INVESTIGATING TECHNIQUES FOR LOW RESOURCE CONVERSATIONAL SPEECH RECOGNITION

Antoine Laurent[1]*, Thiago Fraga-Silva*[1]*, Lori Lamel*[2]*, Jean-Luc Gauvain*[2]

[1]Vocapia Research, 28 rue Jean Rostand, 91400 Orsay, France
[2]CNRS/LIMSI, Spoken Language Processing Group, 91405 Orsay Cedex, France
{laurent,thfraga}@vocapia.com, {lamel,gauvain}@limsi.fr

ABSTRACT

In this paper we investigate various techniques in order to build effective speech to text (STT) and keyword search (KWS) systems for low resource conversational speech. Subword decoding and graphemic mappings were assessed in order to detect out-of-vocabulary keywords. To deal with the limited amount of transcribed data, semi-supervised training and data selection methods were investigated. Robust acoustic features produced via data augmentation were evaluated for acoustic modeling. For language modeling, automatically retrieved conversational-like Webdata was used, as well as neural network based models. We report STT improvements with all the techniques, but interestingly only some improve KWS performance. Results are reported for the Swahili language in the context of the 2015 OpenKWS Evaluation.

Index Terms— low-ressource languages, speech recognition, keyword spotting, conversational speech

1. INTRODUCTION

State-of-the-art speech recognition systems are usually trained on large amounts of acoustic and text data. It is well known that the system performance can considerably degrade when the amount of training data is limited (see for example [1, 6, 23]). Recently, there has been a growing interest in developing technologies for low-resource languages[1]. A variety of approaches have been proposed, from such as bootstrapping with models from well-resourced languages to complete self-discovery of linguistic units for unwritten languages (see for example [14, 20, 22]).

Generally speaking, low-resource languages are those with a low presence on the Internet, and more generally, limited textual resources especially in electronic form. There is in general little knowledge about the language, with very little or essentially no available audio data and small pronunciation dictionaries (if any).

In this paper, we investigate the use of six different approaches to develop speech-to-text (STT) and keyword search (KWS) systems for low-resource conversational speech. 1) *Semi-supervised training* (SST) [13, 24] is used here to cope with the lack of annotated acoustic data. In this case, the automatic transcripts are directly used for acoustic model training. 2) *Data selection* is also used to get relevant training data [4, 5]. In contrast to SST, here the goal is to select data for which accurate manual transcripts will be created.

We used acoustic features produced using bottleneck features extracted from deep-neural network (DNN) models [9, 15]. These features have been shown to outperform raw features in a number of reported works (see for example [9, 22]). Both, monolingual and multilingual bottleneck features are used. 3) *Data augmentation* was assessed in order to increase the system robustness. The method described in [15] is as follows. First, distorted copies of the original acoustic data are created by adding artificial noise and by varying the pitch. These copies and the original data are then used to train the DNNs for feature extraction.

4) *Documents automatically gathered from the Web* are used for language modeling [25]. Conversational-like queries were submitted to a search engine in order to retrieve texts that better match the conversational speech data. Additionally, 5) *neural network language models* (NNLMs) [21] are assessed and compared to standard backoff models in terms of STT and KWS performance.

Specifically for keyword search, we used the methods proposed in [11], which aim to increase the search performance on the out-of-vocabulary (OOV) keywords. Two approaches were combined, 6) *decoding with sub-word units and cross-word search.*

This work was performed in the context of the IARPA-Babel program [10]. The results are reported in this paper for the Swahili language. The investigated techniques were also applied to 6 other languages: Kurmanji, Tok-Pisin, Cebuano, Kazakh, Telugu and Lithuanian [2].

[1]http://www.mica.edu.vn/sltu2008 through sltu2014

[2]Swahili (IARPA-Babel202b-v1.0d), Kurmanji (IARPA-Babel205b-v1.0a), Tok-Pisin (IARPA-Babel207b-v1.0b), Cebuano (IARPA-Babel301b-v2.0b), Kazakh (IARPA-Babel302b-v1.0a), Telugu (IARPA-Babel303b-v1.0a), Lithuanian (IARPA-Babel304b-v1.0b)

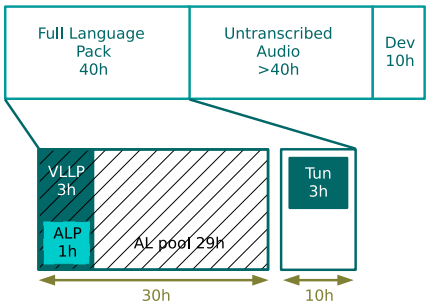

Fig. 1. Available data for system training and evaluation in the IARPA-Babel period OP2.

