Zero-Cost Speech Recognition Task at Mediaeval 2016

Igor Szoke
Brno University of Technology
Brno, Czech Republic
szoke@fit.vutbr.cz

Xavier Anguera
ELSA Corp.
Lisboa, Portugal
xavier@elsanow.io

ABSTRACT

The main goal of the Zero-Cost Speech Recognition task is to bring researchers together on the topic of training ASR systems using only publicly available data. In particular, the task consists on the development of either an LVCSR or a subword speech recognizer on a given target language. For this year we selected Vietnamese as the target language The organizers provided participants several sets of publicly available data combined with one proprietary set gathered for this evaluation. Participants are free to find and use other publicly available resources (free for research purposes). These resources must be shared with other participants till end of July. The data-set for the evaluation is then fixed and no outside data can be used.

1. INTRODUCTION

There are more than thousand spoken languages. Any research lab or "Speech company" that wants to develop technology in a new language usually needs to buy a speech database (audio + carefully hand made transcripts) to get started [1]. The cost of such databases range from 5k to 20k EURs (provided by LDC / ELRA / Appen etc.). Alternatively, if you have enough budget, you can collect your own data and cover some specific topics, acoustic environments etc. This brings a huge gap between "top" labs and companies having enough budget to afford such expenditures with "the other" small players, endowed to freely available data, tools or even systems.

The main goal of this task is to challenge participating teams to come up and experiment with bootstrapping techniques, which allow to train initial ASR system for "free". We are interested in the exploration of techniques to allow researchers to train ASR systems on public multimedia resources (texts, audios, videos, dictionaries etc.), without the need to buy expensive (ideally any) data-sets. Participants may inspire in approaches used in under-resourced languages [5, 8]. There were also other initiatives close to the topic of this tasks: Zero Resource Speech Challenge in 2015 [4] and The Third Frederick Jelinek Memorial Summer Workshop 2016 [2].

[1]We understand any speech tokenizer under abbreviation ASR – i.e. including phoneme recognizer, word recognizer, automatic unit discovery

MediaEval 2016 Workshop, Oct. 20-21, 2016, Hilversum, Netherlands.

2. THE ZERO-COST 2016 DATA-SET

The target language selected for this year is Vietnamese. One of the reasons why we choose Vietnamese is that it was one of the languages of OpenKWS/BABEL in 2013 [1, 7] and there are many papers reporting results in ASR [14, 10, 11, 12, 6, 15, 13]. BUT as the task co-organizer provided "upper-bound" results using their BABEL system [9] and "calibrate" results of Zero-Cost participants to the larger world wide speech community. The BUT baseline was trained only on conversational telephone speech without any adaptation on target domain (Zero-Cost).

Other reasons for choosing Vietnamese are that it is a low-resourced language with limited resources available on-line (more difficult for participants to "cheat" with data other than what is provided), it is a tonal language (with its inherent difficulties) but it is a syllabic language (simpler to treat for zero-resources algorithms where clear phoneme sequences can be seen repetitively in the data).

Task organizers provided participants with an initial set of free multimedia resources – a mix of audio data and imperfect transcripts like audios/videos with subtitles:

- Forvo.com – Download of Vietnamese data from Forvo.com service. It is composed of a collection of short recordings with one or more word pronunciations each.
- Rhinospike.com – Download of Vietnamese data from Rhinospike.com service. It is a collection of recordings consisting between one short to several long sentences.
- ELSA – Proprietary prompted data recorded with a mobile application by Vietnamese students. It contains several read sentences obtained from a book of Vietnamese quotes. This data simulates a case where participant is able to collect small amount of data themselves.
- Other "surprise" test data – Surprise data aiming at evaluating how robust participant systems are to new data. This data is a download of 35 YouTube videos (broadcast news, presentations, talks) mostly containing one speaker. The first 2 minutes of each videos were transcribed and used as ground truth. The rest of video was let in the test set to augment the data for possible unsupervised adaptation.

