Research Article
Journal of Research and Review in Science, 121–128
Volume 4, December 2017
DOI:10.36108/jrrslasu/7102/40(0171)

ORIGINAL RESEARCH

Information Retrieval Metrics for Speech Based Systems: A Systematic Review

Micheal Adenibuyan[1,2], Oluwatoyin Enikuomehin[2], Benjamin Aribisala[2]

[1]Department of Computer Science & Information Technology, College of Natural and Applied Sciences, Bells University of Technology, Ota, Ogun State, Nigeria

[2]Department of Computer Science, Faculty of Science, Lagos State University, Nigeria

Correspondence
Oluwatoyin Ayokunle Enikuomehin, Department of Computer Science, Faculty of Science, Lagos State University, Nigeria.
Email:toyinenikuomehin@gmail.com

Abstract:

Introduction: Information Retrieval (IR) allows identification of relevant information from connected repositories, however their performance have been of research interest leading to investigations in the modalities by which the accuracy of the retrievals are evaluated. Metrics such as Precision, Recall, F-score and Mean Average Precision (MAP) are commonly used for evaluating the performance of text-based IR system. These same metrics are also used for evaluating speech-based system while failing to realize the difference that could have occurred in the process of transcription, especially in the voice to text search, which is the most common speech-based search paradigm. This limits the performance and applications of speech-based system.

Aims: To review and identify the strengths and weaknesses of existing metrics for measuring the performances of speech based.

Materials and Methods: A total of 179 articles were retrieved from Google Scholar repository and carefully examined. Only 25 articles were selected for analysis in this study after applying our predefined inclusion and exclusion criteria.

Results: Results show that MAP is the most frequently used metric for assessing the performance of speech-based IR system and the values ranged from 0.4191 to 0.620. Results also show there is an inverse relationship between IR performance and transcription error.

Conclusion: This suggests that transcription error has a significantly negative effect on retrieval accuracy. Hence there is a need to develop a specialised speech-based metric for measuring IR performance.

Keywords: Information Retrieval, Performance Metric, Spoken Query.

All co-authors agreed to have their names listed as authors.

1. INTRODUCTION

The advent of internet and the continuous adoption of same by many has led to increased availability of data scattered around millions of repositories world over. This growth exists across dynamic platforms. A consequence of this is improved information sharing, dissemination and overload. An important task is the identification of information that meets user's specific need from the repositories. The field of Information Retrieval (IR) offers tools for achieving these tasks. IR requires querying of the repositories, so the method of querying is a crucial task. Poorly formulated query will yield results that will not satisfy the user's information need. Lots of research works have been done in the area of text query to text document on the challenges of retrieving the precise document relevant to the user's intention or need from large multidimensional repositories [1, 2].

With advances in technology and increasing multimedia data, it became relevant to retrieve spoken documents or text documents using spoken or voice query, as it is natural and easy for users to express their information need via voice, because it eliminates the challenge of translating thought to textual query. A number of algorithms (such as vector space model, probabilistic model) have been developed for retrieving spoken words and the performance of the algorithms are very important in order to meet the needs of the users interested in retrieving spoken words. This has attracted some research works and has led to the development of IR performance metrics. IR evaluation involves conducting an experiment using same parameters on different retrieval algorithms, indexing techniques and ranking them.

Some of the commonly used metrics are precision, recall, mean average precision (MAP) and mean reciprocal rank (MRR). It has been observed that the performance metrics developed for evaluating text based IR systems are also used for speech IR based systems[12]. This limits the performance and acceptability of speech-based IR systems because the metrics developed for text-based IR systems do not factor consider important factors like agglutination in languages, noise and data loss at the level of transcription prior to retrieval.

The focus of this systematic review was to review and identify the strengths and weaknesses of existing metrics for measuring the performances of speech-based IR systems. This could form the basis for developing a robust and novel performance metric developed specifically for speech-based IR systems.

2. METHODS

This study adopts a systematic review method, the steps involved are as follows: study design, search strategy and information sources; study selection and data collection process, and quality assessment, data extraction and synthesis[13].

