Penalty Functions for Evaluation Measures of Unsegmented Speech Retrieval

Petra Galuščáková, Pavel Pecina, and Jan Hajič

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

Abstract. This paper deals with evaluation of information retrieval from unsegmented speech. We focus on Mean Generalized Average Precision, the evaluation measure widely used for unsegmented speech retrieval. This measure is designed to allow certain tolerance in matching retrieval results (starting points of relevant segments) against a gold standard relevance assessment. It employs a Penalty Function which evaluates non-exact matches in the retrieval results based on their distance from the beginnings of their nearest true relevant segments. However, the choice of the Penalty Function is usually ad-hoc and does not necessary reflect users' perception of the speech retrieval quality. We perform a lab test to study satisfaction of users of a speech retrieval system to empirically estimate the optimal shape of the Penalty Function.

1 Introduction and Motivation

The quantity of speech data has been increasing rapidly in the last decades. Successful and efficient search in speech data requires the use of high-quality information retrieval (IR) systems which, in turn, are impossible to construct without reliable evaluation of the quality of these systems. IR from speech data (speech retrieval) differs substantially from IR from text documents (document retrieval) and thus special-purpose evaluation techniques are required.

Speech retrieval is defined as retrieving information from a collection of audio data (recordings) in response to a given query – modality of the query could be arbitrary, either text or speech. This task is usually being solved as text retrieval on transcriptions of the audio obtained by automatic speech recognition (ASR). IR systems reported being used for such speech retrieval are e.g. Lemur [11], SMART [10], Terrier [10] and InQuery [9].

Speech retrieval systems based on ASR must deal with a number of issues unknown to the traditional text retrieval: Automatic speech transcriptions are not 100% accurate and contain errors, i.e. misrecognized words. The vocabulary used in speech is usually different from the one used in written text (including colloquial and informal words [11], etc.). Speech contains additional elements such as word fragments, pause fillers, breath sounds, long pauses and it is usually not segmented into topically coherent passages, not even paragraphs or sentences.

T. Catarci et al. (Eds.): CLEF 2012, LNCS 7488, pp. 100–111, 2012.

Evaluation of speech retrieval requires special measures designed specifically for this purpose. In this work, we focus on speech retrieval from recordings not segmented to passages which could serve as documents in the traditional IR. The main objective of this work is to verify whether the methods currently used for evaluation of speech retrieval in unsegmented recordings are appropriate and possibly modify these methods to better correspond to users' expectations. We focus on Mean Generalized Average Precision (mGAP) [7], which is de-facto standard measure for evaluation of unsegmented speech retrieval. mGAP has been used for several years but to our best knowledge such verification has not been reported yet. This work is the first attempt to do so.

First, we review evaluation of speech retrieval in general, then we describe a lab test carried out in order to measure satisfaction of the users with simulated results of a speech retrieval system. Based on an analysis of the survey results we propose a modification of the mGAP measure (or more precisely, its Penalty Function). Evaluation is performed on the results of the Cross-Language Speech-Retrieval track at CLEF 2007 [11], which includes a test collection, evaluation measure, and document rankings from the participating retrieval systems.

2 Evaluation of Speech Retrieval

The standard IR evaluation methods can be theoretically applied to speech retrieval but only if the speech collection is segmented to passages which can play the role of documents. If such a segmentation is not available, they cannot be used directly and need to be modified.

2.1 Segmented Speech Retrieval

In segmented speech retrieval, the collection consists of topically coherent passages which can be judged to be relevant or non-relevant to a particular query (or topic) as a whole. In that case, standard evaluation metrics, such as Mean Average Precision, can be used in the same way as for text document retrieval. This method was for example used in Unknown Story Boundaries Condition Track of TREC-8 [3], in which unknown boundaries of segments were converted to the known ones.

