Query-by-Example Spoken Term Detection ALBAYZIN 2012 evaluation: overview, systems, results and discussion

Javier Tejedor*[1] , Doroteo T. Toledano[2], Xavier Anguera[3] , Amparo Varona[4] , Lluís-F. Hurtado[5] , Antonio Miguel[6] and José Colás[1]

[1]Human Computer Technology Laboratory (HCTLab), Universidad Autónoma de Madrid, Spain
[2]Biometric Recognition Group - ATVS, Universidad Autónoma de Madrid, Spain
[3]Telefónica Research, Barcelona, Spain
[4]Working Group on Software Technologies (GTTS), University of the Basque Country, Spain
[5]Departament de Sistemes Informàtics i Computació, Universitat Politècnica de València, Spain
[6]Voice Input Voice Output Laboratory (ViVoLab), Aragon Institute for Engineering Research (I3A), University of Zaragoza, Spain

Email: Javier Tejedor*- javier.tejedor@uam.es; Doroteo T. Toledano - doroteo.torre@uam.es; Xavier Anguera - xanguera@tid.es; Amparo Varona - amparo.varona@ehu.es; Lluís-F. Hurtado - lhurtado@dsic.upv.es; Antonio Miguel - amiguel@unizar.es; José Colás - jose.colas@uam.es;

*Corresponding author

Abstract

Query-by-Example Spoken Term Detection (QbE STD) aims at retrieving data from a speech data repository given an acoustic query containing the term of interest as input. Nowadays, it has been receiving much interest due to the high volume of information stored in audio or audiovisual format. QbE STD differs from Automatic Speech Recognition (ASR) and Keyword Spotting (KWS)/Spoken Term Detection (STD) since ASR is interested in all the terms/words that appear in the speech signal and KWS/STD relies on a textual transcription of the search term to retrieve the speech data. This paper presents the systems submitted to the ALBAYZIN 2012 QbE STD evaluation held as a part of ALBAYZIN 2012 evaluation campaign within the context of the IberSPEECH 2012 conference[1]. The evaluation consists of retrieving the speech files that contain the input queries, indicating their start and end timestamps within the appropriate speech file. Evaluation is conducted on a Spanish spontaneous speech database containing a set of talks from MAVIR workshops[2] which amount at about 7 hours of speech in total. We present the database, metric, systems submitted along with all results and some discussion. 4 different research groups took part in the evaluation. Evaluation results show the

[1]http://iberspeech2012.ii.uam.es/
[2]http://www.mavir.net

difficulty of this task and the limited performance indicates there is still a lot of room for improvement. The best result is achieved by a DTW-based search over Gaussian posteriorgrams/posterior phoneme probabilities. This paper also compares the systems aiming at establishing the best technique dealing with that difficult task, and looking for defining promising directions for this novel task.

Keywords: Query-by-Example, Spoken Term Detection, International Evaluation, Search on Spontaneous Speech

Introduction

The ever-increasing volume of heterogeneous speech data stored in audio and audiovisual repositories promotes the development of efficient methods for retrieving the stored information. Much work has addressed this issue by means of spoken document retrieval (SDR), keyword spotting, spoken term detection (STD), query-by-example (QbE) or spoken query approaches.

Spoken Term Detection aims at finding individual words or sequences of words within audio archives. Therefore, it relies on a text-based input, commonly the phone transcription of the search term. STD systems are typically composed of three different stages: first, the audio is decoded in terms of word/sub-word lattices from an Automatic Speech Recognition subsystem. Next, a Term Detection subsystem employs the phone transcription corresponding to the search term to find the term within those word/sub-word lattices and hence to hypothesize detections. And finally, confidence measures can be applied to output reliable detections.

Query-by-Example can be defined as "a method of searching for an example of an object or a part of it in other objects". QbE has been widely used in audio applications like sound classification [1–3], music retrieval [4, 5] and spoken document retrieval [6]. In QbE STD, we consider the scenario where the user has found some interesting data within a speech data repository (for example, by random browsing or some other method). His/her purpose is to find similar data within the repository. In doing so, the user selects one or several speech cuts containing the term of interest (henceforth, query) and the system outputs him/her other putative hits from the repository (henceforth, utterances). Another scenario for QbE STD considers one or several user speech recordings of the term of interest. Therefore, QbE STD differs from the STD defined previously, so called *text-based STD*, in that the former uses an acoustic query as input, instead of a text-based representation of the term. This, on the one hand, offers a big advantage for devices without text-based capabilities, which can be effectively used under the QbE STD paradigm. On the other

hand, QbE STD can be also employed for building language-independent STD systems [7, 8], which is mandatory when no or very limited training data are available to build a reliable speech recognition system, since a prior knowledge of the language involved in the speech data is not necessary.
QbE STD has been addressed in the literature from two different points of view:

1. Methods based on phonetic transcription of the query speech signal [7, 9–17], for which the *text-based STD* technology can be next applied. Therefore, these methods decode the query with an automatic speech recognizer to get its word/sub-word representation which can be next employed to hypothesize detections in a *text-based STD*-like system.

2. Methods based on template matching of features extracted from the query/utterance speech signal [7, 8, 17–29]. They usually borrow the idea from Dynamic Time Warping (DTW)-based speech recognition and were found to outperform phonetic transcription-based techniques on QbE STD [18].

Given the high amount of information stored in speech format, automatic systems that are able to provide access to this content are necessary. In this direction, several evaluations including SDR, STD and QbE STD evaluations have been proposed recently [30–36]. Taking into account the increasing interest in the QbE STD evaluation around the world, we organized an international evaluation of QbE STD in the context of ALBAYZIN 2012 evaluation campaign. This campaign is an internationally open set of evaluations supported by the Spanish Network of Speech Technologies (RTTH[3]) and the ISCA Special Interest Group on Iberian Languages (SIG-IL) every 2 years from 2006. Evaluation campaigns provide an objective mechanism to compare different systems and to promote research on different speech technologies such as: speech segmentation [37], speaker diarization [38], language recognition [39], and speech synthesis [40] in the ALBAYZIN 2010 evaluation campaign. This year, this campaign has been held during IberSPEECH 2012 conference[4], which integrated "VII Jornadas en Tecnología del Habla" and "III Iberian SLTech Workshop".
The rest of the paper is organized as follows: next section presents the QbE STD evaluation that includes an evaluation description, the metric used, the database released for experimentation and the participants involved in the evaluation. Next, we present the different systems submitted to the evaluation. Results along with some discussion are presented in Section "Results and Discussion" and the work is concluded in the last section.

