IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Text-Based Spoken Term Detection on the QUESST 2014 Multilingual Benchmark: A Comparative Evaluation Against Speech-to-Speech Acoustic Matching Methods

Vrushali Ravindra Deshpande [1] , Sushil Venkatesh Kulkarni[2] , Suryakant B Kendre[3]

[1] M.Tech. 4th semester Student, Department of Computer Science And Information Technology, MBES College of Engineering Ambajogai, India

[2] Professor and Head , Department of Computer Science And Engineering, MBES College of Engineering Ambajogai, India

[3] Assistant Professor and Guide , Department of Computer Science And Engineering, MBES College of Engineering Ambajogai, India

Abstract—This paper presents a complete end-to-end text-based Spoken Term Detection (STD) system evaluated on the QUESST 2014 multilingual benchmark, covering six low-resource European languages: Albanian, Basque, Czech, non-native English (NNEnglish), Romanian, and Slovak. The proposed system employs the HappyScribe cloud-based Automatic Speech Recognition (ASR) service for audio transcription, followed by the Terrier Information Retrieval (IR) framework with the Hiemstra language model (Hiemstra_LM) for inverted index construction and probabilistic ranked retrieval. Detection performance is assessed using the official QUESST 2014 metrics: minimum Normalised Cross-Entropy (minCNXE, lower is better) and Maximum Term-Weighted Value (MTWV, higher is better), across three query types: Type 1 (exact match), Type 2 (lexical variation), and Type 3 (multi-word expression). The text-based system is systematically compared against four speech-to-speech acoustic matching baselines: Dynamic Time Warping (DTW), Convolutional Neural Network (CNN) matching, End-to-End neural matching, and the Water Chicken Swarm Optimisation-based Deep Segmental Neural Network (WCSO-DSNN). The text-based system achieves a best multilingual minCNXE of 0.2489 and MTWV of 0.8337 for Type 1 evaluation queries, substantially outperforming all four acoustic baselines. Language-specific analysis establishes Romanian (minCNXE=0.1012, MTWV=0.9123) and Slovak (minCNXE=0.1115, MTWV=0.9123) as the highest-performing languages, while NNEnglish (best minCNXE=0.3983) is the most challenging due to multi-speaker accent variability. The findings demonstrate that text-based STD is significantly more mature than speech-to-speech acoustic matching for settings where reliable ASR is available, while acoustic methods retain a practical advantage in true zero-resource environments.

Keywords—Spoken Term Detection, Text-Based STD, QUESST 2014, Hiemstra Language Model, Terrier IR, minCNXE, MTWV, Multilingual Speech Retrieval, Acoustic Matching, Zero-Resource STD.

I. INTRODUCTION

The rapid proliferation of digital spoken content — spanning broadcast archives, parliamentary proceedings, online lectures, call-centre recordings, and surveillance audio — has created an urgent need for scalable, language-agnostic systems capable of detecting specific spoken terms within large audio repositories. Spoken Term Detection (STD) addresses this need by locating all occurrences of a user-specified query within a speech archive and returning a ranked list of matching segments with associated confidence scores [1].

Two principal paradigms dominate the STD landscape. Text-based STD transcribes audio using an Automatic Speech Recognition (ASR) system and indexes the resulting transcripts using information retrieval (IR) techniques. This approach achieves high accuracy when reliable ASR is available but is inherently language-dependent. Speech-to-speech acoustic matching bypasses transcription entirely, comparing the spoken query directly against archive segments using acoustic similarity measures. This zero-resource paradigm is language-independent and applicable where ASR systems are unavailable [2].

Despite the practical importance of both paradigms, their comparative maturity has not been systematically quantified across a standardised multilingual benchmark with multiple query complexity levels. Most existing studies evaluate only one paradigm, limiting cross-paradigm assessment. Furthermore, the interaction between ASR transcript quality, morphological complexity, recording conditions, and detection performance has not been comprehensively analysed across typologically diverse languages.

