DEVELOPING KEYWORD SEARCH UNDER THE IARPA BABEL PROGRAM

Jonathan Mamou[1], *Jia Cui*[2], *Xiaodong Cui*[2], *Mark J. F. Gales*[3],
Brian Kingsbury[2], *Kate Knill*[3], *Lidia Mangu*[2], *David Nolden*[4], *Michael Picheny*[2],
Bhuvana Ramabhadran[2], *Murat Saraclar*[2], *Ralf Schlüter*[4], *Abhinav Sethy*[2], *Philip C. Woodland*[3]

[1]IBM Haifa Research Labs, Haifa 31905, Israel
[2]IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, USA
[3]Cambridge University Engineering Department, Trumpington St., Cambridge CB2 1PZ, U.K.
[4]Chair of Computer Science 6, RWTH Aachen University, Ahornstr. 55, D-52056 Aachen, Germany

ABSTRACT

Spoken content in languages of emerging importance needs to be searchable to provide access to the underlying information. Keyword search (KWS), also known as spoken term detection (STD), is a speech processing task in which the goal is to find all the occurrences of a textual "keyword", a sequence of one or more words, in a large corpus of speech data. In 2006, the U.S. National Institute of Standards and Technology (NIST) created the STD evaluation initiative to facilitate research and development of technology for retrieving information from archives of speech data. The STD 2006 evaluation revealed a close relationship between the performance of KWS technology and the state of the art for automatic speech recognition (ASR) technology on a given combination of language and genre. Reducing the performance gap between high-resource, well studied languages and low-resource, lightly studied languages is one of the primary aims of the IARPA Babel program: "the goal of the Babel Program is to be able to rapidly develop speech recognition capability for keyword search in a previously unstudied language, working with speech recorded in a variety of conditions with limited amounts of transcription." [1]. To achieve this goal, in each period of Babel, program performers work with a diverse set of development languages to gain experience with KWS. In the first period, the development languages are Cantonese, Tagalog, Pashto and Turkish. At the end of each period, there are evaluations of KWS performance on the development languages, and on a surprise language, Vietnamese, where performers have a limited period of time for system development.

In this paper, we present our system for KWS on noisy speech for low-resource languages developed in the framework of the IARPA Babel program. Our approach achieves good performance by combining postings lists produced by diverse speech recognition systems from three different research groups. We apply these methods tolanguages that present some new issues in terms of reduced resources and query lengths. This approach is used for the NIST Open Keyword Search 2013 (OpenKWS13) Evaluation [2], as part of the Babel base period evaluation. Experiments and analysis are provided on the evaluation data. First, we show score normalization methodology that improves in average by 26% keyword search performance. Second, we show that properly combining the outputs of diverse ASR systems performs in average 21% better than the best normalized ASR system.

***Index Terms*—** spoken term detection, keyword search, data fusion, system combination, score normalization

1. INTRODUCTION

The rapidly increasing amount of spoken data calls for solutions to index and search this data. In 2006, NIST created the STD evaluation [3] initiative to facilitate research and development of technology for retrieving information from archives of speech data. However, it focused on large-resource languages and most of the effort focused on clean speech.

It has recently been demonstrated that significant improvement on STD task can be obtained by deliberately designing diverse and complementary ASR components (i.e., front ends, acoustic models, etc) [4]. We show that similar approach works on noisy speech for low resource languages with low target false alarm rate. We have presented score normalization and system combination methodologies in [5]. In this paper, we apply these methods to IARPA Babel base period evaluation languages: Cantonese, Pashto, Turkish, Tagalog and Vietnamese.

The basic processing flow is as follows. The audio data is transcribed using diverse ASR systems. Each ASR system output is indexed separately. Each query is searched against the different indices. The scores of the hits are possibly normalized. The hit lists returned by the different indices are merged to form a single *meta-hit* list for the query and a score is attributed to the meta-hit.

The paper is organized as follows. After presenting the task (Section 2), we present score normalization methods and

system combination approaches (Section 3). Experiments are presented (Section 4). Finally, we conclude (Section 5).