[3]http://www.rthabla.es/
[4]http://iberspeech2012.ii.uam.es/

Query-by-Example Spoken Term Detection evaluation

Evaluation description and metric

This evaluation involves searching *for* audio content *within* audio content *using* an audio content query. Therefore, this is suitable for groups working on speech indexing and retrieval and on speech recognition as well. In other words, this task focuses on retrieving the appropriate audio files, with the occurrences and timestamps, which contain any of those queries. Therefore, the input to the system is an acoustic example per query and hence a prior knowledge of the correct word/phone transcription corresponding to each query cannot be made. This is the same task that the one proposed in MediaEval 2011 and 2012 Search on Speech evaluations [33, 34]. However, this is the first QbE STD evaluation that deals with Spanish language. This makes our evaluation different compared to that proposed in MediaEval 2011 and 2012 evaluations [33, 34], which dealt with Indian and African languages. In addition, participants of our evaluation could make use of the language knowledge (i.e., Spanish) when building their system/s. Participants could submit a primary system and up to 2 contrastive systems. No manual intervention is allowed for each system developed to generate the final output file and hence, all the developed systems must be fully automatic. Listening to the test data, or any other human interaction with the test data is forbidden before all the results have been submitted. The standard XML-based format corresponding to the NIST STD 2006 evaluation [31] has been used for building the system output file.

In QbE STD, a hypothesized occurrence is called a *detection*; if the detection corresponds to an actual occurrence, it is called a *hit*, otherwise it is a *false alarm* (FA). If an actual occurrence is not detected, this is called a *miss*. The Actual Term Weighted Value (ATWV) [31] has been used as metric for the evaluation. This integrates the hit rate and false alarm rate of each query term into a single metric and then averages over all search query terms:

$$ATWV = \frac{1}{|\Delta|} \sum_{K \in \Delta} \left(\frac{N_{hit}^K}{N_{true}^K} - \beta \frac{N_{FA}^K}{T - N_{true}^K} \right) \quad (1)$$

where Δ denotes the set of search query terms and $|\Delta|$ is the number of query terms in this set. N_{hit}^K and N_{FA}^K respectively represent the numbers of hits and false alarms of query term K and N_{true}^K is the number of actual occurrences of K in the audio. T denotes the audio length in seconds, and β is a weight factor set to 999.9.

ATWV represents the TWV for the threshold set by the QbE STD system (usually tuned on development data). An additional metric, called Maximum Term Weighted Value (MTWV) [31] can be also used to evaluate the performance of a QbE STD system. This MTWV is the maximum TWV achieved by a given

QbE STD system and does not depend on the tuned threshold. Although it was not used for the evaluation, results based on this metric are also presented to measure the threshold calibration in the submitted systems.

In addition to ATWV and MTWV, NIST also proposed a detection error tradeoff (DET) curve [41] to evaluate the performance of a QbE STD system working at various miss/FA ratios. Although DET curves were not used for the evaluation itself either, they are also presented in this paper for system comparison.

Database

The database used for the evaluation consists of a set of talks extracted from the Spanish MAVIR workshops[5] held in 2006, 2007 and 2008 (Corpus MAVIR 2006, 2007 and 2008) corresponding to Spanish language (henceforth MAVIR database).

This MAVIR database includes 10 spontaneous speech files, each containing a single different speaker, which amount at about 7 hours of speech and are further divided into training/development and test sets. These data were also manually annotated in an orthographic form. The speech data were originally recorded in several audio formats (PCM mono and stereo, MP3, etc). All data were converted to PCM, 16khz, single channel, 16 bits per sample using Sox tool[6]. The spontaneous speech of this database made this appealing enough for our evaluation.

Training/development data amount at about 5 hours of speech extracted from 7 out of the 10 speech files of the MAVIR database. However, there is no constraint in the amount of training/development data that can be employed to build the systems. The training/development list of queries consists of 60 queries, which were chosen based on their occurrence rate in the training/development speech data. Each query is composed of a single word whose length varies between 7 and 16 single graphemes. Ground truth labels and evaluation tools were provided to the participants by the date of the release.

Test data amount at about 2 hours of speech extracted from the other 3 speech files not used as training/development data. The test list of queries consists of 60 queries, which were chosen based on their occurrence rate in the test speech data. Each query is composed of a single word whose length varies between 7 and 16 single graphemes. No ground truth labels corresponding to the test data were given to the participants until all the systems were submitted to the evaluation. Table 1 includes information related to the test queries.

[5] http://www.mavir.net
[6] http://sox.sourceforge.net/

Participants

4 different systems (Systems 1-4) were submitted from 3 different research groups to ALBAYZIN 2012 Query-by-Example Spoken Term Detection evaluation. In addition, one additional research group submitted a system (named Text-based STD system in this paper) that is capable of *text-based STD*. This system will be used in this paper as a reliable baseline to be compared with the systems submitted to the main QbE STD evaluation. Participants are listed in Table 2. About 3 months were given to the participants for system design. Training/development data were released at the end of June 2012, test data were released at the beginning of September 2012 and the final system submission was due at the end of September 2012.

Systems

In this section, systems submitted to the evaluation are described. The systems appear in the same order that they are ranked in Table 2. A full description of the systems can be found in IberSPEECH 2012 online conference proceedings [42].

System 1

The system is based on a DTW zero-resource matching approach. First, acoustic features (13 Mel frequency cepstral coefficients (MFCCs) along with their first and second derivatives) were extracted from the speech signal for each frame. To solve the speaker dependent issue that these features suffer from [8], these MFCC features are used to train a posterior Gaussian Mixture Model (GMM). This GMM is trained from a combination of Expectation-Maximization and K-means algorithms aiming at maximizing the discovery and separation of automatically derived acoustic regions in the speech signal, as described in [43]. Finally, Gaussian posteriorgram features are extracted from this model as final features. Next, a GMM-based speech/silence detector is applied to filter out non-speech segments. The resulting features (i.e., those corresponding to speech segments) are next sent to the subsequence-DTW (SDTW) [44] matching algorithm, which hypothesizes query detections within the utterances. The minus logarithm of the cosine distance has been employed as similarity measure between each query frame and each utterance frame. This SDTW algorithm allows any query to appear at any time within the utterance. After the matching algorithm returns all possible detections and their scores, an overlap detection algorithm is executed where all those matches that overlap with each other more than 50% of the detection time are post-processed by keeping the detection with the highest score (i.e., the lowest distance) in the output file

along with the non-overlapped detections. It must be noted that this system can be considered language-independent, since it does not make use of the target language and can be effectively used for building language-independent STD systems.

