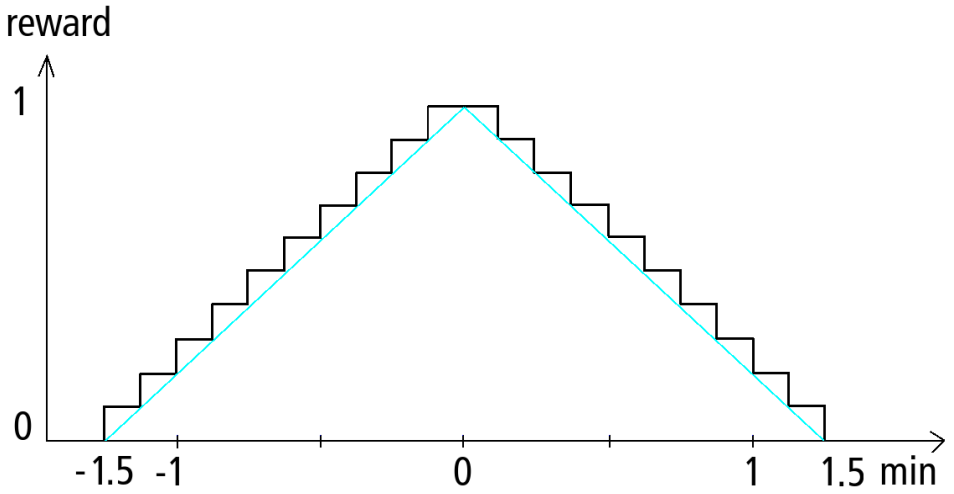
p_k is the value of Precision for the position k calculated as:

$$p_k = \frac{\sum_{i=1}^{k} R_i}{k}$$

mGAP = arithmetic mean of the GAP values for the set of the queries

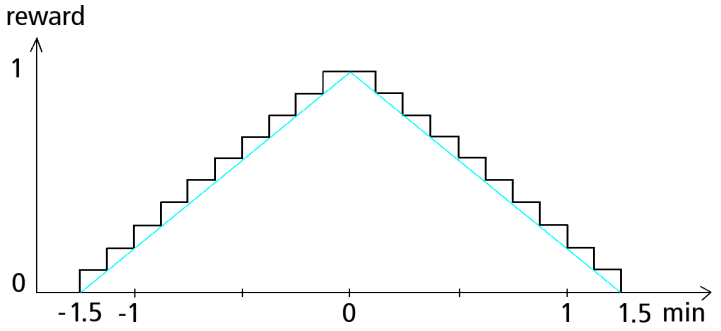
Evaluation of speech retrieval - mGAP score

- Time difference between the starting point of the topic determined by the system and the true starting point of this topic obtained during relevance assessment
- The actual shape of the function can be chosen arbitrarily
- The Penalty Function used in the mGAP measure in the Cross-Language Speech Retrieval Track of CLEF 2006 and 2007



Evaluation of speech retrieval - mGAP score

- Has been widely used, however, the measure (and the Penalty Function itself) have not been adequately studied
- Questions:
  - the Penalty Function is symmetrical and starting points retrieved by a system in the same distance before and after a true starting point are treated as equally good (or bad)
    - "**shape**" of the function itself
  - "**width**" of the Penalty Function, i.e. the maximum distance for which the reward is non-zero



Penalty Function Proposal

Methodology

- **Lab test** to study the behaviour of users
- IR system **simulation**
- Users were presented the topics from the test collection and playback points randomly generated in a vicinity of a starting point of a relevant segment
- Users should have navigated in the recording and indicate when the speaker started to talk about the given topic
- After they found the relevant segment, the participants were asked to indicate their satisfaction with the playback point

Number of participants	24
Number of processed starting points	263

MP3 Player 1.01

Nahrávka | Instrukce | O programu | Podmínky užití

Židovské zdravotní sestry v koncentračních táborech

Ošetřovatelství a péče o pacienty poskytované židovskými sestrami v koncentračních táborech v době holokaustu.

Systém zdravotní péče mohl být spíše neformální než formální. Zajímá nás toto téma jak obecně, tak konkrétně v případě tábora Ohrdruf.