2. DATA AND METHODOLOGY

The STT systems were trained on data provided within the IARPA-funded Babel program [10]. In the OP2 program phase, systems were built for the 6 development languages (Cebuano, Kazakh, Kurdish, Lithuanian, Telugu and Tok-Pisin) and the surprise language (Swahili). As a total about 50 hours of transcribed conversational telephone speech data are provided for each language. This data is divided into different subsets which are illustrated in Figure 1.

2.1. Language Packs and tasks

The techniques investigated in this paper were assessed on the three main tasks of the IARPA-Babel OP2 phase. 1) The Full Language Pack (FLP) contained about 40 hours of transcribed data available for training. 2) The Very-Limited Language Pack (VLLP) is a 3-hour subset of the FLP (see Figure 1), and was selected by NIST to have about the same speech duration for each speaker. 3) The Active Learning (AL) task is defined as follows. A pre-defined 1-hour training set is used to build a bootstrap system. This system is used to decode an untranscribed 29-hour pool data set. Based on the decoding hypotheses and a selection criterion, 2 hours of data are selected from the data pool for manual transcription. An AL-based STT system is then built using the available 3 hours (initial 1h + selected 2h) of data. The AL based systems were created using the selection methods that we described in [5].

For the VLLP and AL tasks, only 3 hours of data are considered to be transcribed and the remainder of the pool data set (27 hours) could be used for semi-supervised training. Additional 40 to 50 hours of untranscribed data were available for each language and could be used for the three tasks. The data available from the Year-1 and Year-2 IARPA-Babel program (11 languages) could be used to develop multilingual models for the VLLP and AL tasks.

In addition to the manual transcriptions associated to the 3-hour training data, a textual corpus was available. It consists of texts collected from the Web (Wikipedia, subtitles and other webtexts). This Webdata was filtered, normalized and provided to the Babelon team by BBN. The size of the Webdata varies between languages from 5.7M to 49M words. For Swahili, about 16M words were available. For the OpenKWS15 Evaluation [18], the Webdata was allowed only for the VLLP and AL tasks. In this paper, we also used it for the FLP condition.

A 3-hour tuning set and a 10-hour development set were used to assess the systems (see Figure 1). In this paper, all the STT and KWS results are reported on the development set.

2.2. Recognition systems

For rapid development, all STT systems are based on graphemic pronunciation units and are built via a flat start. GMM/HMM and DNN/HMM based acoustic models (AM) were used in this study. The GMM/HMM models are triphone-based left-to-right 3-state HMMs [7]. The VLLP and AL models contain about 2k tied-states and 20k mixtures, while the FLP models contain about 10k tied-states and 150k mixtures. The DNN/HMM models have about 10M parameters, 4 hidden layers and a softmax output layer targeting HMM states. Position dependent and position independent AMs were used for word and sub-word decoding respectively.

Both, GMM and DNN based models are built using discriminative features produced with stack bottleneck DNNs and provided to the Babelon team by BUT [9]. Three sets of features were used in this work. The first set was extracted from a DNN trained on the original data, and the other two from DNNs trained on augmented data [15]. Both, noise addition and pitch variation were used to create up to eight copies of the original data.

Language models (LM) are trained with the LIMSI STK toolkit. The models are obtained via interpolation of component models estimated on the manual transcriptions of the audio training data and the webtexts. Both, backoff n-gram models and a feed-forward neural network models were used. The NNLMs contain 2 hidden layers and use a 12k-word shortlist softmax output [21].

2.3. Keyword search method

The keyword search method used in this work is described in [11]. Special attention is given to the detection of OOV keywords. First, a word and a sub-word consensus network (CN) are generated from decoding lattices [17]. Both CNs are searched to locate all sequences of words and sub-words that correspond to each keyword. Word boundaries are ignored during search. Substitution of pronunciation units was also considered during search, as inspired by [2]. However, instead of estimating phone confusion models, some few mappings were performed (e.g $/p/$ and $/b/$).

Keyword hits from word and sub-word CNs are combined based on time-codes. The keyword scores are then normalized and calibrated using the BBN KST normalization

System	Keyword search	All	IV	OOV
VLLP, + SST + Webdata	Word	0.436	0.458	0.268
	Sub-word (5-gram)	0.371	0.367	0.409
	Sub-word (6-gram)	0.375	0.369	0.419
	Sub-word (7-gram)	0.367	0.362	0.409
	4-way combination	**0.458**	0.461	0.456

Table 1. KWS results for a single VLLP system using word and sub-words units, as well as their combined outputs. ATWV is reported on all, in-vocabulary (IV) and out-of-vocabulary (OOV) keywords.

tool [12]. Decision about keeping or ignoring keyword hits is based on a defined threshold. In this work, sub-word units have up to 5, 6 or 7 letters (5-, 6- or 7-grams).