System 2

This system looks for an exact match of the phone sequence output by a speech recognition process given a spoken query, within the phone lattices corresponding to the utterances. Brno University of Technology phone decoders for Czech, Hungarian and Russian have been employed [45]. In this way, this system does not make use of a prior knowledge of the target language (i.e., Spanish) and hence, as the previous system, is language-independent and suitable for building a language-independent STD system.

The system integrates different stages, as follows: first, Czech, Hungarian and Russian phone decoders have been used to produce phone lattices both for queries and utterances. Then, the phone transcription corresponding to each query is extracted from the phone lattice by taking the highest likelihood phone sequence using the *lattice tool* of SRILM [46]. Next, *Lattice2Multigram* tool [47] [48] [49][7] has been used to hypothesize detections that perform an exact match of the phone transcription of each query within each utterance. In this way, three different output files that contain the detections from each phone decoder are obtained. The score given by the *Lattice2Multigram* tool for each detection is normalized by the length of the detection (in number of frames) and by all the detections found within the phone lattices except the current one. Overlapped detections that are hypothesized by two or more phone decoders are merged so that the most likely detection (i.e., the one with the highest score) remains along with the non-overlapped detections. As a post-process, just the best K hypothesis for each utterance are kept in the final output file. K was set to 50 which got the best performance in training/development data.

Two different configurations for this system were submitted. The first one, refers as System 2a in Table 3, combines the detections from the Hungarian and Russian phone decoders, since they got the best performance in the training/development data. The second one, refers as System 2b in Table 3, merges the detections from all the phone decoders (i.e., Czech, Hungarian and Russian) in the final output file.

System 3

The system is based on a search on phoneme lattices generated from *a posteriori phoneme probabilities*. This is composed of different stages, as follows: first, these probabilities are obtained by combining the

[7] http://homepages.inf.ed.ac.uk/v1dwang2/public/tools/index.html

acoustic class probabilities estimated from a clustering procedure on the acoustic space, and the conditional probabilities of each acoustic class with respect to each phonetic unit [50]. The clustering makes use of standard GMM distributions for each acoustic class, which are estimated from the unsupervised way of the Maximum Likelihood Estimation procedure. The conditional probabilities are obtained from a coarse segmentation procedure [51]. An acoustic class represents a phone in the target language (i.e., Spanish) and hence this system employs the knowledge of the target language. Second, the phoneme lattices are obtained for each query and utterance from an ASR process that takes as input the phoneme probabilities computed in the previous stage. This ASR process examines if each vector of phoneme probabilities contains probabilities for each phoneme above a predefined *detection* threshold (tuned on training/development data) to output a specific phoneme for each frame. Start and end time marks for each phoneme are assigned from backward/forward procedures that mark frames before/after the current one with a probability for that phoneme higher than an *extension* threshold (tuned on training/development data as well) stopping when the probability is lower than this threshold to assign the corresponding start and end timestamps. The accumulated frame phoneme probability is used as score for each phoneme in the lattice. In a third step, a search of every path in the lattice corresponding to the query within the phoneme lattice corresponding to the utterance is conducted to hypothesize detections. Substitution, deletion and insertion errors in those query lattice paths are allowed when hypothesizing detections. Score for each detection is computed by accumulating the individual score for each phoneme both in the query and the utterance lattice paths. Overlapped detections are discarded in the final output file by keeping the best, and detections with a score lower than a predefined threshold (tuned on the training/development data) are also filtered out the final output. This threshold is query-dependent since a query detection is considered a hit if its score is lower than the mean of all the scores of this query minus the standard deviation of these scores computed from all the occurrences of the detected query in all the speech files.
Two different configurations were submitted. The first one, referred as System 3a in Table 3, tuned all the thresholds so that at least a 6% of hits are produced. The second one, referred as System 3b in Table 3, is a *late submission* and tuned the thresholds for ATWV performance. This second configuration allows a fair comparison with the rest of the systems submitted.

System 4

This system employs the same phoneme probabilities used in the first stage to build System 3 as query/utterance representation and hence it makes use of the target language. To hypothesize detections, a

SDTW search [52] is conducted with the Kullback-Leibler (KL) divergence as similarity measure between each query frame and each utterance frame. The SDTW algorithm allows any query to appear at any point within the utterance. Overlapped detections found by the SDTW search and detections with a score lower than a predefined threshold (tuned on the training/development data) are filtered out the final output. As in System 3, this threshold is query-dependent and a query detection is considered a hit if its score is lower than the mean of all the scores of this query minus the standard deviation of these scores computed from all the occurrences of the detected query in all the speech files.
As in the previous system, two different configurations were submitted. The first one, referred as System 4a in Table 3, optimizes the system so that at least a 10% of hits are produced. The second one, referred as System 4b in Table 3, is a *late submission*, optimizes the system according to ATWV metric and hence only allows a query to have 2 detections as many in all the speech files. This system optimization towards the ATWV metric allows a fair comparison with the rest of the systems submitted.