2.1 Study Design, Search Strategy and Information Sources

The search strategy is composed of the following processes: (a) Establish search terms by using keywords that aligns with the purpose of this research, (b) Identify alternate words that have same meaning as the major keywords used, (c) Identify some studies relevant to our review, (d) Establish exclusion criteria to filter studies that are not of relevant to this study, (e) Use boolean operators to develop the needed search term[13].

For this research, we used Google scholar repository to identify the articles related to the performance metrics of speech-based IR systems.

2.1.1 Study Selection and Data Collection Process

This systematic review of articles was conducted using the Preferred Reporting Item for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [14]. Figure 2 shows the study selection process in a PRISMA flow diagram.

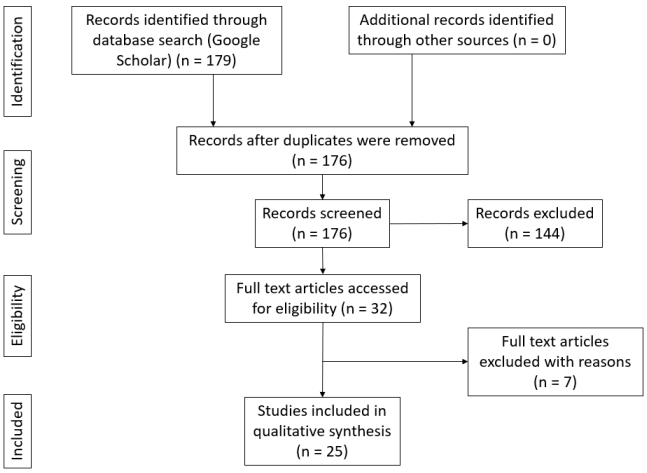

Figure 1: Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram of included studies.

2.1.2 Inclusion Criteria

Studies on IR systems for spoken queries and their performance metrics were identified and reviewed. For the purpose of this systematic review, our studies were restricted to spoken information retrieval systems. Our inclusion criteria consist of terms that include: Spoken Queries, Evaluation Metrics, and Performance. The following search terms were used to search through Google scholar for related literature; Speech corpus, text corpus, retrieval model, performance evaluation, rank, voice query, English, metric, transcription and baseline.

3. RESULTS

3.1 LITERATURE SEARCH RESULTS

The basic search returned 179 articles out of which 3 were duplicates. After the removal of duplicate articles, the remaining 176 were screened using titles and abstracts leaving use with only 25 articles relevant for the review.

10 studies were on Mean Average Precision (MAP), 1 study applied Average Precision and R-Precision, 4 applied a combination of Precision, Recall and F-Measure, 1 combined Average Inverse Rank (AIR) and MAP, 1 combined Average Precision and Recall, 2 combined Precision and Recall, 1 adopted F-measure and MAP, 1 applied the Generalised Average Precision (GAP) and MAP, 1 applied Precision Recall Curve, F-Measure and MAP, 2 used Precision, Recall and MAP while 1 combined Normalized Discounted Cumulative Gain (NCDG), Mean Reciprocal Rank (MRR) and MAP. Table 1 present the characteristics of included studies.

Table 1. Characteristics of Included Studies

Characteristics	**Number of Papers = 25**
Median year of Publication	2005 (1998-2015)
Country of Publication:	
Canada	2
China	6
Finland	1
Ireland	1
Germany	1
UK	3
United State of America (USA)	4
Japan	2
Singapore	1
Thailand and Indonesia	1
Language	
English	25