2.4. Performance metrics

STT performance is measured using the well-known word error rate (WER) metric. The KWS performance is reported here in terms of the Actual Term-Weighted Value (ATWV) [3, 18]. The keyword specific ATWV for the keyword k at a specific threshold t is computed as:

$$ATWV(k,t) = 1 - P_{FR}(k,t) - \beta P_{FA}(k,t) \qquad (1)$$

where P_{FR} and P_{FA} are respectively the probability of a false reject (miss) and false accept. The constant β mediates the trade off between false accepts and false rejects and is set to 999.9 for the OpenKWS15 Evaluation (see Section 5.2 of the KWS15 Keyword Search Evaluation Plan [19]).

3. EXPERIMENTS & RESULTS

3.1. Word and subword based keyword search

Table 1 shows in detail the effect of the KWS methods applied here for a VLLP system. As previously reported [11], the word decoding leads to overall good performances, especially on in-vocabulary (IV) keywords. A part of the OOV are also correctly located due to the combined use of cross-word search and graphemic mappings. Without the mappings, the overall ATWV score drops by about 0.01 absolute.

Sub-word decoding allows for a better detection of OOV keywords, but degrades the performance on IV keywords. The performance slightly varies with the maximum size of the subword units, the best ATWV being obtained for 6-gram subwords (0.375). The combination of the four system outputs leads to an ATWV absolute improvement of 0.02 compared to the word based system alone (from 0.436 to 0.458). In the remainder of this paper, all the ATWV results are given for the combined outputs of word and sub-word keyword hits.

Condition	Without Webdata		With Webdata	
	WER	ATWV	WER	ATWV
VLLP	58.5	0.419	52.4	0.454
AL	57.4	0.417	51.8	0.457
VLLP + SST	57.9	0.421	50.5	**0.458**
AL + SST	56.4	0.428	**50.2**	**0.458**

Table 2. Comparison of VLLP and AL based systems on different conditions. WER(%) and ATWV are given on the Swahili dev set.

3.2. VLLP and AL system comparison

A comparison of VLLP and AL-based systems was performed in different conditions, with and without semi-supervised acoustic model training, and with and without Webdata for language modeling. Table 2 presents the results obtained. All systems use purely multilingual features.

The system vocabulary has about 5k words without Webdata and 200k words with Webdata. Adding Webdata improves the WER performance from 5.6% to 7.4% absolute, and the ATWV on about 0.03–0.04 absolute for both, VLLP and AL systems. The SST brings an absolute improvement of about 0.6%–1.9% in terms of WER, but almost no gain in terms of ATWV. AL-based data selection obtains limited improvements over the VLLP baseline. For the best condition (with SST and Webdata), the WER gain is only 0.3% absolute and there is no different in terms of ATWV performance.

The same trends could be observed for SST with the other 6 IARPA-Babel OP2 languages tested (1% absolute WER gain). The gain with Webdata varies across languages, ranging from 0.3% to 7.4%. Improvements depend on the amount of Webdata available, the morphological aspects of the language, vocabulary coverage, etc. Data selection brought ATWV absolute improvements from 0.01 to 0.04 for the other OP2 languages as reported in [5].

3.3. Data augmentation and fine tuning

Data augmentation [15] and fine tuning of the DNN feature extractors were assessed for the VLLP systems. Three systems were built in similar conditions, using SST for acoustic modeling and Webdata for language modeling. Only the acoustic features changed. The results are shown in Table 3.

Data augmentation and fine tuning allows to produce robust features and generate more accurate acoustic models. The combined use of adding noise, varying the pitch and fine tuning allows to improve the WER performance on 3.8% absolute (from 50.5% to 46.7%). However, no improvements were observed in terms of ATWV with these features.

DNN bottleneck features	WER	ATWV
Multilingual (11 languages)	50.5	**0.458**
+ fine tuning + noise (x4)	47.0	0.458
+ pitch variation (x4)	**46.7**	0.453

Table 3. Comparison of techniques used to train DNNs for feature extraction. WER(%) and ATWV with VLLP systems using SST and Webdata.