Text-based Spoken Term Detection system

For comparison with the systems presented before, we present a system that can conduct STD since it employs the phone transcription corresponding to each query to hypothesize detections. It must be noted that the correct phone transcription corresponding to each search term has been employed.
The STD system consists of four different stages: in the first stage, a phone recognition is conducted to output phone lattices based on two different speech recognizers: (1) a standard triphone context-dependent Hidden Markov Model (HMM) speech recognizer with mixtures of diagonal covariance Gaussians as observation density functions in the states, and (2) a biphone context-dependent HMM speech recognizer where the observation probabilities are obtained from a multilayer perceptron (MLP). In the second stage, a STD subsystem hypothesizes detections from each speech recognizer. The 1-best output of each phonetic recognizer is used as source text for an edit distance search. In doing so, each putative detection could be any substring which has a phonetic edit distance with the searched word of less than 50% of its length. Next, we take all the detections found from the different phonetic recognizers and merge them. For overlapped detections, the best detection (i.e., the one with the minimum edit distance) remains. In the third stage, two different confidence measures based on minimum edit distance and lattice information are used as confidence scores for each putative detection. The former is computed from standard substitution, insertion and deletion errors in the 1-best phone sequence given by each speech recognizer and normalized by the length of the word. The latter is computed as follows: (1) we determinize each lattice by using

HLRescore from HTK [53] so that a smaller and more useful graph is used next, (2) we run the *lattice-tool* from the SRILM toolkit [46] to obtain the corresponding acoustic mesh graph and (3) the confidence calculated in the acoustic mesh graph is used in a modified edit distance algorithm where, instead of all costs equal to 1, we simply sum the confidence of matching phones with the searched word. Then, the score of a putative detection is the sum of the confidences through the acoustic mesh of the searched word between the time limits where the detection resides. This score is also normalized by the length of the word. And the fourth stage makes use of the Bosaris toolkit [8] to fuse both scores obtained in the previous stage to compute the final confidence for each detection.

Results and discussion

Results of the QbE STD evaluation are presented in Table 3 for every system submitted by the participants along with the system applied on STD in terms of MTWV and ATWV.

By analyzing the systems submitted to the QbE STD evaluation at due time (i.e., not considering the late submissions), System 1 achieved the best performance both in terms of MTWV and ATWV. This reflects the good calibration approach used to score each detection. It must be noted that both the difficulty of the task itself (searching acoustic queries on spontaneous data) and the non-prior knowledge of the target language produce this low performance. However, this system is worse than the Text-based STD system. This, as expected, is due to the use of the correct phone transcription for each query and hence the knowledge of the target language employed to build the Text-based STD system.

Special mention requires the late submission corresponding to System 4b. Although this system performance is not the best in terms of MTWV, this achieves the best ATWV. This is caused by the near MTWV and ATWV system performance which reflects the fact that the threshold tuned on the training/development data performs very well on unseen (test) data. This maybe due to several factors: (1) first, the 2 occurrences per query limitation produces less detections in the final output, which seriously limits the MTWV system performance, and (2) the query-dependent threshold plays a very important role as *score normalization*. The best ATWV performance of this system maybe due to the similarity measure used to conduct the SDTW search, being the Kullback-Leibler divergence, perfectly fits the posterior probabilities computed in the first stage. The use of the target language to estimate these posterior probabilities also contributes to this. However, in case of System 1, a prior knowledge of the target language was not applied, and the cosine distance may not fit the Gaussian posterior probabilities as well

[8] https://sites.google.com/site/bosaristoolkit/

as the KL divergence, which cause a worse score calibration, and hence, the higher gap between MTWV and ATWV. Again, this System 4b still underperforms the Text-based STD system.

It can be also seen that System 2a underperforms System 2b. This means that the addition of the Czech decoder is actually helping the QbE STD system. However, in the development data, the opposite occurred (0.013 vs. 0.009). This maybe due to the different development and test queries provided by the organizers. Systems 1 and 2a,b do not make use of the target language whereas Systems 3a, 3b, 4a and 4b do. In particular, it is highly remarkable the best overall performance of the System 1 in terms of MTWV, which can be employed to build language-independent STD systems. A better score calibration of this system is necessary to get nearer MTWV to ATWV system performance.

Although these results cannot be directly compared with those obtained in MediaEval 2011 and 2012 Search on Speech evaluations [33, 34], since the database used for experimentation is different, we can mention that our results are worse than those. This maybe due to the generous time windows allowed in MediaEval 2011 Search on Speech Evaluation and the equal weight given to miss and FA detections when scoring MediaEval 2012 Search on Speech Evaluation systems, which got higher the ATWV performance. In our case, we have been 100% compliant with the ATWV setup, parameters, and scoring provided by NIST. The fast speaking speed or noise background in some test queries maybe also causing this worse system performance.

DET curves are also presented in Figure 1. They show the system performance working at different miss/FA ratios. System 1 clearly outperforms the rest of the QbE STD systems for almost every operating point, except when the Miss rate is low, where System 4a performs the best, and at the best operating point of System 4b. As expected from the ATWV results, by comparing the Text-based STD system with the rest, the former outperforms the others except when the FA rate is low, where System 1 performs the best.

A more detailed analysis is presented in Figure 2 in terms of hit/FA performance for the different systems. As expected from the ATWV results, the late submission corresponding to System 4b achieves the best tradeoff between hits and FAs between those submitted to the QbE STD evaluation. Systems 2a and 2b just output a few detections which results in a worse ATWV performance. It must be noted that these two systems (2a and 2b) dramatically increase the number of FAs as long as more detections are hypothesized, in such a way that the best ATWV result is achieved with a small number of hits and FAs. System 3b exhibits a similar behavior. Systems 3a and 4a achieve such a high number of FAs that their ATWV performance decreases dramatically. This is due to both systems were developed by producing at least a 6% and a 10% coverage of hits in the training/development data respectively, which increases both the

number of hits and FAs. However, the increase in the number of FAs is much higher than the increase in the number of hits, resulting in an overall worse ATWV performance. Again, System 1 achieves the best result in terms of hit/FA performance comparing the systems submitted at due time to the main QbE STD evaluation. Looking at the performance of the Text-based STD system (out of the main QbE STD evaluation), which conducts STD and employs the correct phone transcription of the search terms when hypothesizing detections, it produces the best ATWV result, since it gets a quite high number of hits and a small number of FAs.

Template matching-based versus phone transcription-based QbE STD

Systems 1 and 4a,b employ a template matching-based approach for QbE STD whereas Systems 2a,b and 3a,b employ a phone transcription-based approach for QbE STD. This means that the best overall performance is achieved by the template matching-based approach proposed both in Systems 1 and 4. This result confirms the conclusion presented in [18] where a template matching-based approach outperformed a phone transcription-based approach for QbE STD.

Results obtained by System 2a,b suggest that building a speech recognizer on a language different to the target language to produce phoneme lattices and a next search within these phoneme lattices is not appropriate when addressing the QbE STD task, since they are not reliable enough to represent the speech content in an out-of-language setup. And the query search algorithm employed in System 3a,b considers so many paths in the lattice that represents the query to hypothesize detections within the utterances that many FAs are generated. A better score calibration for this System 3a,b is necessary to reject as many FAs as possible.

