# Developing STT and KWS systems using limited language resources

*Viet-Bac Le*[1], *Lori Lamel*[2], *Abdel Messaoudi*[1,2], *William Hartmann*[2],
*Jean-Luc Gauvain*[2], *Cécile Woehrling*[1], *Julien Despres*[1], *Anindya Roy*[2]

[1]Vocapia Research, 28 rue Jean Rostand, 91400 Orsay, France
[2]CNRS/LIMSI, Spoken Language Processing Group, 91403 Orsay cedex, France
{levb,abdel,woehrling,despres}@vocapia.com, {lamel,hartmann,gauvain,roy}@limsi.fr

## Abstract

This paper presents recent progress in developing speech-to-text (STT) and keyword spotting (KWS) systems for the 2014 IARPA-Babel evaluation. Systems have been developed for the limited language pack condition for four of the five development languages in this program phase: Assamese, Bengali, Haitian Creole and Zulu. The systems have several novel characteristics that support rapid development of KWS systems. On the STT side different acoustic units are explored based on phonemic or graphemic representations, and system combination is used to improve STT performance. The acoustic models are trained on only 10 hours of speech data with manual transcriptions, completed with unsupervised training on additional untranscribed data. Both word and subword units (morphologically decomposed, syllables, phonemes) are used for KWS. The KWS systems are based on the multi-hypotheses produced by a consensus network decoding or searching word lattices. The word error rates of the individual STT systems are on the order of 50-60%, and the KWS systems obtain Maximum Term Weighted Values ranging from 30-45% for all keywords (in-vocabulary and out-of-vocabulary (OOV)). Sub-word units are shown to be successful at locating some of the OOV keywords, and system combination improves system performance.

**Index Terms**: STT, KWS, semi-supervised training, lattice, consensus network, sub-word lexical units, Morfessor

## 1. Introduction

This paper describes our recent research carried out within the context IARPA Babel program and aimed at developing speech-to-text (STT) and keyword spotting (KWS) systems for low-resourced languages. The IARPA Babel program [1] aims to support the rapid development of speech technologies for effective word-based search in varied audio data in a variety of languages. The languages are chosen to present challenges at different levels (written scripts & writing conventions, phonological, morphological, dialectal). For each targeted language the program provides a build pack, which contains transcribed speech data, a pronunciation dictionary and a brief descriptive "Language Specific Peculiarities" document [2]. The techniques developed in the program on what are referred to as development languages are also applied a surprise language as part of the NIST Open Keyword Search Evaluation (OpenKWS13,OpenKWS14) [3, 4].

During the base phase of the IARPA-funded BABEL program [5], we built full language pack [3, 4] (FullLP) STT systems for five languages: 4 development languages (Cantonese, Pashto, Tagalog, Turkish); and 1 surprise language: (Vietnamese). The FullLP STT systems, trained on all available resources, for these languages obtained CER/WER ranging from 37.8% (for Cantonese) to 51.4% (for Pashto). We built a limited language pack (LLP) system only for Vietnamese.

This year our effort has focused on the LLP condition and STT and KWS systems have been developed for four languages. A variety of ideas have been explored with the aim of improving both STT and KWS. Different acoustic units are explored for STT based on phonemic or graphemic representations, and system combination is used to improve STT performance. The acoustic models are trained on only about 10 hours of manually transcribed speech data complemented with unsupervised training on additional untranscribed data [6], and then adapted to the supervised portion. We also report on some initial limited experiments with crosslingual and multilingual modeling. For KWS, both word and subword units (morphologically decomposed, syllables, automatically discovered) are explored. The automatically discovered lexical units [7] are found to be effective for detecting OOV keywords. The KWS systems are based on the multi-hypotheses produced by a consensus network (CN) decoding or searching word lattices. In all KWS experiments presented in this paper, the no test audio re-use (NTAR) condition is used. That means the KWS system does not re-process the test audio after keywords are provided [3].