3.4. FLP systems

The techniques previously assessed for the VLLP condition were also applied to the FLP systems. The FLP and VLLP are not directly comparable, since the VLLP systems were trained using multilingual features, while the FLP use monolingual features. SST for acoustic modeling did not bring STT or KWS improvements on FLP. A possible explanation is that the amount of untranscribed data is about the same of the transcribed set. In general, relatively larger amounts of data are required to obtain improvements with SST [16].

Results obtained with the other techniques are shown in Table 4. Webdata leads to absolute WER gains from 1.1% to 1.5% and ATWV gains of about 0.01 for the different systems. These gains are smaller than observed for VLLP. As more transcribed data is available for FLP (290k *vs.* 26k words), the LMs trained only on transcriptions are more accurate than the VLLP ones, therefore, taking less advantage from the additional Webdata.

Data augmentation via noise addition leads to WER absolute gains of about 1% and ATWV absolute gains of about 0.02. The use of a feed-forward NNLM gave additional gains in WER (about 1.4%-1.8% absolute), but no significant improvement in terms of KWS.

For comparison, a DNN/HMM based acoustic model was trained using the BUT bottleneck features (without data augmentation). The DNN based model outperforms the equivalent GMM one in terms of WER (from 41.5% to 39.3) and ATWV (from 0.520 to 0.539). In this work, we did not explore using data augmentation for DNN/HMM training, and did not assess the DNN/HMM with neural network language models. Both techniques can be expected to obtain additional gains to the best DNN system.

4. CONCLUSION

In this paper we explored various techniques aiming to improve the speech recognition and keyword search performances on low resource conversational speech. Keyword search was performed with special attention on OOV words. The combined use of word and sub-word decoding, cross-word search and graphemic mappings allowed to detect OOV keywords as well as in-vocabulary words, greatly increasing the overall KWS performance.

Condition	Without Webdata		With Webdata	
	WER	ATWV	WER	ATWV
FLP GMM/HMM	43.1	0.507	41.5	0.520
+ noise (4x)	42.0	0.524	40.5	0.538
+ NNLM	40.2	0.528	**39.1**	**0.540**
FLP DNN/HMM	41.0	0.514	39.3	0.539

Table 4. Comparison of the different techniques used on for the FLP system. WER(%) and ATWV on the Swahili dev set.

Automatic training data selection was assessed. For Swahili, data selection obtained limited ATWV improvements over a strong baseline, the VLLP condition. For the other 6 OP2 languages assessed (see [5]), data selection obtained absolute ATWV improvements between 0.01 and 0.04.

Semi-supervised acoustic model training obtained WER absolute gains between 1.0% and 1.5% when only 3 hours of transcribe data were available (VLLP and AL tasks). No gain was observed for the 40-hour training condition. SST did not lead to improvements in KWS performance in any of the conditions assessed.

Data augmentation for feature extraction was also evaluated. With the GMM/HMM acoustic models, data augmentation leads to a significant WER reduction over the baseline for two different conditions. Absolute gains of 3.8% and 1.0% were observed for VLLP (3-hour training set) and FLP (40-hour training set) respectively. ATWV improvements were also observed for the FLP (0.02 absolute).

In terms of language modeling, we explored the use of Webdata and neural network models. Webdata leads to STT and KWS improvements in all cases. It is important to note that a careful pre-processing of the texts is required. The absolute gains vary depending on the language and the amount of transcribed data. The largest gains are obtained for the 3-hour condition (7.4% WER and 0.03 ATWV absolute). The use of feed-forward NNLMs led to a significant WER reduction for FLP (1.4% absolute), but did not lead to improvements in KWS performance.

5. ACKNOWLEDGMENTS

We would like to thank our Babelon partners for sharing resources (BUT for the bottle-neck features and BBN for the Webdata), and Grégory Gelly for providing the VADs.

This research was in part supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense US Army Research Laboratory contract number W911NF-12-C-0013. 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.

6. REFERENCES

[1] L. Besacier, E. Barnard, A. Karpov, T. Schultz, "Automatic speech recognition for under-resourced languages : A survey," *Speech Communication Journal*, vol. 56, pp. 85-100, January 2014.

[2] G. Chen, O. Yilmaz, J. Trmal, D. Povey, and S. Khudanpur, "Using proxies for OOV keywords in the keyword search task," *IEEE ASRU*, 2013.

[3] J. G. Fiscus, J. Ajot, J. S. Garofolo, G. Doddington, "Results of the 2006 spoken term detection evaluation," *ACM SIGIR*, pp. 51–55, 2007.