Precision (P) is defined as the ratio of the number of relevant retrieved documents to all retrieved documents and *Recall* (R) is the ratio of the number of relevant retrieved documents to all relevant documents. If an IR system also returns a relevance score for each retrieved document, these can be sorted in a descending order according to this score in a ranked list (for a given topic). For such a ranked list, one can compute the Average Precision (AP) as an arithmetic mean of the values of precision for the set of first m most relevant retrieved documents. This score is calculated for each new retrieved relevant document (d_m) [8]. Let S_k be the set of the first k retrieved documents for a given query and:

$$AP(d_m) = \frac{1}{m} \cdot \sum_{k=1}^{m} precision(S_k). \tag{1}$$

Mean Average Precision (MAP) is then calculated as an arithmetic mean of the AP values for the set of the queries Q on the set of documents D, formally:

$$MAP(Q) = \frac{1}{|Q|} \cdot \sum_{j=1}^{|Q|} AP_{Q_j}(D). \tag{2}$$

If no relevant document was retrieved, then the MAP value is equal to zero.

2.2 Unsegmented Speech Retrieval

If the collection consists of recordings with no topical segmentation, the system is expected to retrieve exact starting (and eventually ending) points of each passage relevant to a given query (or topic). The main issue with evaluation of such retrieval results is that failing to match a starting point exactly cannot be interpreted as a complete failure, which is the case in document retrieval.

Only a few measures targeting unsegmented speech retrieval have been proposed. Liu and Oard in [7] proposed the Mean Generalized Average Precision (mGAP) measure, a modification of MAP for unsegmented speech retrieval. This measure was used for example for the evaluation of Cross-Language Speech Retrieval Track of CLEF [10] and Rich Speech Retrieval Task of MediaEval Benchmark [6]. Eskevich et al. [2] introduced two measures for search in informally structured speech data: Mean Average Segment Precision (MASP) and Mean Average Segment Distance-Weighted Precision (MASDWP). MASP is a modification of MAP, inspired by MAiP [5] designed for evaluation of retrieval of relevant document parts. This measure evaluates retrieval systems with respect to segmentation quality and ranking of the results. MASDWP measure, similarly to mGAP, takes into account the distance between the start of a relevant segment and the retrieved segment [2], but employs segment precision too.

Mean Generalized Average Precision was designed to allow certain tolerance in matching search results (starting points of relevant segments) against a gold standard relevance assessment. This tolerance is determined by the Penalty Function, a function of the 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. *Generalized Average Precision* is defined formally as:

$$GAP = \frac{\sum_{R_k \neq 0} p_k}{N}, \tag{3}$$

where N is the number of assessed starting points, R_k is a reward calculated according to the Penalty Function for the starting point retrieved on the position k and p_k is the value of Precision for the position k calculated as:

$$p_k = \frac{\sum_{i=1}^{k} R_i}{k}. \tag{4}$$

Each annotated point is used in the Penalty Function calculation only once.

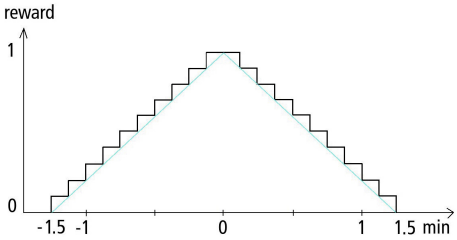

Fig. 1. mGAP Penalty Function used the CL-SR track at CLEF 2006 and 2007

mGAP is then defined analogically as in Equation (2) as an arithmetic mean of the values GAP for a set of queries Q and a set of documents D.

Values of Penalty Function are always non-negative and they decrease with increasing distance from the true starting points. For exact matches the Penalty Function returns 1 as a maximum reward. From a certain distance the function values are equal to zero. Apart from this, 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 [10] and 2007 [11] is shown in Figure 1. This function is not smooth, for each 9 seconds of time between the retrieved and true starting points the function decreases by 0.1. Thus, the interval for which the function gives non-zero scores is [-1.5,1.5] minutes.

The proposed mGAP measure has been widely used in recent evaluation campaigns [10,11] and research papers [4]. However, the measure (and the Penalty Function itself) have not been adequately studied as of yet. It is not clear to what extent mGAP scores correlate with human satisfaction of retrieval results.

For example, 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). We do not have enough empirical evidence whether this assumption is correct. Another point which needs to be verified is the "width" of the Penalty Function, i.e. the maximum distance for which the reward is non-zero, and the actual "shape" of the function itself.