This paper directly addresses these gaps by implementing a complete text-based STD pipeline and evaluating it on the QUESST 2014 multilingual benchmark [3], [4], followed by systematic comparison against four speech-to-speech acoustic matching baselines. The QUESST 2014 benchmark provides a standardised zero-resource evaluation framework covering six European languages under identical experimental conditions, making it the most comprehensive available platform for cross-paradigm maturity assessment.

The principal contributions of this paper are: (i) a complete reproducible text-based STD pipeline using HappyScribe ASR [5] and the Terrier IR framework [6] with the Hiemstra language model [7]; (ii) comprehensive evaluation across six languages, three query types, and both development and evaluation query sets, producing 32 detailed performance tables; (iii) systematic comparative analysis against four acoustic baselines — DTW [8], CNN matching [9], End-to-End neural matching [10], and WCSO-DSNN [21] — using identical QUESST 2014 conditions; and (iv) language-specific analysis establishing ASR transcript quality as the primary determinant of text-based STD performance, with morphological complexity as the primary secondary factor.

II. RELATED WORK

A. Text-Based STD

Text-based STD transforms spoken content search into a classical IR problem. Early systems employed LVCSR-generated word lattices indexed for rapid keyword lookup [11]. Norouzian [12] demonstrated that two-stage retrieval combined with acoustic verification improves out-of-vocabulary (OOV) recall. Singh et al. [13] proposed a TF-IDF-based QbE-STD approach achieving improved ATWV, recall, and MAP on QUESST 2014. Yusuf and Saraçlar [14] demonstrated that written term detection improves spoken term detection, confirming text-based retrieval advantages for exact-match queries. Tejedor and Toledano [15] showed that Whisper-based transcription substantially improves text-based STD for the ALBAYZIN evaluation. The Terrier IR framework [6] has been widely adopted for text-based STD evaluations, with the Hiemstra language model providing probabilistic query-document similarity based on background-smoothed interpolation [7].

B. Speech-to-Speech Acoustic Matching

Zhang and Glass [8] introduced DTW-based matching on Gaussian posteriorgrams, establishing the fundamental zero-resource baseline. Ram et al. [9] extended this to CNN-based matching using convolutional layers for local acoustic pattern modelling. The same group subsequently proposed an end-to-end neural approach jointly learning query and document embeddings in a shared latent space [10]. Kamper et al. [16] introduced unsupervised word segmentation using acoustic word embeddings. Recent self-supervised representations — Wav2Vec 2.0 [17] and HuBERT [18] — have substantially improved zero-resource acoustic matching. The WCSO-based Deep Segmental Neural Network [19] combines Water Chicken Swarm Optimisation with segmental modelling through three parallel sub-networks (Left Context DNN, Segment Representative DNN, Right Context DNN).

C. Research Gap

The literature lacks systematic comparative evaluation of text-based and speech-to-speech STD under identical multilingual benchmark conditions across multiple query types and primary metrics (minCNXE and MTWV simultaneously). This paper fills this gap.

III. THE QUESST 2014 BENCHMARK

A. Dataset Description

The QUESST 2014 (Query by Example Search on Speech Task 2014) dataset [3], [4] was introduced as part of the MediaEval 2014 Benchmark Evaluation campaign to evaluate QbE-STD systems in a zero-resource setting. It comprises 12,492 spoken audio documents across six European languages with a total duration of 1,385 minutes. Table I presents the complete dataset statistics.

TABLE I: QUESST 2014 Dataset Statistics

Language	Duration (min)	Documents	Dev Queries	Eval Queries	Speech Type
Albanian	127	968	50	50	Read
Basque	192	1,841	70	70	Broadcast
Czech	237	2,653	100	100	Conversational
NNEnglish	273	2,438	138	138	TEDx Multi-Accent
Romanian	244	2,272	100	100	Read
Slovak	312	2,320	102	97	Parliamentary
TOTAL	1,385	12,492	560	555	Mixed

B. Query Types

Three query types of increasing complexity are defined. Type 1 (exact match): the query term appears verbatim in the archive. Type 2 (lexical variation): the query may appear with morphological inflections or pronunciation variants. Type 3 (multi-word expression): all constituent words of the query must be detected, possibly with reordered constituents or intervening filler content. The query distribution across development and evaluation sets is 307/190/155 and 307/179/156 for Types 1/2/3 respectively.