## 2. Experimental setup

### 2.1. Data

All data used to train the STT systems were provided in the context of the IARPA-funded BABEL program. Five development languages were targeted in this second program phase: Assamese, Bengali, Haitian Creole, Lao, and Zulu. For the LLP condition about 10 hours of transcribed speech are provided from 60 speakers. The data contain both conversational speech ($\sim$80%) and scripted speech ($\sim$20%).[1] In this work, the language packs used are: Assamese (iarpa_babel_op1_103), Bengali (iarpa_babel_op1_102), Haitian Creole (iarpa-babel201b-v0.2b) and Zulu (iarpa-babel206b-v0.1d). As a total about 60 hours of transcribed speech are provided for each language for the FullLP condition, the remaining audio could be used in an unsupervised manner to complement the transcribed speech for the LLP condition. The speech data were collected for 3 dialects in Assamese, Bengali, and Creole, and one dialect for Zulu.

[1] A small proportion of the recordings are wide-band for some of the languages. For the moment these were simply filtered to telephone band and pooled with the remaining data.

Table 1: *# INV and OOV keywords in dev & eval keyword lists.*

| KWD | Assamese | | Bengali | | Creole | | Zulu | |
|---|---|---|---|---|---|---|---|---|
| set | inv | oov | inv | oov | inv | oov | inv | oov |
| dev | 1436 | 564 | 1373 | 627 | 1680 | 320 | 771 | 1229 |
| eval | 2405 | 648 | 2369 | 684 | 2019 | 677 | 2165 | 985 |

Results are reported using the official development (dev) and evaluation (eval) lists provided by NIST for the 2014 IARPA-Babel evaluation. From the lists, in-vocabulary (INV) and out-of-vocabulary (OOV) sublists are extracted relative to the decoding vocabulary of the respective STT systems: a keyword is considered as OOV keyword if it contains at least one word which is OOV. Otherwise, it is considered as an INV keyword. The number of INV and OOV keywords in the dev and eval data for each language are shown in Table 1.

### 2.2. Language characteristics

The languages share or differ with respect to several characteristics. The Zulu language is particularly challenging having a complex morphology (agglutinative, with extensive inflection), many borrowed English words, being tonal and having clicks among the phonological sound inventory. In contrast, Haitian Creole has a minimal derivational morphology and shares many characteristics with the French language, with a simpler grapheme-to-phoneme correspondence.

The Bengali and Assamese languages basically share the same written script. The two differences observed in the Babel transcripts are: *ra*, which is represented slightly differently in Bengali (র) and Assamese (ৰ); and an additional letter in the Assamese script (ৱ) corresponding to the sound 'wo' which is absent in Bengali. In both Assamese and Bengali scripts, there are distinct characters representing alveolar and dental versions of [t, d, n]. However, according to the Appen language description, in spoken Assamese this place distinction is no longer made. Note that these scripts are abugidas (alphasyllabary) where vowel graphemes are almost always realized as diacritics attached to consonant graphemes if present inside a word. There are 11 vowels and 35 (36 for Assamese) consonants in the scripts. Both Creole and Zulu are written using Latin characters, with 28 and 40 graphs respectively.

At the phonological level, Bengali and Assamese are also quite similar, with 33 or 30 consonants, 10 or 9 oral vowels and 9 nasal vowels. Bengali differentiates three alveolar and dental phonemes [t, d, n], a distinction not made in Assamese, and also has a schwa not present in Assamese. In the Appen dictionary for Assamese the two graphemes are mapped to a single alveolar phoneme. Haitian Creole has 20 consonants and 12 vowels (including 4 nasal vowels and one diphthong), where as Zulu has 28 consonants, 7 vowels 9 clicks. The clicks differ in terms of place of articulation (dental, post-alveolar, alveolar lateral) and manner of articulation (+/-voicing, +/-aspiration).

In order to explore complementary STT systems for combination, and to avoid having unseen or very few tokens of rare phones or graphemes, the acoustic unit sets for some languages were reduced by merging some closely related units. For example, smaller phone sets for Bengali and Assamese were created by merging aspirated and unaspirated plosives and nasal/nonnasal vowels and splitting complex consonants (diphthongs, afficates) into a sequence of phonemes.

### 2.3. Baseline recognition systems

For rapid development of STT systems and in the context of IARPA-Babel program, all phonemic and graphemic STT systems are built using a flat start. The acoustic models are tied-state, left-to-right 3-state HMMs with Gaussian mixture observation densities (typically 32 components) [8]. The triphone-based models are word position-dependent. Initial sets of context-independent models are first estimated on the transcribed training data using a 42-component feature vector [9, 10]. Larger acoustic models covering more context-dependent units are successively estimated using the same features. The final models are trained using discriminative features produced with a stacked bottle-neck multilayer perceptron and provided to the Babelon team by BUT [11]. Our systems also use the BBN voice activity detection [12].