Despite the bad performance exhibited by the configuration 4a corresponding to System 4, it must be noted that this was not optimized for the final metric (i.e., ATWV) but to get a predefined hit coverage, which greatly affects the final ATWV performance [54] and hence a fair comparison with the rest of the systems cannot be made.

Set of features for QbE STD

Different sets of features have been employed as speech signal representation: Gaussian posteriorgrams for System 1, a posteriori phoneme probabilities for Systems 3a,b and 4a,b, and 3-state MLP-based phoneme probabilities for System 2a,b. Although all these features should be fed within all the search algorithms to derive a more powerful conclusion, we can observe that Gaussian posteriorgram features are suitable for

speech signal representation due to the best performance of System 1 when no prior knowledge of the target language is used. We can also mention that the posterior phoneme probabilities used in the language-dependent setup corresponding to the late submission of System 4b are an effective representation of the speech signal due to their best ATWV performance.

Towards a language-independent STD system

From the systems submitted to this evaluation, an analysis aiming at deciding the feasibility of a language-independent STD system can be conducted. By comparing the best language-independent QbE STD system (System 1) with the Text-based STD system, we can claim that building a language-independent STD system is still a far milestone. This means that more research is needed in this direction to get nearer language-dependent to language-independent STD systems.

Challenge of the QbE STD task

By inspecting the results of all the systems submitted to the QbE STD evaluation, we can claim that building a reliable QbE STD system is still far from being a solved problem. The low ATWV performance exhibited by the best system (ATWV=0.0217) confirms this. There are many issues that must be still solved in the future. First, a robust feature extraction process is necessary to represent in an accurate way the query/utterance speech content. Next, a suitable search algorithm that hypothesizes detections is also necessary to output as many hits as possible while maintaining a reasonable low number of FAs. In addition, the spontaneous speech, inherent to QbE STD systems, is an important drawback since phenomena such as disfluences, hesitations, noises, etc, are very difficult to deal with. Some pre-processing steps that deal with these phenomena could enhance the final performance. From the systems submitted to this evaluation, we can claim that Gaussian posteriorgrams or, generally speaking, posterior phoneme probabilities, as features and a subsequent DTW-based search is a reasonable *good* starting point when facing QbE STD.

Conclusions

We have presented the 4 systems submitted to ALBAYZIN 2012 Query-by-Example Spoken Term Detection evaluation along with the system that conducts STD. 4 different Spanish research groups (TID, GTTS, ELiRF and VivoLab) took part in the evaluation. There were two different kinds of systems submitted to the evaluation: template matching-based systems and phone transcription-based systems.

Systems 1 and 4a,b belong to the former group and Systems 2a,b and 3a,b belong to the latter. It has been shown a better performance of the template matching-based systems over the systems that employ the phone transcription of each query got from a phone decoding and next a *text-based STD*-like search to hypothesize detections. The best system employs Gaussian posteriorgram/a posteriori phoneme probability features and a DTW-like search to hypothesize detections.

We have also shown that QbE STD systems (Systems 1 and 4b) are still far from systems that deal with text-based STD (Text-based STD system).

This evaluation is the first that has been conducted for Spanish language so far, which represents a good baseline for future research in this language. In addition, the spontaneous speech database chosen for the experimentation made the evaluation attractive enough. Results presented in this paper indicate that there is still a big room for improvement which encourages us to maintain this evaluation in next ALBAYZIN evaluation campaigns.

Author's contributions

- Systems submitted to the first Query-by-Example Spoken Term Detection evaluation for Spanish language are presented.
- Systems are categorized into template matching-based systems and phone transcription-based systems.
- Analysis of system submission results is presented.
- Systems are compared with a text-based Spoken Term Detection system.

References

1. Zhang T, Kuo CCJ: **Hierarchical classification of audio data for archiving and retrieving**. In *Proceedings of ICASSP* 1999:3001–3004.
2. Helén M, Virtanen T: **Query by example of audio signals using Euclidean distance between Gaussian Mixture Models**. In *Proceedings of ICASSP* 2007:225–228.
3. Helén M, Virtanen T: **Audio query by example using similarity measures between probability density functions of features**. *EURASIP, Journal on Audio, Speech and Music Processing* 2010, **2010**:2:1–2:12.
4. Tzanetakis G, Ermolinskyi A, Cook P: **Pitch histograms in audio and symbolic music information retrieval**. In *Proceedings of the Third International Conference on Music Information Retrieval: ISMIR* 2002:31–38.
5. Tsai WH, Wang HM: **A query-by-example framework to retrieve music documents by singer**. In *Proceedings of the IEEE International Conference on Multimedia and Expo* 2004:1863–1866.
6. Chia TK, Sim KC, Li H, Ng HT: **A lattice-based approach to query-by-example spoken document retrieval**. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* 2008:363–370.
7. Tejedor J, Fapšo M, Szöke I, Černocký H, Grézl F: **Comparison of methods for language-dependent and language-independent query-by-example spoken term detection**. *ACM Transactions on Information Systems* 2012, **30**(3):18:1–18:34.
8. Muscariello A, Gravier G, Bimbot F: **Zero-resource audio-only spoken term detection based on a combination of template matching techniques**. In *Proceedings of Interspeech* 2011:921–924.
9. Lin H, Stupakov A, Bilmes J: **Spoken keyword spotting via multi-lattice alignment**. In *Proceedings of Interspeech* 2008:2191–2194.
10. Parada C, Sethy A, Ramabhadran B: **Query-by-Example Spoken Term Detection for OOV terms**. In *Proceedings of ASRU* 2009:404–409.
11. Shen W, White CM, Hazen TJ: **A comparison of Query-by-Example Methods for Spoken Term Detection**. In *Proceedings of Interspeech* 2009:2143–2146.
12. Lin H, Stupakov A, Bilmes J: **Improving multi-lattice alignment based spoken keyword spotting**. In *Proceedings of ICASSP* 2009:4877–4880.
13. Barnard E, Davel M, van Heerden C, Kleynhans N, Bali K: **Phone recognition for Spoken Web Search**. In *Proceedings of MediaEval* 2011:5–6.
14. Buzo A, Cucu H, Safta M, Ionescu B, Burileanu C: **ARF@MediaEval 2012: A Romanian ASR-based Approach to Spoken Term Detection**. In *Proceedings of MediaEval* 2012:7–8.
15. Abad A, Astudillo RF: **The L2F Spoken Web Search system for Mediaeval 2012**. In *Proceedings of MediaEval* 2012:9–10.
16. Varona A, Penagarikano M, Rodríguez-Fuentes L, Bordel G, Diez M: **GTTS System for the Spoken Web Search Task at MediaEval 2012**. In *Proceedings of MediaEval* 2012:13–14.
17. Szöke I, Fapšo M, Veselý K: **BUT2012 Approaches for Spoken Web Search - MediaEval 2012**. In *Proceedings of MediaEval* 2012:15–16.
18. Hazen TJ, Shen W, White CM: **Query-by-Example spoken term detection using phonetic posteriorgram templates**. In *Proceedings of ASRU* 2009:421–426.
19. Zhang Y, Glass JR: **Unsupervised spoken keyword spotting via segmental DTW on Gaussian Posteriorgrams**. In *Proceedings of ASRU* 2009:398–403.
20. Chan C, Lee L: **Unsupervised spoken-term detection with spoken queries using segment-based dynamic time warping**. In *Proceedings of Interspeech* 2010:693–696.
21. Anguera X, Macrae R, Oliver N: **Partial sequence matching using an unbounded dynamic time warping algorithm**. In *Proceedings of ICASSP* 2010:3582–3585.
22. Anguera X: **Telefonica System for the Spoken Web Search Task at Mediaeval 2011**. In *Proceedings of MediaEval* 2011:3–4.