C. Evaluation Protocol and Metrics

Systems submit a stdlist.xml file containing, for each (query, document) trial, a continuous similarity score and a binary YES/NO detection decision. The primary metric is minimum Normalised Cross-Entropy (minCNXE, lower is better), defined as CNXE = C_{xe} / C_{prior} where C_{xe} is the empirical cross-entropy of the detection scores and C_{prior} is the prior entropy of a trivial reference system. minCNXE is the best CNXE achievable through optimal linear score calibration and measures score-level discrimination independently of any decision threshold. The secondary metric is Maximum Term-Weighted Value (MTWV, higher is better): MTWV = $1 - (1/|Q|) \times \Sigma[P_{Miss}(q,\theta) + \beta \times P_{FA}(q,\theta)]$ at the

optimal threshold θ^*, where $\beta = (C_{FA}/C_{Miss}) \times (1-P_{target})/P_{target}$. For QUESST 2014: $P_{target}=0.0008$, $C_{FA}=1$, $C_{Miss}=100$, giving $\beta=12.49$. MTWV=1 indicates perfect detection; MTWV=0 indicates a trivial system with no detections. Evaluation is performed separately for development and evaluation query sets, and for each query type per language.

IV. PROPOSED TEXT-BASED STD SYSTEM

A. System Architecture

The proposed system shown in figure 1 follows a sequential six-stage pipeline transforming raw speech audio into ranked detection results. The stages are: (1) ASR Transcription, (2) TREC Format Conversion, (3) Terrier Inverted Index Construction, (4) Hiemstra_LM Retrieval, (5) Decision Thresholding and stdlist.xml Generation, (6) QUESST 2014 Evaluation.

B. Stage 1: ASR Transcription using HappyScribe

All 12,492 spoken audio documents and 1,115 spoken queries (development + evaluation) are transcribed using the HappyScribe cloud-based ASR service [5]. HappyScribe accepts audio uploads and returns word-level transcripts. The same ASR service is applied to both document and query tracks, ensuring internal consistency. The quality of ASR transcription directly bounds the upper limit of text-based STD performance [11]. Language-specific ASR accuracy varies substantially: languages with regular phonology and formal-register speech (Romanian, Slovak) yield higher-quality transcripts than those with complex morphology and conversational conditions (Czech) or multi-accent heterogeneity (NNEnglish).

C. Stage 2: TREC Format Conversion

ASR transcripts are converted into five TREC-compatible formats for the Terrier IR framework:

(i) TrecDoc — each archive document formatted with <DOCNO>, <HEAD>, and <BODY> tags;

(ii) TrecQuery — each spoken query formatted as a TREC topic with <title>, <desc>, and <narr> fields all carrying identical ASR-transcribed text;

(iii) TrecQrel — ground-truth relevance annotations mapping query IDs to relevant document IDs;

(iv) TrecRes — evaluation output storage; and

(v) TrecRun — ranked retrieval results formatted as: qid Q0 docno rank score tag, subsequently converted to stdlist.xml.