Table 2: *Transcribed data characteristics.*

| Language | #words | vocab. | % 3g-hits | 3g-ppl | % oov |
|---|---|---|---|---|---|
| Assamese | 73.4k | 8.8k | 15.0 | 241.0 | 8.46 |
| Bengali | 81.9k | 9.7k | 15.1 | 248.8 | 8.78 |
| Creole | 103.1k | 5.7k | 24.0 | 147.6 | 4.39 |
| Zulu | 68.3k | 15.9k | 15.9 | 239.9 | 21.8 |

Language model training is performed with LIMSI STK toolkit which allows model training without any pruning or cutoff [10], thus keeping all information in the training data. The number of training words, vocabulary size, 3-gram hit rates and perplexity of the development data with 3-gram LMs are shown in Table 2.

Decoding is carried out in a single-pass, using case sensitive language models for all languages. Word decoding generates a word lattice followed by consensus decoding with a 3-gram or 4-gram and with/without pronunciation probabilities.

### 2.4. Performance Metrics

STT system performance is measured with the common metric Word Error Rate (WER) which is defined by a function of insertion, deletion and substitution rates.

For KWS, the Maximum Term-Weighted Value (MTWV) and Actual Term-Weighted Value (ATWV) are defined as the measures of interest for IARPA-Babel program [4]. ATWV was also used in the NIST 2006 Spoken Term Detection evaluation [13]. The keyword specific ATWV for the keyword $k$ at a specific threshold $t$ can be computed by:

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

where $P_{FR}$ and $P_{FA}$ are the probability of a false reject (miss) and false accept, respectively. The constant $\beta$, set to a value of 999.9, mediates the trade off between false accepts and false rejects. The MTWV represents the maximum score over the range of all possible values of $t$ (score decision threshold). Since the experiments reported here are performed only on dev data, we report the KWS performance in terms of MTWV. This metric weights all keywords equally regardless of its frequency. Missing a single occurrence of a rare word can affect the final score as much as missing a more common word dozens of times. This is why substantial effort is devoted to detecting OOV keywords.

## 3. Developing limited resourced STT

### 3.1. Investigating different phonesets

Table 3 reports the WER for the phonemic and graphemic based STT systems for the four languages. For all systems, three units correspond to silence, breath noise and fillers. All models are trained using features (version v3) provided to the Babelon team by BUT [11, 14]. For each language, the column 'Sup' corresponds to the baseline models estimated using only

Table 3: *Word error rates of STT systems.*

| Language | Acoustic Unit | % WER Sup | Semi-sup | SysComb |
|---|---|---|---|---|
| Assamese | 49 phones | 58.6 | 57.6 | 57.1 |
| | 29 phones | 59.3 | 58.2 | |
| | 49 graphs | 58.5 | 57.4 | |
| Bengali | 55 phones | 59.2 | 58.8 | 58.4 |
| | 34 phones | 59.5 | 58.8 | |
| | 48 graphs | 60.6 | 59.4 | |
| Creole | 35 phones | 52.0 | 50.8 | 50.7 |
| | 31 graphs | 52.3 | 51.4 | |
| Zulu | 40 phones | 65.4 | 65.2 | 64.6 |
| | 26 graphs | 65.2 | 65.1 | |

Table 4: *WER (in %) of common Bengali-Assamese STT system.*

| System PLP+f0, ML-SAT | monolingual AM+LM | multilingual AM+LM | mono AM multi LM |
|---|---|---|---|
| Bengali 28ph | 72.0 | 75.4 | 72.3 |
| Bengali 45gr | 72.7 | 76.1 | 72.4 |
| Assamese 28 ph | 70.7 | 76.8 | 72.3 |
| Assamese 45 gr | 71.0 | 75.7 | 71.8 |