23. Muscariello A, Gravier G: **Irisa MediaEval 2011 Spoken Web Search System**. In *Proceedings of MediaEval* 2011:9–10.

24. Szöke I, Tejedor J, Fapšo M, Colás J: **BUT-HCTLab approaches for Spoken Web Search - MediaEval 2011**. In *Proceedings of MediaEval* 2011:11–12.

25. Wang H, Lee T: **CUHK System for the Spoken Web Search task at Mediaeval 2012**. In *Proceedings of MediaEval* 2012:3–4.

26. Joder C, Weninger F, Wöllmer M, Schuller B: **The TUM Cumulative DTW Approach for the Mediaeval 2012 Spoken Web Search task**. In *Proceedings of MediaEval* 2012:5–6.

27. Vavrek J, Pleva M, Juhár J: **TUKE MediaEval 2012: Spoken Web Search using DTW and Unsupervised SVM**. In *Proceedings of MediaEval* 2012:11–12.

28. Jansen A, Durme BV, Clark P: **The JHU-HLTCOE Spoken Web Search System for MediaEval 2012**. In *Proceedings of MediaEval* 2012:17–18.

29. Anguera X: **Telefonica Research System for the Spoken Web Search task at Mediaeval 2012**. In *Proceedings of MediaEval* 2012:19–20.

30. NIST: *The Ninth Text REtrieval Conference (TREC 9)* 2000, [http://trec.nist.gov].

31. NIST: *The spoken term detection (STD) 2006 evaluation plan*. 10, National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA 2006, [http://www.nist.gov/speech/tests/std].

32. Sakai T, Joho H: **Overview of NTCIR-9**. In *Proceedings of NTCIR-9 workshop* 2011:1–7.

33. Rajput N, Metze F: **Spoken Web Search**. In *Proceedings of MediaEval* 2011:1–2.

34. Metze F, Barnard E, Davel M, van Heerden C, Anguera X, Gravier G, Rajput N: **Spoken Web Search**. In *Proceedings of MediaEval* 2012:1–2.

35. Tokyo University of Technology: *Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access* 2013, [http://research.nii.ac.jp/ntcir/ntcir-10/].

36. NIST: *The OpenKWS13 Evaluation Plan*. 1, National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA 2013, [http://www.nist.gov/itl/iad/mig/openkws13.cfm].

37. Taras B, Nadeu C: **Audio segmentation of broadcast news in the Albayzin-2010 evaluation: overview, results, and discussion**. *EURASIP Journal on Audio, Speech, and Music Processing* 2011, **1**:1–10.

38. Zelenák M, Schulz H, Hernando J: **Speaker Diarization of Broadcast News in Albayzin 2010 Evaluation Campaign**. *EURASIP Journal on Audio, Speech, and Music Processing* 2012, **19**:1–9.

39. Rodríguez-Fuentes LJ, Penagarikano M, Varona A, Díez M, Bordel G: **The Albayzin 2010 Language Recognition Evaluation**. In *Proceedings of Interspeech* 2011:1529–1532.

40. Méndez F, Docío L, Arza M, Campillo F: **The Albayzin 2010 text-to-speech evaluation**. In *Proceedings of FALA* 2010:317–340.

41. Martin A, Doddington G, Kamm T, Ordowski M, Przybocki M: **The DET Curve In Assessment Of Detection Task Performance**. In *Proceedings of Eurospeech* 1997:1895–1898.

42. Iberspeech 2012: *"VII Jornadas en Tecnología del Habla" and "III Iberian SLTech Workshop"*[http://iberspeech2012.ii.uam.es/IberSPEECH2012_OnlineProceedings.pdf].

43. Anguera X: **Speaker independent discriminant feature extraction for acoustic pattern-matching**. In *Proceedings of ICASSP* 2012:485–488.

44. Anguera X, Ferrarons M: **Memory Efficient Subsequence DTW for Query-by-Example Spoken Term Detection**. In *Proceedings of ICME* 2013.

45. Schwarz P: **Phoneme recognition based on long temporal context**. *PhD thesis*, FIT, BUT, Brno, Czech Republic 2008.

46. Stolcke A: **SRILM - An Extensible Language Modeling Toolkit**. In *Proceedings of Interspeech* 2002:901–904.

47. Wang D, King S, Frankel J: **Stochastic Pronunciation Modelling for Out-of-Vocabulary Spoken Term Detection**. *IEEE Transactions on Audio, Speech, and Language Processing* 2011, **19**(4):688–698.

48. Wang D, Tejedor J, King S, Frankel J: **Term-dependent Confidence Normalization for Out-of-Vocabulary Spoken Term Detection**. *Journal of Computer Science and Technology* 2012, **27**(2):358–375.

49. Wang D, King S, Frankel J, Vipperla R, Evans N, Troncy R: **Direct posterior confidence for out-of-vocabulary spoken term detection**. *ACM Transactions on Information Systems* 2012, **30**(3):1–34.