3.2 MAIN STUDY RESULTS

3.2.1 Table 2. Deduction from some included study

S/No	Title of Study	Authors	Year	Methodology	Strength	Weakness	Contribution
1.	The THISL Spoken Document Retrieval System	Renals, Steve, and Dave Abberley.	1998	ABBOT large vocabulary continous speech recognition (LVCSR) was used to transcribe audio documents. This was then applied to the IR engine	The performance of the spoken retrieval system with respect to MRR gives 0.75 which is good in the face of WER of 40% or less.	Performance begins to fall when the WER is above 40%	Development of a spoken document retrieval system and measure the performance on audio documents.
2.	1998 TREC-7 Spoken Document Retrieval Track 1998 TREC-7 Spoken Document Retrieval Track Overview and Results	JS Garofolo, EM Voorhees, CGP Auzanne	1999	Four retrieval conditions were made for experiment control, for a corpus of 100 hours Broadcast News.	MAP is the preferred TREC Metric for Speech based IR. There is a linear relationship between retrieval accuracy and the speech transcription.		There is a linear relationship between recognition word error rate and retrieval performance.
3.	Subword-based approaches for spoken document retrieval	Ng, Kenney, and Victor W. Zue	2000	Audio data of radio broadcast was orthographically transcribed by professionals then implemented on an IR engine based on Vector Space Model (VSM).	Retrieval performance measured in MAP improves for difference subword unit.	MAP is a single figure metric.Hence, it hides some performance information	Modifying spoken queries via relevance feedback and document expansion using N-best recognition hypotheses improves IR performance.
4.	Document Expansion using a Side Collection for Monolingual and Cross-Language Spoken Document Retrieval	Yuk-Chi Li, Helen M. Meng	2003	VSM was applied on a local television news broadcast and evaluated using the average inverse rank (AIR). Document expansion was applied to the indexing.	Average Inverse Rank (AIR) which is same as MRR was used to measure retrieval performance. The retrieval performance improved from 0.479 to 0.747 using document expansion		Monolingual and cross-language tasks proves that document expansion improves IR performance.
5.	Syllable-based Language Models in Speech Recognition for English Spoken Document Retrieval	Christian Schrumpf, Martha Larson, and Stefan Eickeler	2005	Radio news with no noise or background sound effects was applied to Hidden Markov Model (HMM) for speech recognition, then a fuzzy word match for retrieval.	IR performs better using syllable recognizer transcription.		Syllable indexing is an efficient suppliment to word features for spoken document retrieval (SDR).
6.	Query-Free News Search	Henzinger, Monika, Bay-Wei Chang, Brian Milch, and Sergey Brin	2005	A query generation algorithm was applied on a news broadcast corpus. Boolean retrieval technique was used.	Precision is improved when single word term is considered.	Pooled recall was used as an absolute measure of a system recall performance. Transcription is not considered in the experiment.	Filtering with similarity to caption offers large improvement in precision.
7.	The Influence of Word Detection	Rispoli, Renato,	2006	Audio recording of an english	Retrieval performance is	AP provides a binary relevance.	Speech based IR has the potential

	Variability on IR Performance in Automatic Audio Indexing of Course Lectures	Richard C. Rose, and Jon Arrowood.		speaking class with the background noise of a typical classroom was used as the corpus. This was then transcribed manually.	measured as Average Precision (AP). AP combines relevance ranking and recall.	IR performance degrades when keyword detection is used.	to improve user productivity.
8.	Integrating recognition and retrieval with user feedback: A new framework for spoken term detection	Lee, Hung-yi, and Lin-shan Lee.	2010	Segments of speech is transcribed into lattice, then computed using the posterior probability.	MAP is the performance metric used for its good stability.		A feedback framework was proposed that improves retrieval performance.
9.	Recent developments in spoken term detection: a survey.	Mandal, Anupam, KR Prasanna Kumar, and PabitraMitra.	2014	LVCSR was used for transcription. Position specific posterior lattice (PSPL) to get a posterior probability of the word lattice, then the retrieval engine.	Performance was measured using MAP. 3-gram phone index gives the high MAP (best performance), as well as an overlapping subword unit.	When none overlapping subwords are used performance (MAP) is poor.	Supervised approach is more effective compared to the unsupervised approaches of spoken term detection in IR.
10.	Using Zero-Resource Spoken Term Discovery for Ranked Retrieval.	White, Jerome, Douglas W. Oard, Aren Jansen, Jiaul H. Paik, and Rashmi Sankepally.	2015	Spoken query was applied on a corpus of AvajOtalo's recorded speech.	It compared three IR metrics (MAP, MRR, NDCG). NDCG was the best performing metric when compared.		Zero-Resource approach is a viable method to retrieve relevant response.

