Overview of the NIST Open Keyword Search 2013 Evaluation Workshop

Nancy F. Chen, Jonathan G. Fiscus

SLTC Newsletter, August 2013

IARPA BABEL PROGRAM

The NIST Open Keyword Search 2013 (OpenKWS13) Evaluation Workshop was the first in a series of community-wide evaluations to test research systems that search for keywords in audio of a "surprise" language whose identity is unveiled at the beginning of the evaluation period. OpenKWS13 made use of the surprise language data set developed for IARPA's Babel Program [1] that was made available to the wider research community. During the past 15 months, the four Babel teams were given four development language build packs to develop keyword search capabilities across languages. This was followed up by an evaluation on a surprise language that was held from March 25 to May 1, 2013; participants had four weeks to build their systems and one week to return keyword search results once the keywords were provided by NIST. While only the Babel teams worked with and were evaluated on the four development language packs, the Surprise Language Evaluation was open to the entire speech research community. The ultimate goal of the Babel program is to be able to rapidly develop keyword search spoken language technologies for ANY language, especially for under-resourced languages where linguistic resources such as word transcriptions are limited [1]. Currently, most spoken language technologies are only extensively investigated using a few resource-rich languages such as English. The Babel initiative is particularly meaningful given that there are over 7,000 languages in the world, of which over 300 have more than one million speakers.

OPENKWS13 EVALUATION SETUP

The surprise language of OpenKWS13 is Vietnamese. Vietnamese is the language with the most speakers (~76million) in the Austroasiatic language family. Vietnamese is a challenging language - it is a monosyllable-based, tonal language, with regional dialects that might not always be mutually intelligible.

Participants of the OpenKWS13 Evaluation are given two language packs: (1) Full Language Pack (FullLP), consisting of 20 hours of word-transcribed scripted speech, 80 hours of word-transcribed conversational telephone speech, and a pronunciation lexicon; and (2) Limited Language Pack (LimitedLP), which consists of a 10-hour subset of FullLP plus the remaining audio without transcription. The contrast between the Full and Limited language pack conditions requires researchers to tackle out-of-vocabulary search terms and enables the exploration of unsupervised methods. There are three language resource (LR) conditions: (1) BaseLR: LRs limited to the supplied language pack, (2) BabelLR: LRs include non-test languages, Babel-supplied language packs, (3) OtherLR: LRs not mentioned in (1) and (2). In addition, there are two audio conditions: (1) No test audio reuse (NTAR), and (2) Test audio reuse (TAR). The contrast between TAR and NTAR is whether or not the keyword search system is allowed to reprocess the audio after obtaining the search keywords. In the NTAR condition, a system is constructed so that it processes the audio one time only without knowing all possible keywords that could be framed by future users; whereas, in the TAR condition, the audio can be reprocessed after the keywords are known to support keyword-specific indexing methods. Participants are encouraged to submit results for as many different conditions as possible, but the minimum submission requirement is FullLP+BaseLR+NTAR. In addition to this condition, the Babel teams are also required to submit results on the LimitedLP+BaseLR+NTAR condition to help investigate how much algorithmic advantage is achieved from the amount of transcribed data.

During the final week of the evaluation, the participants are given 75 hours of un-segmented conversational speech and a list of 4,065 keywords in orthographic format. Systems returned the time region of each detected keyword along with an estimate how likely the keyword was to occur and a detection decision.

There were 12 participating teams, four of which are members of the IARPA's Babel program (namely:

BABELON, LORELEI, RADICAL, SWORDFISH). The remaining eight teams were from around the world including China, Israel, Singapore, and the USA. Participating teams were from both academia and industry. The organizers of the workshop expressed gratitude for the efforts of all the participating teams, including the 8 non-Babel teams, who participated as volunteers in the evaluation.

SELECTED TECHNICAL HIGHLIGHTS

The OpenKWS13 results on the required submission of FullLP+BaseLR+NTAR are plotted in Fig. 1, which shows the detection error trade-off curves of the submitted keyword search (KWS) systems. Most participants used automatic speech recognition (ASR) as the backbone of the keyword search engine. Acoustic modeling was crucial in achieving superior ASR performance, as there is limited text data to train good language models. Most top performing teams incorporated deep neural networks to enhance their ASR capabilities. Robust acoustic features were also found to be helpful in obtaining good ASR and KWS results. Most top-performing teams integrated tonal features. In addition to traditional pitch features, the RADICAL team (led by CMU) also included other novel features such as fundamental frequency variation (FFV) spectrum [2, 3].

Misses are detrimental to KWS systems. This is especially true for languages like Vietnamese, where many words are only one syllable in length, and thus very short in duration. The LORELEI team (led by IBM) exploited cascaded search, which first searches the word index for a particular keyword, and continues to search the phonetic index if no result is returned in the first stage [4]. Score normalization was also shown to be crucial in improving performance and reducing the probability of missing rare search terms [5, 6]. The BABELON team (lead by BBN) showed that with proper score normalization, results can improve by 10% absolute or more. In addition, they showed that the effect of normalization is greater if applied earlier in the pipeline (e.g., before system combination.)