The main purpose of the study is therefore to verify the appropriateness of the mGAP Penalty Function by examining the correlation of its scores and actual human behaviour and satisfaction in a simulated environment of a speech retrieval system.

3 Methodology

We have designed a lab test to study the behaviour of users when presented results of a speech retrieval system – i.e. a starting point of a segment which should be relevant to a particular topic. The users did not use a real speech retrieval system. Instead, they were presented a topic description and a starting point randomly generated in the vicinity of a starting point of a true relevant segment in an interface allowing basic playback functions. We measured a subjective satisfaction of the users with the retrieved starting point (whether it pointed to a

Table 1. Translation of a topic from the Malach speech-retrieval test collection

Id	1148
Title	Jewish resistance in Europe
Description	Provide testimonies or describe actions of Jewish resistance in Europe before and during the war.
Narrative	The relevant material should describe actions of only-or mostly Jewish resistance in Europe. Both individual and group-based actions are relevant. Type of actions may include survival (fleeing, hiding, saving children), testifying (alerting the outside world, writing, hiding testimonies), fighting (partisans, uprising, political security). Information about undifferentiated resistance groups is not relevant.

passage relevant to the given topic or not and/or how difficult it was to find one) and the time they spent doing this.

3.1 Test Collection

Data for the survey (including recordings, topic descriptions, and relevance assessments) was taken from the test collection [4] used for Cross-Language Speech-Retrieval track of the CLEF 2007 [11]. This collection was built from a part of oral history archive from the Malach Project[1]. This archive consists of 52,000 Holocaust survivors' testimonies in 32 languages. A subset of 357 testimonies recorded in Czech was manually processed by human assessors and passages relevant to 118 topics were identified for the purposes of the CLEF evaluation campaign. 32 topics were assessed by at least two assessors in parallel. The assessors identified 5 436 relevant segments with an average duration of 167 seconds. An example of a test topic is given in Table 1. The description consists of four parts – numerical ID, title, short description, and a more verbose narrative. All the topics are related Holocaust, Word War II, etc. An average length of a testimony in the test collection is approximately 95 min.

3.2 User Interface

For the purpose of our survey we have developed a custom user interface, implemented as an on-line application in the Flex programming language[2] to be easily used over the Internet (in a web browser). Participants of the survey did not have to download the application and data to their computers what reduced their effort. A screenshot of the interface is displayed in Figure 2.

The key component of the interface is an audio player which allows the survey participants to listen and navigate through the presented recordings. The interface also displays the topics. The player control buttons include the standard play and pause buttons, volume indicator, and a large slider for precise navigation in the recording, as well as buttons for fast forward and backward jump