Mariana Tylová

Místo narození: Prague (Czechoslovakia)

Rok narození: 1920

Aktuální čas v nahrávce: 45:38

Celková délka nahrávky: 81:18

> || >| - +

Nalezeno Nenalezeno

Odhlásit Ukončit

Data

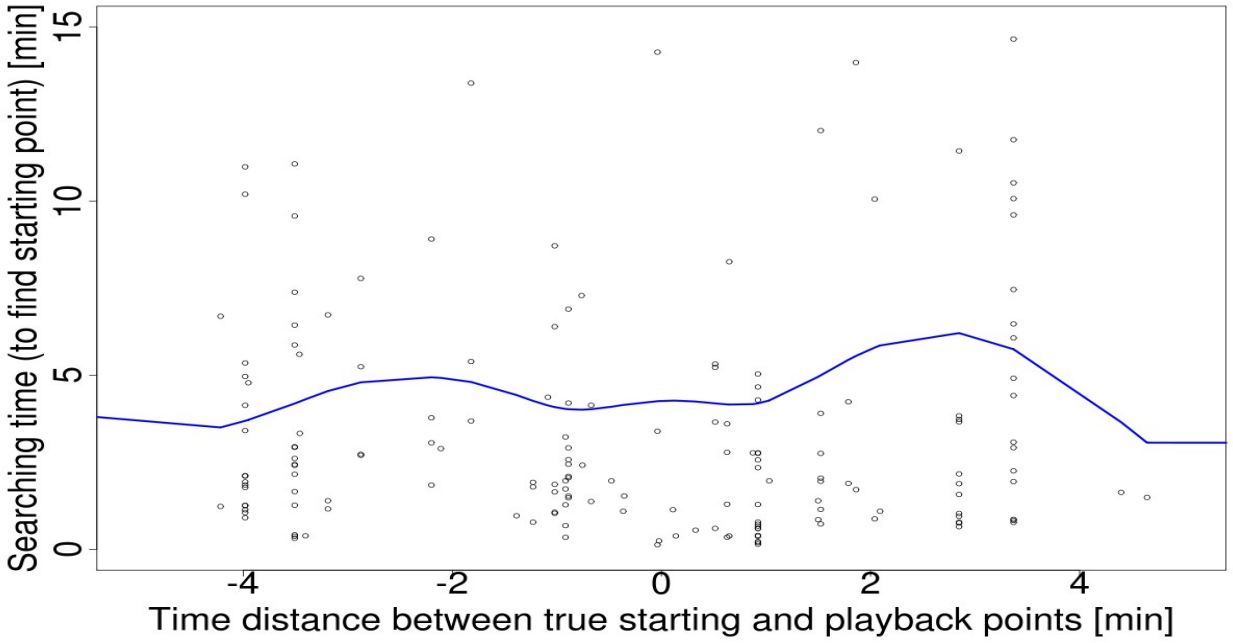
- Test collection used for Cross-Language Speech-Retrieval track of CLEF 2007
- **Manually processed** by human assessors – relevant passages for given topics were identified
- Part of oral history archive from the **Malach** Project (Holocaust testimonies)
- Recorded in Czech

Recordings in Malach Project	52 000
Czech recordigs in Malach Project	700
Assessed Czech recordings	357
Average length of the recording	95 min
Processed topics	116