3.2.2 Types of IR Evaluation Measures

3.2.2.1 Precision and Recall

Precision is the positive predictive value retrieved, while Recall is the sensitivity of the system to retrieve relevant items from the collection. Precision and Recall are based on the idea of relevance. How many relevant items are retrieved (Recall), how many retrieved items are relevant (precision). This account for 8% of publications reviewed.

3.2.2.2 Mean Average Precision (MAP)

This is the most used evaluation metric for an IR System performance. It account for 60% of publication studied for this review.[1, 2, 6, 10, 14-22] all applied MAP, it ranks relevant documents higher than irrelevant documents. MAP result range from 0.4191 to 0.620 with relative improvement in retrieval performance by 13.8%.

3.2.2.3 Average Inverse Rank (AIR) / Mean Reciprocal Rank (MRR)

AIR and MRR are used interchangeably; it is the inverse of the ranked retrieval result. [2, 6, 7] applied AIR/MRR to measure retrieval performance and achieved as much as 14% improvement in IR performance with range between 0.747 to 0.826. MRR account for 12% of publications reviewed.

3.2.2.1 Normalized Discounted Cumulative Gain (NDCG)

4% of publications studied applied the NDCG metric to measure IR performance, specifically [6] applied this and result range from 0.089 to 0.284.

Table 3. Performance Metric included in this study

Retrieval Metrics	Number of Studies	Studies
MAP	10	[1, 10, 13, 15-18, 20]
MRR	1	[7]
Precision, Recall, F-Measure	4	[23]
Precision, Recall	2	[2]
MRR, MAP	1	[2]
Average Precision, Recall	1	[9]
F-Measure, MAP	1	[19]
Precision Recall – Curve, F-Measure, MAP	1	[14]
Precision Recall, MAP	2	[22]
Average Precision, MAP	1	[21]
NDCG, MRR, MAP	1	[6]

4. RESULTS AND DISCUSSION

4.1 GENERAL DEDUCTION BASED ON STUDY

The aim of this research to review existing metrics for measuring the performance of speech-based IR system as well as to identify strengths and weaknesses of these metrics. Result shows that metrics used to measure text-based IR system performance are exactly the same applied in speech-based IR system. Though precision and recall are the traditional metrics but they do not satisfy in evaluating the performance when results need to be ranked as they are binary relevance measures. MAP is the most used metric to measure speech-based IR system performance though it hides some performance information since it is a single numeric metric.

In the face of transcription errors, retrieval performance varies. Tonal/language agglutination, background noise when passing the voice query is not part of the properties used to determine the performance of a speech-based IR system. We say, transcription is a challenge in speech-based IR performance. An inverse relationship is seen between transcription error and retrieval performance.

Existing metrics used to measure performance of speech-based IR system does not consider transcription errors as a property that have significant impact on the retrieval performance.

The strength of this study is that systematic approach was used for identifying all previous work on performance metric for information retrieval. This allows us to use a set of carefully chosen search terms in the study. Also, a standard approach for determining the inclusion and exclusion criteria was used. This enabled us to retain and to review only the relevant articles.

Another strength of this study is the use of google scholar repository. The benefit of this is that google scholar is connected to all other repositories, thereby ensuring that all the relevant articles were identified.

5. CONCLUSION

In this study, a systematic review of speech-based IR system was conducted. We found that most research work concurs to the fact that same metric used for text-based IR systems are the exact metrics used for speech-based IR system and these metrics are not good enough for measuring speech IR as they are sensitive to transcription errors. We hereby suggest that more research be done in this field to enable the development of new IR metrics or modification of existing metrics for speech IR.

ACKNOWLEDGEMENTS

We will like to show our gratitude to the faculty of science of Lagos State University, specifically the department of Computer Science for the opportunity to present our systematic review.

COMPETING INTERESTS

The authors declare that there are no competing interests.

AUTHORS' CONTRIBUTIONS

Micheal Adenibuyan conducted literature search, analyzed them and wrote the first draft of the manuscript. Oluwatoyin Enikuomehin and Benjamin Aribisala designed the study, revised the manuscript, wrote the final copy of the manuscript and supervised the study.

REFERENCES

1. Ng, K. and V.W. Zue, *Subword-based approaches for spoken document retrieval.* Speech Communication, 2000. **32**(3): p. 157-186.