[1] `http://malach.umiacs.umd.edu`
[2] `http://www.adobe.com/products/flex.html`

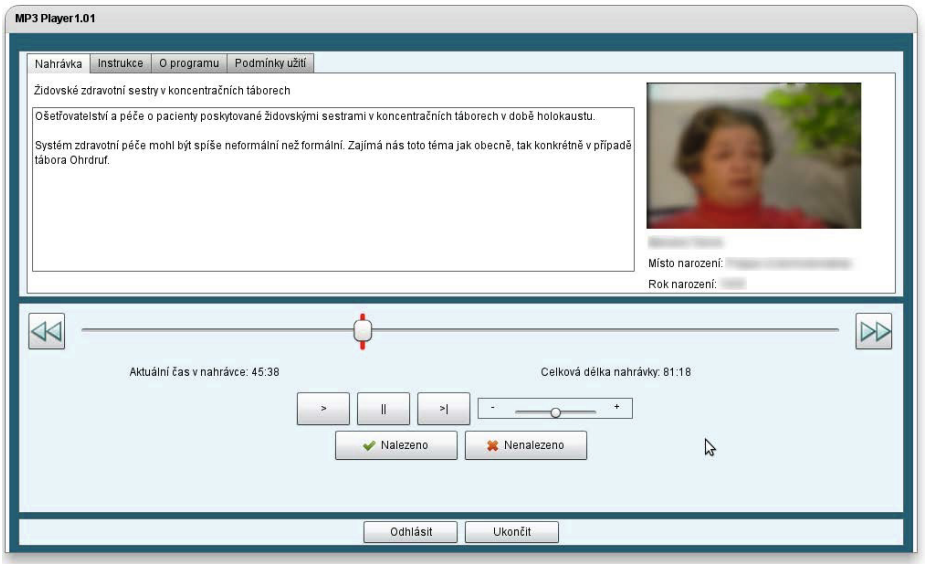

Fig. 2. A screenshot of the user interface used during the survey focused on behaviour of users analysing simulated results of a speech retrieval system

(by 30 seconds). The randomly generated starting points are indicated by a red icon on the slider (one at a time). When users identified a relevant passage they were instructed to press the "Found" ("Nalezeno") button below the control bar and indicate their level of satisfaction in a newly opened pop-up window. If the users were not satisfied with a presented starting point (and could not find a relevant passage nearby) they were allowed to proceed with the next starting point by pressing the "Not Found" ("Nenalezeno") button, but they could not return back. Some additional information was accessible through the interface: description of the topic being processed, details of the current speaker (picture and some basic information), survey instructions, etc. All actions of the participants, such as the movement of the slider, playing and stopping the record were recorded in order to study the behaviour of the participants. As all the data used in the survey were in Czech, the language of the interface was Czech too.

3.3 Survey

The survey was designed to simulate results of a retrieval system. The participants did not input any query; instead, they were presented the topics from the test collection and playback points randomly generated in a vicinity of a starting point of a relevant segment. The survey data was prepared as follows. First, we removed topics which were assessed by one assessor only and topics which had less than 5 assessed relevant segments. For each of the remaining topics, we randomly selected a set of seven relevant segments and their starting points. For each of the true starting points we randomly generated one simulated starting point which was

presented to the participants. The absolute position of this point was drawn from a normal distribution with mean set to the position of the true starting point and variance empirically set to reflect the real lengths of relevant passages identified in the test collection: the mean of the length of the segments is 2.73 minutes and the standard deviation value is 2.92. The resulting pool of randomly generated playback starting points consisted of 257 playback times in 157 recordings. The order of the playback points presented to the survey participants was random but identical for each participant.

The participants of the survey were volunteers who were asked to work for at least 15 minutes. A total of 24 users participated in the survey and they analysed 263 starting points. The average time spent per participant was 1 hour.

Randomly placed playback points were displayed one per record to the participants of the survey. Each playback point was marked on the time slider of the audio player. The true starting point of the topic was hidden from the participants. The participants were instructed to get familiar with the given topic first. Then, they started to play the audio from the simulated playback point and listened. Participants were allowed to navigate in the recording and instructed to indicate when the speaker started to talk about the given topic (beginning of a relevant passage) or when they were not able to find a relevant passage. After they found the relevant segment, the participants were asked to indicate their satisfaction with the playback point (how easy it was to find a beginning of a relevant passage) on a four-point scale: *very good*, *good*, *bad*, and *very bad*.

4 Results

As we have mentioned earlier, we consider two factors as indicators of the quality of the (simulated) retrieval results: a) the time needed to find the starting point of a passage relevant to the given topic and b) the overall satisfaction with the retrieval result (i.e. the location of the playback point). This allows us to analyse correlation of these two factors with the relative position of the starting point of the true relevant passages.

4.1 Time Analysis

Time needed for finding the relevant information is an important measure of quality of an IR system [1]. In our user study, 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. Figure 3 visualizes these values on the vertical axis with respect to the difference between the simulated playback points and true starting points on the horizontal axis.

The key observation is that the respondents generally need less time to complete the task when the playback point is located *before* the true starting point. For the playback points generated 3 minutes before the true starting point the

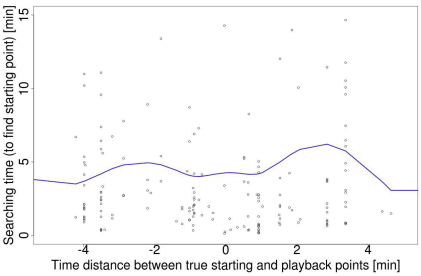

Fig. 3. Time needed for indicating that a relevant passage was found versus distance of the playback points from the true starting point.