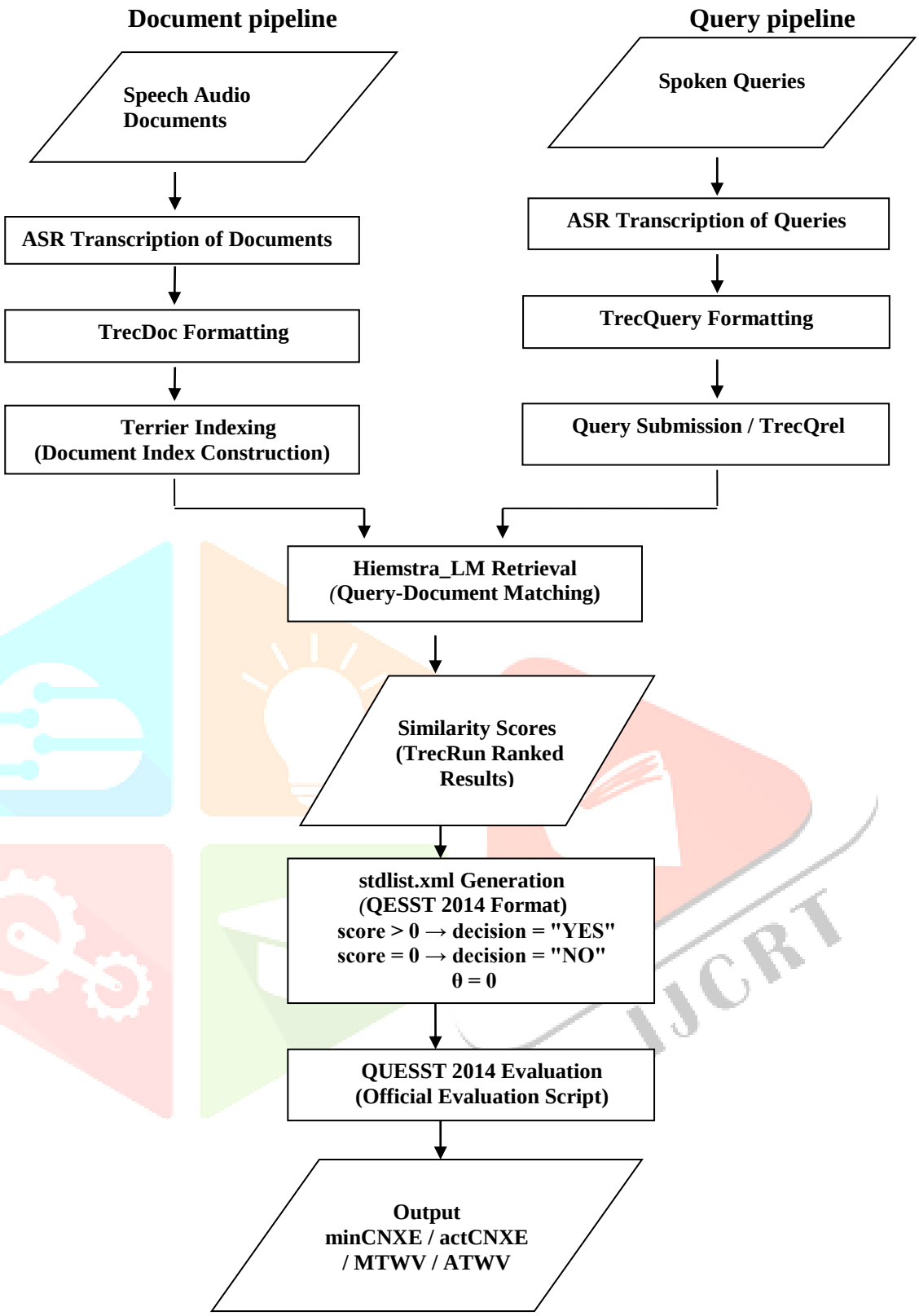

Figure 1: Workflow of Text-Based STD System.

D. Stage 3: Terrier Inverted Index Construction

The Terrier IR framework [6] constructs an inverted index from all TrecDoc-formatted document transcripts. The indexing pipeline comprises: document collection specification via TRECCollection class, tokenisation of transcript text, stop-word removal, stemming to morphological roots, posting accumulation of term-document statistics, and index serialisation. The resulting inverted index maps each unique term to a postings list recording the documents in which it appears along with term frequency statistics. The index is constructed from all 12,492 QUESST 2014 document transcripts.

E. Stage 4: Hiemstra_LM Retrieval

The Hiemstra language model [7] computes similarity scores between each query transcript and all indexed document transcripts. The scoring function is:

$$\text{Score}(Q, D) = \Sigma_{\{t \in Q\}} \log[\lambda \cdot P(t|D) + (1-\lambda) \cdot P(t|C)]$$

where $P(t|D) = f(t,D) / |D|$ is the term probability in document D normalised by document length, $P(t|C) = f(t,C) / |C|$ is the collection background model probability, and λ is the interpolation parameter (set to $\lambda=0.25$ in all experiments). Higher scores indicate stronger query-document term overlap. The probabilistic background smoothing makes the model robust to variable-length transcript segments and partial keyword matches.