50. Gómez J, Sanchis E, Castro-Bleda M: **Automatic speech segmentation based on acoustical clustering**. In *Proceedings of the joint IAPR International Conference on Structural, Syntactic, and Statistical Pattern Recognition* 2010:540–548.

51. Gómez J, Castro M: **Automatic segmentation of speech at the phonetic level**. In *Proceedings of the joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition* 2002:672–680.

52. Park A, Glass J: **Towards unsupervised pattern discovery in speech**. In *Proceedings of ASRU* 2005:53–58.

53. Young S, Evermann G, Gales M, Hain T, Kershaw D, Liu X, Moore G, Odell J, Ollason D, Povey D, Valtchev V, Woodland P: *The HTK Book*. Engineering Department, Cambridge University 2006.

54. Mertens T, Wallace R, Schneider D: **Cross-site Combination and Evaluation of Subword Spoken Term Detection Systems**. In *Proceedings of CBMI* 2011:61–66.

Figures

Figure 1 - DET curves of the QbE STD systems.

The broken black curve represents System 1, the red dot curve represents System 2a, the dark blue curve represents System 2b, the green curve represents System 3a, the solid black curve represents System 3b, the light blue curve represents System 4a, the red curve represents System 4b and the pink curve represents the Text-based STD system. Systems 3b and 4b represent late submissions. Systems 1-4 are on QbE STD and Text-based STD is on STD.

Figure 2 - Hit/FA performance of the QbE STD systems.

The blue column represents the hit performance and the brown column represents the FA performance. Both hit and FA values are represented as single values. Systems 3b and 4b represent late submissions. Systems 1-4 are on QbE STD and Text-based STD system is on STD.

Tables

Table 1 - Test queries along with the time length per query (in hundredth of seconds) and the number of occurrences in the test data.

Query (Time)	# occurrences	Query (Time)	# occurrences
acuerdo (29)	7	lenguaje (39)	6
análisis (37)	18	mecanismo (47)	7
aproximación (85)	7	metodología (81)	10
buscador (58)	7	motores (34)	6
cangrejo (49)	7	necesario (65)	6
castellano (57)	9	normalmente (32)	6
conjunto (49)	7	obtener (38)	9
conocimiento (49)	6	orientación (60)	6
desarrollo (46)	6	parecido (40)	6
detalle (28)	7	personas (54)	6
difícil (41)	12	perspectiva (49)	7
distintos (45)	21	porcentaje (66)	8
documentos (75)	7	precisamente (68)	6
efectivamente (29)	10	presentación (58)	15
ejemplo (55)	54	primera (29)	19
empezar (34)	7	principio (48)	9
encontrar (35)	19	propuesta (44)	19
entidades (67)	28	realidad (27)	10
estudiar (80)	7	reconocimiento (66)	6
evaluación (48)	15	recurso (52)	7
fuenlabrada (57)	15	referencia (47)	13
general (42)	11	resolver (42)	6
gracias (40)	13	segunda (52)	8
idiomas (29)	27	seguridad (35)	6
implicación (60)	31	siguiente (37)	11
importante (68)	19	simplemente (65)	8
incluso (41)	12	también (24)	93
información (56)	92	textual (59)	15
intentar (42)	13	trabajar (38)	39
interfaz (48)	10	utilizar (50)	15

Table 2 - Participants in the Query-by-Example Spoken Term Detection ALBAYZIN 2012 evaluation.

Team ID	Research Institution
TID	Telefonica Research, Barcelona, Spain
GTTS	University of the Basque Country, Bilbao, Spain
ELiRF	Politechnical University of Valencia, Spain
VivoLab	University of Zaragoza, Spain

Table 3 - Results of the QbE STD ALBAYZIN 2012 evaluation. Systems 1-4 are on QbE STD and Text-based STD system is on STD.

System ID	MTWV	ATWV	p(FA)	p(Miss)
System 1	0.0436	0.0122	0.00000	0.952
System 2a	0.0055	0.0031	0.00001	0.983
System 2b	0.0075	0.0047	0.00000	0.990
System 3a	0.0000	-2.1471	0.00000	1.000
System 3b (late submission)	0.0000	-0.0678	0.00000	1.000
System 4a	0.0000	-0.6416	0.00000	1.000
System 4b (late submission)	0.0238	0.0217	0.00007	0.909
Text-based STD system	0.1120	0.0638	0.00007	0.815