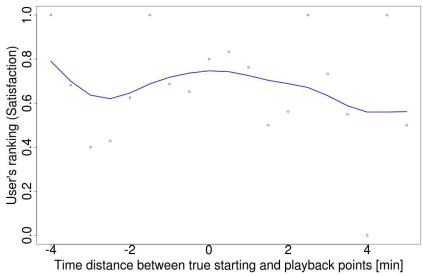

Fig. 4. Average retrieval satisfaction of respondents versus distance of the playback starting points from the true starting points.

average time needed to find the relevant segment is 1.7 minutes. For the playback points generated 3 minutes after the true starting point the average time needed to find the relevant segment is 2.1 minutes. With increasing distance from the true starting points the situation changes. The average time needed to find a relevant segment is 3 minutes when the playback point lies 5 minutes before the true starting point and 2.6 minutes when the playback point lies 5 minutes after starting point. However, this is very biased by a number of cases when the respondents gave up searching the relevant passages at all. There were 68 such cases (26%) and most of them happened when the generated playback points appeared 5 to 3 minutes before the true starting point.

When a playback point is placed closer than one minute to the true starting point, the time needed to mark the starting point is almost the same as if the playback reference and true starting points were coincident. When the time between true starting and reference points is more than four minutes, the values are distorted due to the smaller number of observations.

4.2 Users' Satisfaction

The second aspect is the overall (subjective) satisfaction with the playback points in terms of retrieval quality. During the survey, 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*. This scale was then transformed into real number values: the responses *very good* and *good* were assigned 1, and *bad* and *very bad* were assigned 0. The cases in which no starting point was found were treated as *very bad* and assigned 0. Then the arithmetic mean of these values from all respondents was calculated for each generated playback point. Visualization of the results is shown in Figure 4.

The trend of the spline function generated from the satisfaction values is not clear. The respondents seem to be most satisfied when the playback reference point lies shortly before the true starting point (negative values). On the other

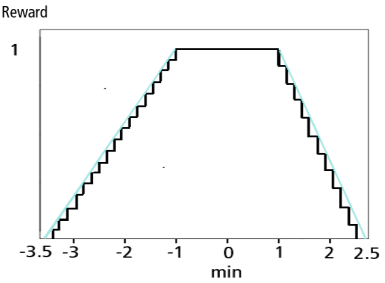

Fig. 5. The proposed modification of Penalty Function

hand, the function value decreases more slowly for positive time when the playback point lies after the true starting point. This means that if a starting point is retrieved, for example, two minutes after the true starting point it is likely that the speaker is still talking about the topic, the participant could guess where the topic starts and he/she judges the retrieval result to be better. This stands against the results of the time needed to mark the starting point, though.

5 Proposed mGAP Modifications

If we want to propose a modification of the mGAP Penalty Function which would better reflect user perception of speech retrieval quality, the following findings of the user study should be taken into account:

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 (negative time distance).
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".

Our proposal of the modified mGAP Penalty Function based on these findings is shown in Figure 5. The "width" of this new function for positive time values is 2.5 minutes. This time corresponds to the average length of a speaker's talk on one topic in Malach data collection[3]. The average length of the topic may differ for various collections. Therefore, the possibility of arranging this time according to the recordings collection specification should be further studied. Because of the better results of the points in negative time we enlarged the width of this function in the negative time region to 3.5 minutes. We decided not to take into

[3] This information comes from the data used in the CLEF evaluation campaign.

Table 2. mGAP scores of the retrieval systems participating in the CLEF 2007 CL-SR track calculated with original and modified Penalty Functions.