F. Stage 5: Decision Thresholding and stdlist.xml Generation

A fixed decision threshold of $\theta=0$ is applied to Hiemstra_LM similarity scores, corresponding to the natural decision boundary of the model: a score of exactly zero indicates no query term appears in the indexed document. Both continuous scores (for minCNXE computation) and binary decisions (for actTWV computation) are included in the stdlist.xml output, enabling the official QUESST 2014 evaluation scripts to compute all four metrics: minCNXE, actCNXE, MTWV, and actTWV.

V. EXPERIMENTAL RESULTS

A. Multilingual Performance — Development Queries

Table II presents the text-based STD performance aggregated across all six languages using development queries. Type 1 exact-match queries achieve the best performance (minCNXE=0.2538, MTWV=0.7898), confirming that verbatim term matching is the most reliable detection scenario. Type 2 lexical variation is the weakest condition (minCNXE=0.6639, MTWV=0.5253), with a PMiss of approximately 44.60% at the optimal threshold, reflecting the challenge of morphological inflection for transcript-based retrieval.

TABLE II: Multilingual Text-Based STD — Development Queries

Measure	All Types	Type 1	Type 2	Type 3
minCNXE	0.4988	0.2538 ★	0.6639	0.5224
actCNXE	1.3465	0.5917	1.9762	1.4169
MTWV	0.6267	0.7898 ★	0.5253	0.4759
actTWV	0.5677	0.6893	0.4960	0.4345

B. Multilingual Performance — Evaluation Queries

Table III presents the evaluation query results. Type 1 again achieves the best performance (minCNXE=0.2489, MTWV=0.8337), with a P_{Miss} of only 14.72% and P_{FA} of 0.1530% at the optimal operating point, confirming that approximately 85% of all genuine exact-match target occurrences are correctly retrieved across all six languages. The actTWV of 0.7734 is close to MTWV, confirming good score calibration. Type 2 remains the weakest condition (minCNXE=0.6301, MTWV=0.5236), with actCNXE of 1.8427 substantially above minCNXE, indicating severely poor score calibration for lexically varied queries.

TABLE III: Multilingual Text-Based STD — Evaluation Queries

Measure	All Types	Type 1	Type 2	Type 3
minCNXE	0.4773	0.2489 ★	0.6301	0.5205
actCNXE	1.2622	0.5613	1.8427	1.4136
MTWV	0.6667	0.8337 ★	0.5236	0.5431
actTWV	0.6311	0.7734	0.5101	0.5003

C. Language-Specific Performance — Evaluation Queries

Table IV presents language-specific Type 1 evaluation results. Romanian and Slovak jointly achieve the best performance with minCNXE of 0.1012 and 0.1115 respectively, and MTWV of 0.9123 for both, corresponding to a PMiss below 9% at the optimal threshold. Czech achieves the best All Types MTWV (0.8056) confirming strong performance even under conversational conditions. Albanian exhibits the most unusual pattern — Type 2 (MTWV=0.7891) outperforms Type 1 (MTWV=0.7290) — confirming that inflected Albanian forms are reliably captured in the indexed transcripts. Basque shows extreme polarisation: exceptional Type 1 (minCNXE=0.2190, MTWV=0.8172) but catastrophic Type 2 (minCNXE=0.7697) due to agglutinative morphology. NNEnglish is uniformly the weakest language, with no condition achieving MTWV above 0.67.