2. TASK DESCRIPTION

The present work addresses the STD task defined by NIST for the 2006 STD Evaluation with some modifications introduced by IARPA's Babel program [1]. The task consists in finding all the exact matches of a specific query in a given corpus of speech data. A query is a textual phrase containing one or several terms. We focus in the NTAR (no test audio reuse) task where the system components and word indices are frozen before the queries are provided. Manual transcripts of the speech are not provided but are used by the evaluators to find true occurrences. By definition, true occurrences of a query are found automatically by searching the manual transcripts using the following rule: the gap between adjacent words in a query must be less than 0.5 seconds in the corresponding speech. For evaluating the results, each system output occurrence is judged as correct or not according to whether it is close in time to a true occurrence of the query retrieved from manual transcripts; it is judged as correct if the midpoint of the system output occurrence is less than or equal to 0.5 seconds from the time span of a true occurrence of the query. KWS performance is measured by the Term-Weighted Value (TWV) metric, which combines missed detection and false alarm error types [3]. More precisely, TWV is 1 minus the weighted sum of the term-weighted probability of missed detection and the term-weighted probability of false alarms. MTWV is the maximum TWV over the range of all possible values of the detection threshold. MTWV ranges from $-\infty$ to +1.

3. SCORE NORMALIZATION AND SYSTEM COMBINATION FOR KEYWORD SEARCH

Score normalization and system combination methodologies are described in details in [5]; they extend data fusion methodologies widely used for document IR [6]. Sum-to-one (STO) normalization takes into account term frequency diversity while system combination promotes ASR system diversity. Therefore, given hit lists from diverse ASR systems, it is necessary to normalize the detection scores and combine the results from all systems to generate a final output.

Suppose that there are N_q hits for the query q according to a given ASR system. Let $s_{q,i}$ denote the score of the i-th hit for the query q. The STO normalization computes new scores as

$$\frac{s_{q,i}}{\sum_{j=1}^{N_q} s_{q,j}}$$

The specific system combination method used in this work is denoted STO-CombMNZ-STO and consists in:

1. applying STO normalization to each hit list,
2. combining the results using CombMNZ data fusion method,
3. applying STO normalization to the fused hit list to produce the final output.

.

Language	Baseline MTWV	STO MTWV	% Improvement
Cantonese	0.402	0.498	24
Pashto	0.355	0.419	18
Turkish	0.484	0.556	15
Tagalog	0.429	0.528	23
Vietnamese	0.267	0.396	48

Table 1. MTWV results for mutliple languages using STO score normalization.

4. EXPERIMENTS

4.1. Experimental Setup

Results are reported on the language collections from the IARPA Babel Program, evalpart1 partition, full language pack: Cantonese (release babel101b-v0.4c), Pashto (release babel104b-v0.4bY), Turkish (release babel105b-v0.4), Tagalog (release babel106b-v0.2g) and Vietnamese (release babel107b-v0.7). The data collection covers multiple aspects of spanning dialects, topics, gender and age. The training data is basically telephone conversational data and some scripted data. The test data is limited to only conversational data. The KWS results are produced by diverse ASR systems generated by Cambridge University, IBM and RWTH Aachen. The description of the ASR systems is beyond the scope of this paper. For Cantonese, ASR systems have been described in details in [7, 8] Our KWS system is implemented using the OpenFst toolkit [9]. MTWV is evaluated using the F4DE NIST Evaluation tool [10].

4.2. Score Normalization

We present KWS performance for STO normalization methodology in Table 1 for a speaker adaptively trained ASR system using deep neural network acoustic model. In the second column, we report MTWV results with un-normalized raw scores (posterior probabilities), and in the last column, we report the relative MTWV improvement of the STO results over the baseline. STO score normalization methodology improves MTWV for all the languages with a strongest effect for languages with lower KWS performance.