Submission	*Team*	*Orig. PF*	*Modif. PF*	*Difference*
UWB_2-1_tdn_l	University of West Bohemia	0.0274	0.0490	0.0216
UWB_3-1_tdn_l	University of West Bohemia	0.0241	0.0517	0.0276
UWB_2-1_td_s	University of West Bohemia	0.0229	0.0383	0.0154
UCcsaTD2	University of Chicago	0.0213	0.0387	0.0174
UCcslTD1	University of Chicago	0.0196	0.0359	0.0163
prague04	Charles University in Prague	0.0195	0.0373	0.0178
prague01	Charles University in Prague	0.0192	0.0370	0.0178
prague02	Charles University in Prague	0.0183	0.0347	0.0164
UWB_3-1_td_l	University of West Bohemia	0.0134	0.0256	0.0122
UWB_2-1_td_w	University of West Bohemia	0.0132	0.0255	0.0123
UCunstTD3	University of Chicago	0.0126	0.0270	0.0144
brown.s.f	Brown University	0.0113	0.0258	0.0145
brown.sA.f	Brown University	0.0106	0.0242	0.0136
prague03	Charles University in Prague	0.0098	0.0208	0.011
brown.f	Brown University	0.0049	0.0131	0.0082

account the fact that users prefer playback points lying before starting points of true relevant segments in a greater distance. Starting point retrieved closer than one minute to the true starting point is considered to be equally good as exact match. This reflects the tolerance of smaller nuances in retrieval which are difficult to recognize even by a human.

The reward assigned by the modified Penalty Function will always be higher than the one from the original Penalty Function. Consequently, the mGAP score calculated using the proposed function will be higher too.

5.1 Comparison with the Original Measure

We evaluate the impact of the proposed modification of the Penalty Function in the setting of the CLEF 2007 Cross-Language Speech Retrieval Track [11]. We have rescored all 15 retrieval systems which participated in the task using mGAP with the modified Penalty Function and we have compared the results with the original scores, see Table 2. Visual comparison is then shown in Figure 6.

The original and new scores are quite correlated, the final rankings of the retrieval systems differ only in a few cases and the absolute changes are relatively small and not significant. The high correlation is mainly caused by the large amount of cases in which is the Penalty Function equal to 0: Almost 98% of all Penalty Function values are equal to 0. Figure 7 illustrates in how many cases the scoring (reward) of individual retrieved points actually changed when the modified Penalty Function was applied. This nicely corresponds with the modified shape of the Penalty Function.

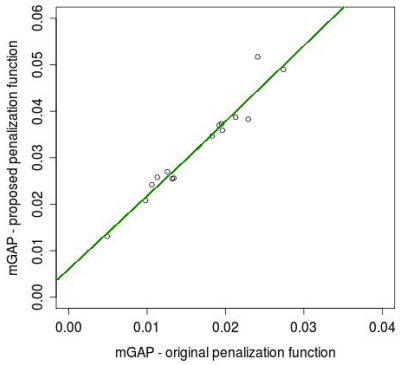

Fig. 6. Comparison of the scores calculated by mGAP with original and modified Penalty Function.

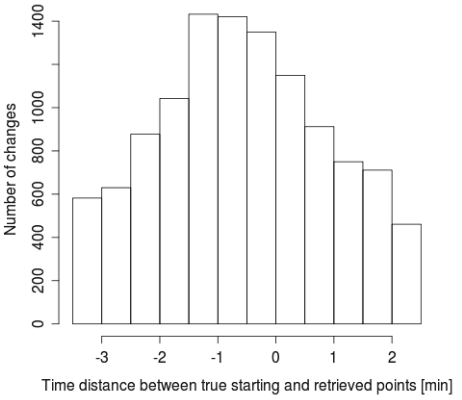

Fig. 7. Distribution of reward changes using the original and modified penalty function on the CL-SR CLEF 2007 results.

6 Conclusion

We have examined metrics used for evaluation of information retrieval from speech recordings. Our main focus was on the mGAP measure, which is currently often used for retrieval of unsegmented recordings. Several drawbacks of this measure were described and an experiment to help to improve this measure was proposed. At the core of the experiment was a human-based lab test in which participants were asked to search for the starting point of a particular topic. A total of 24 respondents participated in this test. A modified Penalty Function to be used in the mGAP measure was proposed based on our test results. The three most significant modifications to the original Penalty Function are that the new Penalty function is "wider" than the original one, the new Penalty Function prefers IR systems which retrieve a topic starting point before the true annotated starting point and if the IR system retrieves a starting point closer than one minute from the annotated point, there is no penalty. Finally, a comparison of the original and modified Penalty Functions was performed using real data from retrieval systems used in CLEF 2007 track and a high correlation between the outputs of the mGAP measure with the two Penalty Functions has been found. As a result, the original ranking of retrieval system from CL-SR CLEF 2007 changed only insignificantly.