TABLE IV: Language-Specific Type 1 Evaluation Performance — Text-Based STD

Language	minCNXE	MTWV	PMiss (%)	PFA (%)	Performance Tier
Romanian	0.1012	0.9123	8.43	0.03	**Best**
Slovak	0.1115	0.9123	6.69	0.17	**Best**
Czech	0.1751	0.8792	7.92	0.33	**Best**
Basque	0.2190	0.8172	17.70	0.05	Good
Albanian	0.3257	0.7290	26.71	0.03	Good
NNEnglish	0.4226	0.6491	34.89	0.02	Weakest

D. Cross-Language Analysis

Three consistent findings emerge from the language-specific results. First, ASR transcript quality is the primary determinant of text-based STD performance. Languages with controlled recording conditions and phonologically regular structures (Romanian, Slovak, Czech) consistently achieve the best minCNXE values, while NNEnglish — where multi-speaker accent variability degrades ASR quality — achieves the weakest results regardless of query type. Second, morphological complexity determines the size of the Type 1-to-Type 2 performance drop. Slovak shows the largest collapse (MTWV 0.9123 → 0.3995, drop of 0.5128) and Basque the second largest (0.8172 → 0.4365, drop of 0.3807), while Albanian shows almost no decline (Type 2 actually exceeds Type 1). Third, multi-accent speaker variability is the only challenge the text-based system cannot overcome — NNEnglish is uniformly the weakest language with no condition achieving MTWV above 0.67 in either development or evaluation query sets.

VI. COMPARATIVE EVALUATION

A. Multilingual Comparison Across Query Types

Table V presents the comparative performance of all five STD systems on multilingual evaluation queries. The text-based system achieves the best Type 1 minCNXE (0.2489) and MTWV (0.8337) across all five methods by a substantial margin. For Type 3 multi-word expressions, the text-based system (minCNXE=0.5205, MTWV=0.5431) also outperforms DTW (0.7601, 0.2138), CNN (0.6569, 0.3603), and End-to-End (0.6278, 0.3617). The WCSO-DSNN achieves MTWV=0.0000 across all multilingual conditions, reflecting poor score calibration rather than complete detection failure, as it achieves consistent minCNXE values of approximately 0.55, confirming genuine score-level discriminative capability.

TABLE V: Comparative Performance — All Methods, Multilingual Evaluation Queries

System	All Types minCNXE	Type 1 minCNXE	Type 1 MTWV	Type 2 minCNXE	Type 3 minCNXE
Text-Based (Ours)	0.4773	0.2489 ★	0.8337 ★	0.6301	0.5205
End-to-End [10]	0.5850	0.3796	0.6499	0.5158	0.6278
CNN Matching [9]	0.5480	0.4121	0.6103	0.5235	0.6569
DTW Matching [8]	0.6600	0.4606	0.5663	0.6013	0.7601
WCSO-DSNN [21]	0.5575	0.5640	0.0000	0.5593	0.5531

B. Language-Specific Comparison (Type 1 Queries)

Table VI presents language-specific Type 1 minCNXE across all five methods. The text-based system achieves the lowest minCNXE across all six languages. The WCSO-DSNN achieves the second-lowest minCNXE for Albanian (0.3572) and NNEnglish (0.5735) among acoustic methods, outperforming DTW, CNN, and End-to-End for these languages. For Basque, Czech, Romanian, and Slovak, End-to-End and CNN methods outperform WCSO-DSNN acoustically, while all acoustic methods are substantially outperformed by text-based STD.

TABLE VI: Language-Specific Type 1 minCNXE — All Five Methods

Method	Albanian	Basque	Czech	NNEnglish	Romanian	Slovak
Text-Based (Ours)	0.3257 ★	0.2190 ★	0.1751 ★	0.4226 ★	0.1012 ★	0.1115 ★
End-to-End [10]	0.3800	0.2200	0.4200	0.6200	0.1700	0.2100
CNN Matching [9]	0.4100	0.2400	0.4600	0.6000	0.1800	0.2200
DTW Matching [8]	0.4200	0.4150	0.5200	0.6100	0.2200	0.2300
WCSO-DSNN[21]	0.3572	0.5086	0.5979	0.5735	0.5361	0.5627

C. Maturity Assessment

Text-based STD is significantly more mature than speech-to-speech acoustic matching for settings where reliable ASR is available. The performance gap is large and consistent: the text-based system achieves Type 1 minCNXE values of 0.10 to 0.33 across all six languages, while the best acoustic methods achieve 0.17 to 0.62 for the same query type. The MTWV gap is particularly striking: text-based STD achieves MTWV values of 0.65 to 0.91 for Type 1 across all languages, contrasted with MTWV=0.0000 for the WCSO-DSNN across all multilingual conditions.