4.3. System Combination

We present KWS improvement for STO-CombMNZ-STO system combination in Table 2. In the second column, we

Language	Best System MTWV	Combined MTWV	Improvement (%)
Cantonese	0.527	0.574	9
Pashto	0.410	0.482	17
Turkish	0.633	0.686	8
Tagalog	0.528	0.583	10
Vietnamese	0.399	0.528	32

Table 2. MTWV results for mutliple languages using system combination.

Language	Query Type	% queries	% Combination Improvement
Cantonese	IV	93	8
Cantonese	OOV	7	27
Pashto	IV	98	16
Pashto	OOV	2	83
Tagalog	IV	91	9
Tagalog	OOV	9	88
Vietnamese	IV	99	32
Vietnamese	OOV	1	14

Table 3. System combination MTWV relative improvement as a function of the query type.

report MTWV results for the best single ASR system, and in the last column, we report the relative MTWV improvement of the system combination over the best single system. System combination methodology improves MTWV for all the languages with a strongest effect for languages with lower KWS performance.

4.4. Query Type Analysis

We analyze the effect of system combination on KWS performance as a function of query type (IV or OOV). Table 3 presents a breakdown of results by query type. For each type, we provide the rate of queries belonging to this type. We report in the last column the relative improvement of STO-CombMNZ-STO system combination over the best normalized single ASR system. System combination improves MTWV for both IV and OOV queries for all the languages, with the effect being strongest for OOV queries.

5. CONCLUSION

These methods have been successively applied to IARPA Babel base period evaluation languages. STO score normalization methodology improves in average by 26% KWS performance, and system combination approach improves in average by 15% KWS performance.

6. ACKNOWLEDGMENT

We are grateful to Janice Kim of IBM Research for providing software support for the keyword search toolkit.

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense U.S. Army Research Laboratory (DoD / ARL) contract number W911NF-12-C-0012. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government.

7. REFERENCES

[1] M. Harper, "IARPA Solicitation IARPA-BAA-11-02," `http://www.iarpa.gov/solicitations_babel.html`, 2011.

[2] "NIST Open Keyword Search 2013 (OpenKWS13) Evaluation," `http://www.nist.gov/itl/iad/mig/openkws13.cfm`.

[3] J.G. Fiscus, J. Ajot, J.S. Garofolo, and G. Doddington, "Results of the 2006 spoken term detection evaluation," in *Proceedings of ACM SIGIR Workshop on Searching Spontaneous Conversational*. Citeseer, 2007, pp. 51–55.

[4] L. Mangu, H. Soltau, H.-K. Kuo, B. Kingsbury, and G. Saon, "Exploiting diversity for spoken term detection," in *Proc. ICASSP*, 2013.

[5] J. Mamou, J. Cui, X. Cui, M. J. F. Gales, B. Kingsbury, K. Knill, L. Mangu, D. Nolden, M. Picheny, B. Ramabhadran, R. Schlüter, A. Sethy, and P. C. Woodland, "A high-performance Cantonese keyword search system," in *Proc. ICASSP*, 2013.

[6] S. Wu, *Data Fusion in Information Retrieval*, vol. 13, Springer, 2012.

[7] B. Kingsbury, J. Cui, X. Cui, M. J. F. Gales, K. Knill, J. Mamou, L. Mangu, D. Nolden, M. Picheny, B. Ramabhadran, R. Schlüter, A. Sethy, and P. C. Woodland, "A high-performance Cantonese keyword search system," in *Proc. ICASSP*, 2013.

[8] J. Cui, X. Cui, B. Ramabhadran, J. Kim, B. Kingsbury, J. Mamou, L. Mangu, M. Picheny, T. N. Sainath, and A. Sethy, "Developing speech recognition systems for corpus indexing under the iarpa babel program," in *Proc. ICASSP*, 2013.

[9] C. Allauzen, M. Riley, J. Schalkwyk, W. Skut, and M. Mohri, "Openfst: A general and efficient weighted finite-state transducer library," *Implementation and Application of Automata*, pp. 11–23, 2007.

[10] "NIST Tools," `http://www.itl.nist.gov/iad/mig/tools/`.