Please note, that transcripts may not match the audio in 100%. In addition, any audio may contain some dropouts, noise or some speech may be missing. This data has been preprocessed, split into Train / Devel / Test, and converted to 16kHz wav + STM references.

	Train	Devel-Local	Devel	Test	SUM
Forvo.com	663.3	2.8	34.3	33.8	731.4
Rhinospike.com	122.4	2.3	10.1	7.8	140.3
ELSA	43.1	8.5	43.2	58.8	145.1
Surprise data	–	–	–	40.7	40.7
SUM	828.8	13.6	87.6	141.1	

Table 1: *Distribution of data (in minutes) according to a set and a data source. Devel-Local is a subset of Devel.*

In addition to the "official" datasets, several participants have provided some free data which we encourage other participants to use. Apart from these, the use of no other data is allowed. Train, Development and Test sets are available already during system training. Participants can use them and adapt their system on them (e.g. unsupervised adaptation on the Test set). However, reference transcripts are not provided for the development / test data and it is not allowed to transcribe or manually analyze it.

2.1 Participants' data description

- I2R - A list of 890k Vietnamese webpage URLs.
- I2R - A Vietnamese wordlist – 80k words.
- I2R - A raw dump of Vietnamese wiktionary – later cleaned by I2R to 750MB of text.
- BUT - A download of Vietnamese-English subtitles [3] – 93MB of text.
- BUT - A set of Vietnamese videos and subtitles – 14 partly subtitled episodes of a Vietnamese telenovel.

3. BRIEF SUB-TASKS DESCRIPTION

Participants of the task are asked to train a speech tokenizer – LVCSR or subword – on a collection of public data (see section 2) in Vietnamese language. Each participant must take part in at least one sub-task.

3.1 Large vocabulary continuous speech recognition (LVCSR) sub-task

This task targets full speech recognition where the output is a sequence of recognized words. Systems will be evaluated on the Word-Error-Rate (WER) metric (using cstk scoring tool). The WER is based on the comparison of transcripts (reference and generated hypothesis) at word level. Both transcripts should be produced in uppercase and without punctuation, hesitation markers etc. There is no other text normalization done. This sub-task main use-case scenario is in areas where full speech transcript is needed.

3.2 Subword speech recognition sub-task

This task aims at building a "light weight" speech recognizer. The output is a sequence of subword tokens/units. We do not define what the tokens should be. It can be phonemes, graphemes, syllables, triphones, automatically estimated units, etc. This sub-task's main use-case scenario is for areas where speech must be converted to a sequence of discrete symbols (LID, SID, KWS, Topic detection, etc.). Phoneme units are used as ground truth for this sub-task.

The evaluation metric used in this sub-task is a Normalized version of the Mutual Information (NMI) also called the *symmetric uncertainty*. It ranges between 0 and 1. When both variables, X and Y, are independent, meaning that the units discovered are completely unrelated to the reference phone labels, the results is zero. The maximum, one, is achieved when one can fully recover the phone sequence from the discovered units AND the entropy of both the discovered units and the reference phone is the same. This means that the metric penalizes systems that have too many units. The evaluation algorithm used to compute NMI takes into account timing of the discovered units. It matches them to reference ones (according to time) first and then calculates the NMI.

4. EVALUATION, SCORING AND LEADER BOARD

Participants are provided with Training, Development and Test (evaluation) data all at once. However they do not have references transcripts for Development and Test data. They can use the on-line leader board to score their systems and compute development results. When the evaluations are over, the results on the Test set will be published. To make the development faster and easier, we defined a Devel-Local subset and provided the ground truth to participants, so that they can perform initial development on their systems locally.