2. Li, Y.-C. and H.M. Meng. *Document expansion using a side collection for monolingual and cross-language spoken document retrieval*. in *ISCA Workshop on Multilingual Spoken Document Retrieval*. 2003.

3. Kupiec, J., D. Kimber, and V. Balasubramanian. *Speech-based retrieval using semantic co-occurrence filtering*. in *Proceedings of the workshop on Human Language Technology*. 1994. Association for Computational Linguistics.

4. Syu, I. and S.-D. Lang, *Adapting a diagnostic problem-solving model to information retrieval.* Information processing & management, 2000. **36**(2): p. 313-330.

5. Benahmed, Y. and S.-a. Selouani. *Robust self-training system for spoken query information retrieval using pitch range variations*. in *Electrical and Computer Engineering, 2006. CCECE'06. Canadian Conference on*. 2006. IEEE.

6. Saracevic, T. *The stratified model of information retrieval interaction: Extension and applications*. in *Proceedings of the Annual Meeting-American Society for Information Science*. 1997. Learned Information (Europe) Ltd.

7. Sutcliffe, A. and M. Ennis, *Towards a cognitive theory of information retrieval.* Interacting with computers, 1998. **10**(3): p. 321-351.

8. Zhai, C. and J. Lafferty. *A study of smoothing methods for language models applied to ad hoc information retrieval*. in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. 2001. ACM.

9. Robertson, S., *Evaluation in information retrieval*, in *Lectures on information retrieval*. 2000, Springer. p. 81-92.

10. Garofolo, J.S., et al. *1998 trec-7 spoken document retrieval track overview and results*. in *Broadcast News Workshop*. 1999.

11. Crestani, F., *An experimental study of the effects of word recognition errors in spoken queries on the effectiveness of an information retrieval system.* 1999.

12. Moreno-Daniel, A., et al. *Spoken query processing for information retrieval*. in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*. 2007. IEEE.

13. Khan, S.U., M. Niazi, and R. Ahmad, *Systematic Literature Review Protocol for Software Outsourcing Vendors Readiness Model (SOVRM).* School of Computing and Maths, Keele University, UK TR/08-01, 2008.

14. Shamseer, L., et al., *Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation.* Bmj, 2015. **349**: p. g7647.

15. Jones, G.J., et al., *Examining the contributions of automatic speech transcriptions and metadata sources for searching spontaneous conversational speech.* 2007.

16. Lin, S.-H., B. Chen, and E.-E. Jan. *Improving the informativeness of verbose queries using summarization techniques for spoken document retrieval*. in *Chinese Spoken Language Processing (ISCSLP), 2010 7th International Symposium on*. 2010. IEEE.

17. Schrumpf, C., M. Larson, and S. Eickeler. *Syllable-based language models in speech recognition for English spoken document retrieval*. in *Proc. of the 7th International Workshop of the EU Network of Excellence DELOS on AVIVDiLib, Cortona, Italy*. 2005.

18. Chia, T.K., et al. *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. ACM.

19. Shokouhi, M., U. Ozertem, and N. Craswell. *Did You Say U2 or YouTube?: Inferring Implicit Transcripts from Voice Search Logs*. in *Proceedings of the 25th International Conference on World Wide Web*. 2016. International World Wide Web Conferences Steering Committee.

20. Vergyri, D., et al. *The SRI/OGI 2006 spoken term detection system*. in *Interspeech*. 2007.

21. Moher, D., et al., *Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement.* PLoS medicine, 2009. **6**(7): p. e1000097.

22. Lee, H.-y. and L.-s. Lee. *Integrating recognition and retrieval with user feedback: A new framework for spoken term detection*. in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. 2010. IEEE.

23. Zhai, C. and J. Lafferty. *A study of smoothing methods for language models applied to ad hoc information retrieval*. in *ACM SIGIR Forum*. 2017. ACM.

24. Kelly, D. and C.R. Sugimoto, *A systematic review of interactive information retrieval evaluation studies, 1967–2006.* Journal of the Association for Information Science and Technology, 2013. **64**(4): p. 745-770.