Results

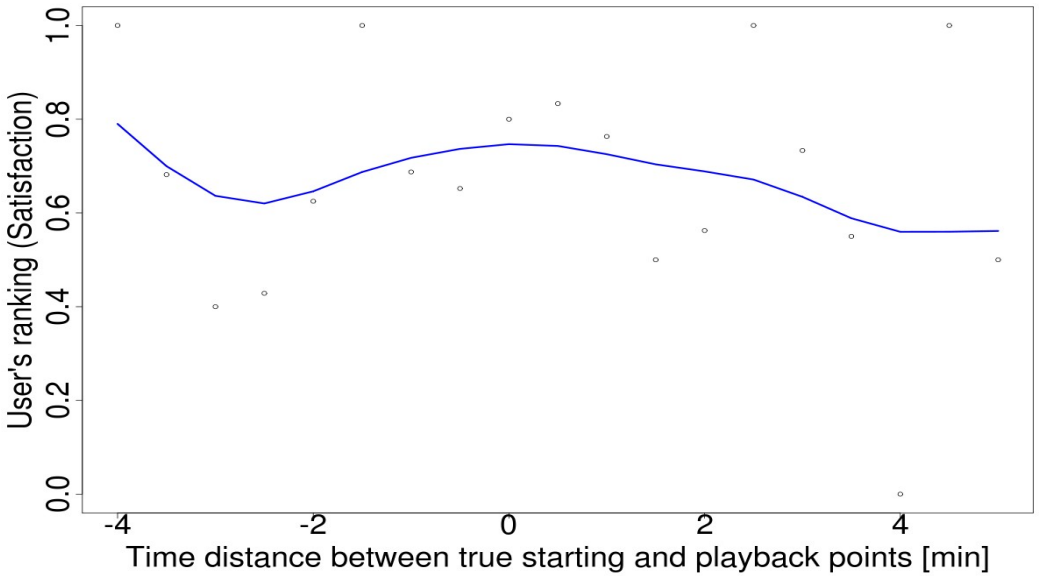
Time analysis

- We measure the elapsed time between the **beginning of playback** and the moment when the participant presses the button indicating that the **relevant passage was found**
- Respondents generally need **less time** to complete the task when the **playback** point is located **before** the true **starting** point



Users' satisfaction

- Participants were requested to indicate to what extent they were happy with the location of the playback points in the scale of: very good, good, bad or very bad
- Trend not clear - most satisfied when the **playback** reference point lies **shortly before** the **true starting** point but function value decreases more slowly for positive time



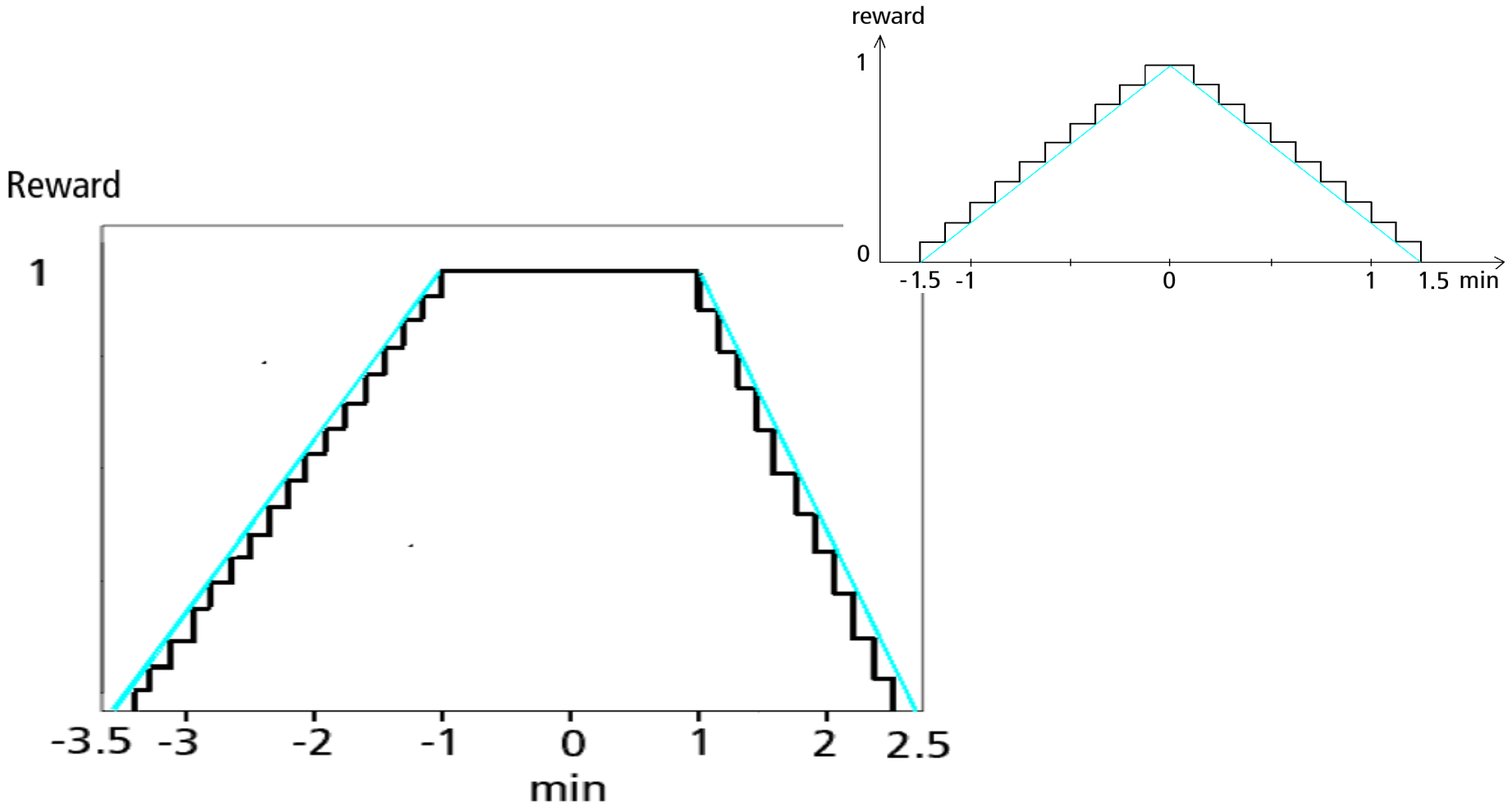
Proposed mGAP Modifications

1) Users prefer **playback points appearing before** the beginning of a true **relevant passages** to those appearing after, i.e. more reward should be given to playback points appearing before the true starting point of a relevant segment