Acknowledgements. This research was supported by the project AMALACH (grant no. DF12P01OVV022 of the program NAKI of the Ministry of Culture of the Czech Republic), the Czech Science Foundation (grant n. P103/12/G084) and SVV project number 265 314.

References

1. Cleverdon, C.W., Mills, J., Keen, M.: Factors determining the performance of indexing systems. Test results, vol. 2. Aslib Cranfield Research Project, Cranfield, England (1966)
2. Eskevich, M., Magdy, W., Jones, G.J.F.: New Metrics for Meaningful Evaluation of Informally Structured Speech Retrieval. In: Baeza-Yates, R., de Vries, A.P., Zaragoza, H., Cambazoglu, B.B., Murdock, V., Lempel, R., Silvestri, F. (eds.) ECIR 2012. LNCS, vol. 7224, pp. 170–181. Springer, Heidelberg (2012)
3. Garofolo, J., Auzanne, C., Ellen, V., Sparck, J.K.: 1999 TREC-8 Spoken Document Retrieval (SDR) Track Evaluation Specification (1999), `http://www.itl.nist.gov/iad/mig/tests/sdr/1999/spec.html`
4. Ircing, P., Pecina, P., Oard, D.W., Wang, J., White, R.W., Hoidekr, J.: Information Retrieval Test Collection for Searching Spontaneous Czech Speech. In: Matoušek, V., Mautner, P. (eds.) TSD 2007. LNCS (LNAI), vol. 4629, pp. 439–446. Springer, Heidelberg (2007)
5. Kamps, J., Pehcevski, J., Kazai, G., Lalmas, M., Robertson, S.: INEX 2007 Evaluation Measures. In: Fuhr, N., Kamps, J., Lalmas, M., Trotman, A. (eds.) INEX 2007. LNCS, vol. 4862, pp. 24–33. Springer, Heidelberg (2008)
6. Larson, M., Eskevich, M., Ordelman, R., Kofler, C., Schmiedeke, S., Jones, G.J.F.: Overview of MediaEval 2011 rich speech retrieval task and genre tagging task. In: Larson, M., Rae, A., Demarty, C.H., Kofler, C., Metze, F., Troncy, R., Mezaris, V., Jones, G.J.F. (eds.) Working Notes Proceedings of the MediaEval 2011 Workshop. CEUR Workshop Proceedings, vol. 807, pp. 1–2. CEUR-WS.org (2011)
7. Liu, B., Oard, D.W.: One-sided measures for evaluating ranked retrieval effectiveness with spontaneous conversational speech. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2006, pp. 673–674. ACM, New York (2006)
8. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, New York (2008)
9. Oard, D.W., Hackett, P.G.: Document Translation for Cross-Language Text Retrieval at the University of Maryland. In: Voorhees, E.M., Harman, D.K. (eds.) The Sixth Text REtrieval Conference (TREC-6), pp. 687–696. U.S. Dept. of Commerce, Technology Administration, National Institute of Standards and Technology (1997)
10. Oard, D.W., Wang, J., Jones, G.J.F., White, R.W., Pecina, P., Soergel, D., Huang, X., Shafran, I.: Overview of the CLEF-2006 Cross-Language Speech Retrieval Track. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 744–758. Springer, Heidelberg (2007)
11. Pecina, P., Hoffmannová, P., Jones, G.J.F., Zhang, Y., Oard, D.W.: Overview of the CLEF-2007 Cross-Language Speech Retrieval Track. In: Peters, C., Jijkoun, V., Mandl, T., Müller, H., Oard, D.W., Peñas, A., Petras, V., Santos, D. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 674–686. Springer, Heidelberg (2008)