Many participants expressed interest in continuing related research inspired from the OpenKWS13 Evaluation. For example, the SINGA team (led by I2R) is extending their Vietnamese transliteration system to other under-resourced Southeast Asian languages to help resolve out-of-vocabulary issues; the STAR-SRI team (led by SRI) is exploring articulatory features derived from synthesized Vietnamese; and the Swordfish team (led by ICSI) is working on injecting higher-level linguistic structure into their KWS system.

FUTURE OPENKWS EVALUATIONS

Due to the success of this year's Open Keyword Search Evaluation, NIST is planning to continue organizing Open Keyword Search Evaluations to the public in the following years. Potentially interested participants are welcome to visit a href="http://www.nist.gov/itl/iad /mig/openkws13.cfm">http://www.nist.gov/itl/iad/mig/openkws13.cfm to familiarize themselves with this year's evaluation information and to join the mailing list by sending an email to openkws-poc@nist.gov.

Disclaimer: Certain commercial equipment, instruments, software, or materials are identified in this article in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the (NIST), nor is it intended to imply that the equipment, instruments, software or materials are necessarily the best available for the purpose.

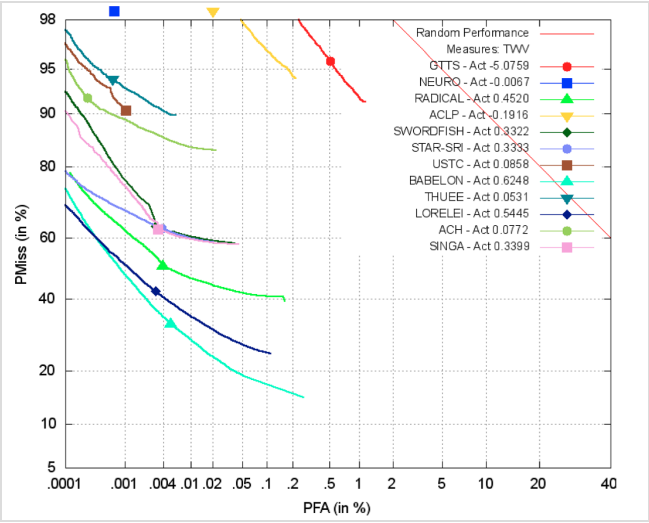

Fig. 1. Detection error trade-off curves of the 12 participating teams of OpenKWS13 Evaluation

August 2013 Newsletter Home

From the IEEE SLTC chair

Call for Nominations: Speech and Language Technical Committee (SLTC) Member Positions

Call for Nominations: Vice Chair of the Speech and Language Technical Committee (SLTC)

IEEE Signal Processing Society Newsletter

Calls for papers, proposals, and participation

Job advertisements

An Overview of ICML 2013 Deep Learning Workshop

Joint IEEE/ACM Transactions on Audio, Speech, and Language Processing

Web-Based Validation for Small Corpora

A make-over of the Signal Processing Letters

Scanning the Special Issue of the Proceedings of the IEEE on Speech Information Processing

Overview of the NIST Open Keyword Search 2013 Evaluation Workshop

Interview: Speech Interfaces for Low-Literate Users

SLT @ SLT: IEEE Spoken Language Technology (SLT) Workshop in South Lake Tahoe (SLT)

Subscribe to the newsletter

SLTC Home

on the FullLP+BaseLR+NTAR condition.

REFERENCES

[1] "IARPA broad agency announcement IARPA-BAA-11-02," 2011, https://www.fbo.gov/utils/view?id=ba991564e4d781d75fd7ed54c9933599.

[2] K. Laskowski, J. Edlund, and M. Heldner, "An Instantaneous Vector Representation of Delta Pitch for Speaker-Change Prediction in Conversational Dialogue Systems," Proceedings of ICASSP, 2008.

[3] K. Laskowski, J. Edlund, and M. Heldner, "Learning Prosodic Sequences Using the Fundamental Frequency Variation Spectrum", Proceedings of Speech Prosody, 2008.

[4] M. Saraclar and R. Sproat, "Lattice-Based Search for Spoken Utterance Retrieval", Proceedings of ACL, 2004.

[5] J. Mamou, et. al., "System Combination and Score Normalization for Spoken Term Detection," Proceedings of ICASSP, 2013.

[6] B. Zhang, R. Schwartz, S. Tsakalidis, L. Nguyen, and S. Matsoukas, "White listing and score normalization for keyword spotting of noisy speech," Proceedings of INTERSPEECH, 2012. *Nancy F. Chen is a scientist at the Institute for Infocomm Research (I2R), Singapore. Her current research interests spoken language processing for under-resourced languages, speech summarization, and dialect/accent recognition. For more information: http://alum.mit.edu/www/nancychen. Jonathan Fiscus is a computer scientist at the National Institute of Standards and Technology. His research focuses on developing technology performance assessment methods for speech, text and video content extraction technologies. Email: jfiscus@nist.gov*