- The Devel-Local is 1/5 subset of Devel. Participants are provided with references and scoring scripts so that they can score their system outputs on this subset. This was done to allow for quick iterations during training period and to overcome the need to upload the system outputs to the leader board too often.
- The Devel consists of the full Devel dataset. Once participants end up with some good enough / improved enough system, they are encouraged to upload their results to the leader board and be scored on much more data. The uploaded scores are available for all participants to see.
- The Test is "unseen" data. It partly contains data similar to training / devel one but also unseen one. Participants are encouraged to adapt their systems on this data (in unsupervised ways).

Each participant has to register and submit their results to the on-line Leader Board (`http://www.zero-cost.org/`). There is no maximum limit on the number of submissions per team. For the official final scoring each participant must define one primary submission by adding the `P-` prefix to their submission and optionally at most 5 others as contrastive systems `C1-` – `C5-`.

The submission deadline is 12th of September 2016.

5. ACKNOWLEDGMENTS

We would like to thank the Mediaeval organizers for their support and all the participants for their hard work.

6. REFERENCES

[1] Openkws/babel in 2013 web: https://www.nist.gov/multimodal-information-group/openkws13-evaluation/.

[2] The third frederick jelinek memorial summer workshop 2016 web: http://www.clsp.jhu.edu/workshops/16-workshop/building-speech-recognition-system-from-untranscribed-data/.

[3] Vietnamese-english subtitles web: http://opus.lingfil.uu.se/.

[4] Zero resource speech challenge in 2015 web: http://www.lscp.net/persons/dupoux/bootphon/zerospeech2014/website/.

[5] L. Besacier, E. Barnard, A. Karpov, and T. Schultz. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85 – 100, 2014.

[6] N. F. Chen, S. Sivadas, B. P. Lim, H. G. Ngo, H. Xu, V. T. Pham, B. Ma, and H. Li. Strategies for vietnamese keyword search. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014*, pages 4121–4125, 2014.

[7] J. G. Fiscus and N. Chen. Overview of the NIST open keyword search 2013 evaluation workshop, 2013.

[8] T. Fraga-Silva, A. Laurent, J.-L. Gauvain, L. Lamel, V. B. Le, and A. Messaoudi. Improving data selection for low-resource STT and KWS. In *Proceedings of ASRU 2015*.

[9] M. Karafiát, F. Grézl, M. Hannemann, and J. Černocký. BUT neural network features for spontaneous Vietnamese in BABEL. In *Proceedings of ICASSP 2014*, pages 5659–5663. IEEE Signal Processing Society, 2014.

[10] H. Q. Nguyen, P. Nocera, E. Castelli, and V. L. Trinh. Large vocabulary continuous speech recognition for Vietnamese, an under-resourced language. In *Proceedings of SLTU 2008*.

[11] T. C. Nguyen and J. Chaloupka. *Phoneme set and pronouncing dictionary creation for large vocabulary continuous speech recognition of vietnamese.*, pages 394–401. Berlin: Springer, 2013.

[12] S. Tsakalidis, R. Hsiao, D. Karakos, T. Ng, S. Ranjan, G. Saikumar, L. Zhang, L. Nguyen, R. M. Schwartz, and J. Makhoul. The 2013 BBN vietnamese telephone speech keyword spotting system. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014*, pages 7829–7833, 2014.

[13] N. T. Vu and T. Schultz. Vietnamese large vocabulary continuous speech recognition. In *2009 IEEE Workshop on Automatic Speech Recognition & Understanding, ASRU 2009, Merano/Meran, Italy, December 13-17, 2009*, pages 333–338, 2009.

[14] T. T. Vu, D. T. Nguyen, C. M. Luong, and J.-P. Hosom. Vietnamese large vocabulary continuous speech recognition. In *Proceedings of INTERSPEECH 2005*.

[15] S. Xiong, W. Guo, and D. Liu. The vietnamese speech recognition based on rectified linear units deep neural network and spoken term detection system combination. In *The 9th International Symposium on Chinese Spoken Language Processing, Singapore, September 12-14, 2014*, pages 183–186, 2014.