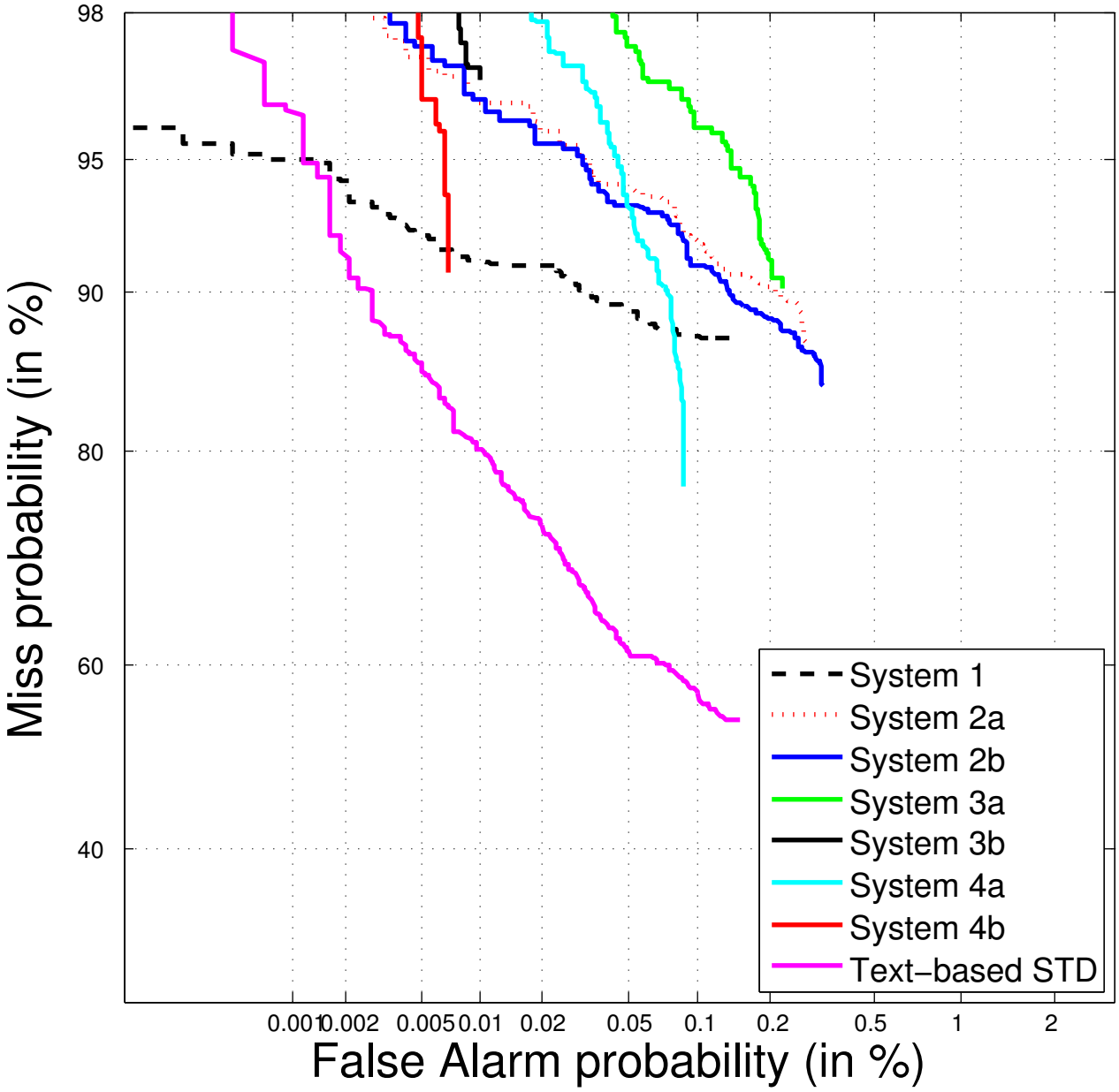
Miss probability (in %)
98
95
90
80
60
40
0.001
0.002
0.005
0.01
0.02
0.05
0.1
0.2
0.5
1
2
False Alarm probability (in %)
System 1
System 2a
System 2b
System 3a
System 3b
System 4a
System 4b
Text−based STD

Figure 1

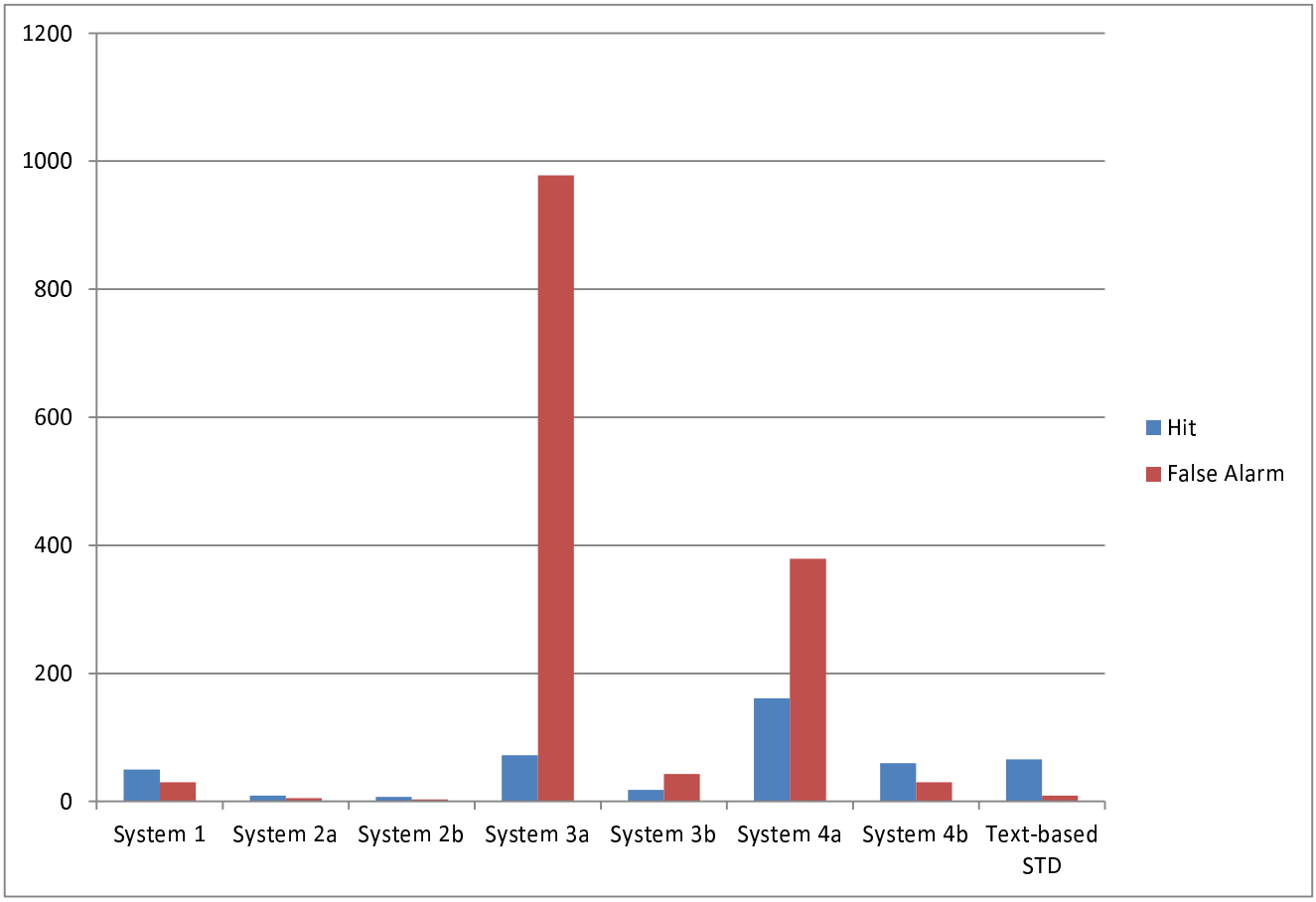
1200
1000
800
600
400
200
0
System 1
System 2a
System 2b
System 3a
System 3b
System 4a
System 4b
Text-based STD
Hit
False Alarm

Figure 2