Speech-to-speech acoustic matching retains a critical practical advantage in true zero-resource settings where ASR is unavailable. The complementarity of the two paradigms is most clearly demonstrated by NNEnglish, where both systems struggle equally, confirming that multi-accent speaker variability is a challenge that neither paradigm has yet solved. These findings establish that text-based STD is the recommended paradigm for multilingual spoken term detection wherever reliable ASR is available, while speech-to-speech acoustic matching remains indispensable for zero-resource environments.

VII. DISCUSSION

A. Effect of ASR Quality on Retrieval Performance

The results confirm that ASR transcript quality is the single most important determinant of text-based STD performance, operating independently of morphological complexity. Czech demonstrates this most clearly: despite featuring conversational speech with seven grammatical cases — typically the most challenging conditions for acoustic matching — it achieves the best All Types MTWV of any language (0.8056) because Czech ASR transcript quality is sufficiently high to support near-perfect term-level matching. Conversely, NNEnglish — with relatively simple English morphology — achieves the worst results because multi-accent speaker variability fundamentally degrades ASR quality for both query and archive transcripts.

B. Morphological Complexity and Type 1-to-Type 2 Degradation

The Type 1-to-Type 2 MTWV collapse pattern reveals the secondary role of morphological complexity. Slovak shows the largest collapse (0.5128 MTWV drop) due to its rhythmic law producing systematically different vowel patterns in inflected surface forms relative to citation-form query transcripts. Basque shows the second largest (0.3807 drop) due to agglutinative suffixation. Czech, despite its seven grammatical cases, shows a smaller collapse than Slovak or Basque because its higher ASR quality ensures richer indexed transcripts with more partial lexical overlap for inflected forms. Albanian is the unique case where Type 2 exceeds Type 1, confirming that its inflected forms are reliably transcribed and indexed.

C. Score Calibration Analysis

The gap between actCNXE and minCNXE provides diagnostic insight into score calibration quality. This gap is smallest for Type 1 (gap=0.3124 in multilingual evaluation) and largest for Type 2 (gap=1.2126), confirming that Hiemstra_LM scores are more directly useful for operational decision-making for exact-match queries than for lexically varied ones. For operational deployment, score normalisation or threshold optimisation would meaningfully improve actTWV for Type 2 and Type 3 conditions.

D. Limitations

The proposed system has three principal limitations. First, it is fundamentally bounded by ASR transcript quality, making it inapplicable in true zero-resource settings. Second, morphological lexical variation (Type 2) poses a structural challenge: the Hiemstra_LM performs term-level matching, so inflected surface forms not present verbatim in the indexed transcripts generate low or zero retrieval scores regardless of semantic equivalence. Third, the system does not provide temporal keyword localisation within audio documents, as Terrier retrieves entire documents rather than timestamp-bounded segments.

VIII. CONCLUSION

This paper presented a complete text-based STD pipeline implemented using HappyScribe ASR and the Terrier IR framework with the Hiemstra language model, evaluated on the QUESST 2014 multilingual benchmark across six languages and three query types. The system achieves a best multilingual Type 1 minCNXE of 0.2489 and MTWV of 0.8337, substantially outperforming DTW, CNN, End-to-End neural, and WCSO-DSNN acoustic matching baselines. Romanian and Slovak achieve near-ceiling performance (MTWV=0.9123) driven by high-quality ASR transcription, while NNEnglish is the only language where neither paradigm achieves satisfactory performance, establishing multi-accent speaker variability as the primary unsolved challenge.