[4] T. Fraga-Silva, J.-L. Gauvain, L. Lamel, A. Laurent, V.-B. Le, A. Messaoudi, "Active Learning based data selection for limited resource STT and KWS," *ISCA Interspeech*, 2015.

[5] T. Fraga-Silva, A. Laurent, J.-L. Gauvain, L. Lamel, V.-B. Le, A. Messaoudi, "Improvind data selection for low-resource STT and KWS," *IEEE ASRU*, 2015.

[6] M. Gales, K. Knill, A. Ragni, S. Rath, "Speech recognition and keyword spotting for low resource languages: BABEL project research at CUED," *SLTU*, 2014.

[7] J. L. Gauvain, L. Lamel and G. Adda, "The LIMSI broadcast news transcription system," *Speech Communication*, vol. 37, no. 1-2, pp. 89–108, 2002.

[8] G. Gelly, J. L. Gauvain. "Minimum Word Error Training of RNN-based Voice Activity Detection," *ISCA Interspeech*, 2015.

[9] F. Grézl, M. Karafiát, "Semi-Supervised bootstrapping approach for neural network feature extractor training," *IEEE ASRU*, pp. 470–475, 2013.

[10] M. Harper, "IARPA Babel Program," `http://www.iarpa.gov/index.php/research-programs/babel`

[11] W. Hartmann, V. B. Le, A. Messaoudi, L. Lamel, J. L. Gauvain, "Comparing decoding strategies for subword-based keyword spotting in low-resourced languages," *ISCA Interspeech*, 2014.

[12] D. Karakos, R. Schwartz, S. Tsakalidis, L. Zhang, S. Ranjan, T. Ng, R. Hsiao, G. Saikumar, I. Bulyko, L. Nguyen, J. Makhoul, F. Grezl, M. Hannemann, M. Karafiat, I. Szoke, K. Vesely, L. Lamel, V.B. Le "Score normalization and system combination for improved keyword spotting," *IEEE ASRU*, pp. 210–215, 2013.

[13] T. Kemp, A. Waibel, "Unsupervised training of a speech recognizer: recent experiments," *ESCA Eurospeech*, pp. 2725–2728, 1999.

[14] T. Kempton, R. Moore. "Discovering the phoneme inventory of an unwritten language: A machine-assisted approach," *Speech Communication Journal*, vol. 56, pp. 152-166, January 2014.

[15] M. Karafiát, F. Grézl, L. Burget, I. Szöke and J. Cernocký, "Three ways to adapt a CTS recognizer to unseen reverberated speech in BUT system for the ASpIRE challenge," *ISCA Interspeech*, 2015.

[16] L. Lamel, J.L. Gauvain, G. Adda, "Lightly supervised and unsupervised acoustic model trainings" *Computer Speech and Language*,vol. 16, pp. 115–129, 2002.

[17] L. Mangu, E. Brill, A. Stolcke, "Finding consensus in speech recognition: Word error minimization and other applications of confusion networks," *Computer, Speech and Language*, 14(4):373-400, 2000.

[18] NIST Open Keyword Search Evaluation (OpenKWS), `http://www.nist.gov/itl/iad/mig/openkws.cfm`

[19] KWS15 Keyword Search Evaluation Plan, `http://www.nist.gov/itl/iad/mig/upload/KWS15-evalplan-v05.pdf`

[20] S. Stücker, M. Müller, Q.B. Nguyen, A. Waibel. "Training time reduction and performance improvements from multilingual techniques on the BABEL ASR task," *IEEE ICASSP*, pp. 6374–6378, 2014.

[21] H. Schwenk, J.L. Gauvain, "Connectionist Language Modeling for Large Vocabulary Continuous Speech Recognition," *IEEE ICASSP*, pp.765-768, Mu 2002.

[22] N. T. Vu, F. Metze and T. Schultz. "Multilingual bottleneck features and its application for under-resourced languages," *SLTU*, Cape Town, South Africa, May 2012.

[23] N. T. Vu, D. Imseng, D. Povey, P. Motlicek, T. Schultz, H. Bourlard, "Multilingual deep neural network based acoustic modeling for rapid language adaptation," *IEEE ICASSP*, 2014.

[24] G. Zavaliagkos and T. Colthurst, "Utilizing Untranscribed Training Data to Improve Performance," *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, pp. 301-305, 1998.

[25] L. Zhang, D. Karakos, W. Hartmann, R. Hsiao, R. Schwartz and S. Tsakalidi, "Enhancing Low Resource Keyword Spotting with Automatically Retrieved Web Documents," *ISCA Interspeech*, 2015