2) Users are **tolerant to playback points** appearing **within a 1-minute** distance from the **true starting points**. i.e. equal (maximum) reward should be given to all playback points which are closer than one minute to the true starting point.

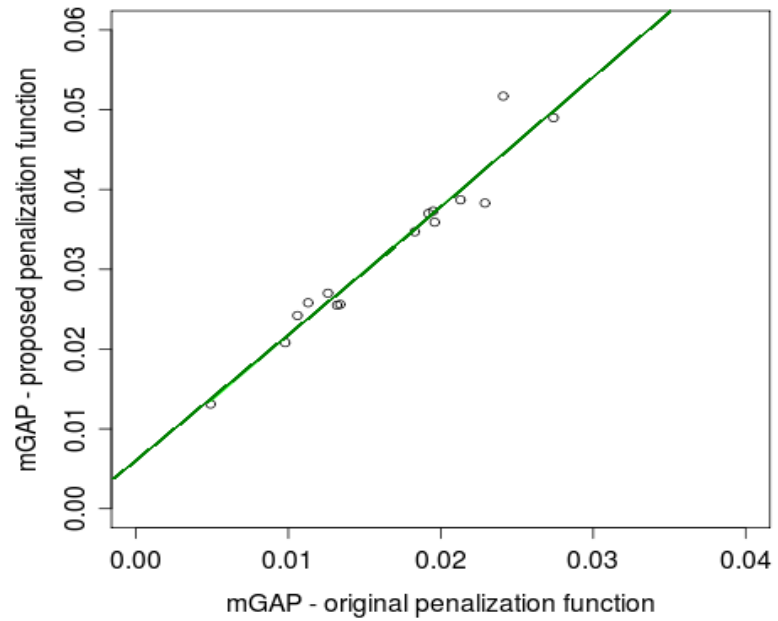
3) Users are still **satisfied** when **playback points** appear in **two- or three- minute** distance from the **true starting** point. i.e. function should be “wider”.

Proposed mGAP Modifications



Comparison with the Original Measure

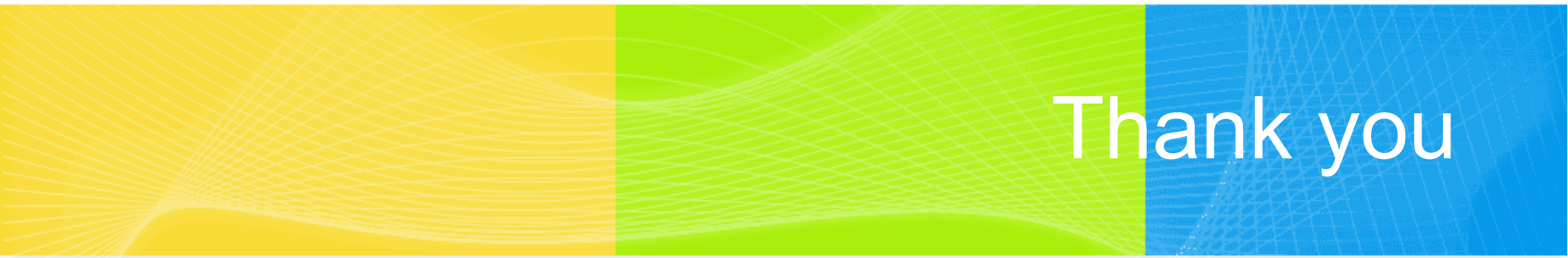
- Outputs of CLEF 2007 Cross-Language Speech Retrieval Track
- 15 retrieval system scored with the original and proposed Penalty functions
- High correlation



Conclusion

Conclusion

- We described evaluation of speech retrieval (segmented/not segmented)
- Described mGAP, penalty function drawbacks
- We organized human-based lab test
- Based on lab test results we modified Penalty Function
- Finally compared modified Penalty Function with the original function


Thank you