Future work will investigate: (i) replacing HappyScribe with Whisper-based transcription to potentially improve performance for morphologically complex and accented-speech languages; (ii) phoneme or subword-level query expansion to address the OOV limitation; (iii) hybrid systems combining acoustic embeddings with text-based retrieval for robustness in both high-resource and zero-resource settings; and (iv) passage-level indexing with forced alignment to add temporal keyword localisation to the IR-based pipeline.

REFERENCES

[1] J. G. Fiscus, J. Ajot, J. S. Garofolo, and G. Doddington, "Results of the 2006 Spoken Term Detection Evaluation," in Proc. NAACL-HLT, 2007, pp. 51–57.

[2] A. Mandal, K. R. Prasanna Kumar, and P. Mitra, "Recent developments in spoken term detection: A survey," Int. J. Speech Technol., vol. 17, no. 2, pp. 183–198, 2014.

[3] QUESST 2014 Challenge Database, "Query by example search on speech (QUESST) 2014," [Online]. Available: http://speech.fit.vutbr.cz/files/quesst14Database.tgz. Accessed: Oct. 2024.

[4] X. Anguera et al., "QUESST2014: Evaluating QbE speech search in a zero-resource setting with real-life queries," in Proc. IEEE ICASSP, Apr. 2015, pp. 5833–5837.

[5] HappyScribe, "Automatic Speech Recognition and Transcription Service," [Online]. Available: https://www.happyscribe.com. Accessed: Jan. 2020.

[6] C. Macdonald, R. McCreadie, R. Santos, and I. Ounis, "From Puppy to Maturity: Experiences in Developing Terrier," in Proc. SIGIR OSIR Workshop, 2012.

[7] D. Hiemstra, Using Language Models for Information Retrieval, Ph.D. dissertation, Univ. Twente, 2001.

[8] Y. Zhang and J. R. Glass, "Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams," in Proc. IEEE ASRU, Dec. 2009, pp. 398–403.

[9] D. Ram, A. Asaei, and H. Bourlard, "Phonetic subspace features for improved QbE-STD," Speech Commun., vol. 103, pp. 27–36, 2018.

[10] D. Ram, L. Miculicich, and H. Bourlard, "Neural network based end-to-end query by example spoken term detection," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 28, pp. 1416–1427, 2020.

[11] J. Li, X. Wang, and B. Xu, "An empirical study of multilingual and low-resource spoken term detection using deep neural networks," in Proc. Interspeech, 2014, pp. 1747–1751.

[12] A. Norouzian, Techniques for Two-Stage Open Vocabulary Spoken Term Detection and Verification, Ph.D. dissertation, McGill Univ., Montreal, Canada, 2015.

[13] A. Singh, V. Arora, and Y. P. P. Chen, "An efficient TF-IDF-based QbE-STD," in Proc. IEEE CAI, Jun. 2024, pp. 170–175.

[14] B. Yusuf and M. Saraçlar, "Written term detection improves spoken term detection," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 32, pp. 3213–3223, 2024.

[15] J. Tejedor and D. T. Toledano, "Whisper-based spoken term detection systems for ALBAYZIN evaluation," EURASIP J. Audio, Speech, Music Process., vol. 2024, Art. no. 15, 2024.

[16] H. Kamper, A. Jansen, and S. Goldwater, "Unsupervised word segmentation and lexicon discovery using acoustic word embeddings," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 24, no. 4, pp. 669–679, 2016.

[17] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in Proc. NeurIPS, 2020.

[18] W.-N. Hsu et al., "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 29, pp. 3451–3460, 2021.

[19] D. Ram, "Language Independent QbE-STD," Ph.D. dissertation, Idiap Research Institute, Martigny, Switzerland, Nov. 2019.

[20] D. Ram, A. Asaei, and H. Bourlard, "Sparse subspace modeling for QbE-STD," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 26, no. 6, pp. 1130–1143, 2018.

[21] Sushil Venkatesh Kulkarni and Sukomal Pal, "Water chicken swarm optimization-based deep segmental

neural network for spoken term detection using bayesian filtering," *Multimedia Tools and Applications*,

Springer Nature,