Table 5: *MTWV results (INV keywords only) for different KWS methods on fused dev+eval keyword lists.*

| Language | Unit | WER (%) | MTWV (INV) lat | MTWV (INV) CN | SysComb 2-way | SysComb 3-way |
|---|---|---|---|---|---|---|
| Assamese | 49 ph | 57.6 | 0.393 | 0.401 | 0.422 | 0.435 |
| | 29 ph | 58.2 | 0.399 | 0.401 | | |
| | 49 gr | 57.4 | 0.376 | 0.382 | | |
| Bengali | 55 ph | 58.9 | 0.396 | 0.396 | 0.421 | 0.429 |
| | 34 ph | 58.8 | 0.398 | 0.404 | | |
| | 48 gr | 59.4 | 0.373 | 0.367 | | |
| Creole | 35 ph | 51.0 | 0.513 | 0.501 | 0.529 | |
| | 31 gr | 51.5 | 0.513 | 0.500 | | |
| Zulu | 40 ph | 65.4 | 0.343 | 0.385 | 0.415 | |
| | 26 gr | 65.2 | 0.344 | 0.379 | | |

the transcribed LLP data for supervised acoustic model training. The WERs range from 52% for Creole to just under 60% for Assamese and Bengali, to 65% for Zulu. Similar performances are obtained with the phonemic and graphemic systems, and slightly better results with the grapheme based system for Zulu. During system development models based on the reduced phone set for Bengali outperformed the full phone set models, but this difference reversed with the most recent BUT features.

Given the similarity of the Bengali and Assamese languages, we explored building a common STT system. We defined a common set of 28 phones, where the aspirated phones were merged with their unaspirated counterparts, the nasal and oral vowels merged, and diphthongs split into a vowel-glide sequence. We retained the schwa present in Bengali but absent in Assamese. For the multilingual system, the AMs are multilingual (trained by pooling of data from both languages) but both multilingual and monolingual LMs were estimated. Recognition results (WER) using PLP+F0 features with ML-SAT training are reported in Table 4. As these first results were not promising, this research direction was put on hold.

### 3.2. Semi-supervised training

Unsupervised training methods have been used for a variety of tasks, but have only relatively recently been successfully applied to conversational speech data with high error rates [6]. Using automatic transcripts produced by a different speech recognizer can also give the added benefit of cross-system adaptation. In this work we have used automatic transcripts (without thresholding on confidence scores) provided to the Babelon team by BBN. The column 'Semi-sup' in Table 3 reports results using AMs trained by pooling the supervised data with the remaining data (about 35 hours per language) using the automatic transcripts, and adapting the resulting models with the supervised data. Using the additional data is seen to reduce the WER by around 1% for most languages and acoustic unit sets. A notable exception Zulu for which almost no improvement is observed for the two systems. The rightmost column reports the results of Rover-based [15] combination of the individual systems. System combination is seen to reduce the WER by 0.1% (Creole) to 0.5% (Zulu) over the best single system. It is interesting to note that system combination gives the largest gain for Zulu, for which semi-supervised training was the least effective.

## 4. Keyword spotting

### 4.1. Search method

Two keyword search methods were investigated in our experiments: lattice based and consensus network (CN) based search, both using an exact match. For the lattice based keyword search, all possible word n-grams (n is limited to 5 in our experiments) with timecode are generated for each lattice. N-grams with near or overlapping timecodes are fusioned to reduce the index size. Then each keyword is matched against all n-grams either ignoring or keeping word boundaries. In CN-based keyword search is performed on a consensus network which created from a word/sub-word lattice [16]. For each keyword, the CN is searched to locate all sequences of word/sub-words either keeping or ignoring word boundaries. The KW hits are combined based on time-codes and ranked using a geometric mean score [17]. The raw KW scores from both search methods are further normalized and calibrated by BBN's KST normalization tool [18, 19]. Score normalization produces significantly better results and approaches the ATWV results to the MTWV ones.

Table 5 compares MTWV results on INV keywords with both search methods, keeping word boundaries. The lattice-based systems are seen to give similar performance to CN-based systems for Assamese and Bengali, to obtain slightly better performance for Haitian Creole and worse performance for Zulu.

Concerning computation time, indexing is much slower for the lattice-based method than the CN-based method, but searching is faster. For example, the total indexing time on dev+eval data (15 hours) for Bengali is about 16 hours for the CN-based and 112 hours for lattice-based method. The total search time with dev+eval keyword list is about 10 hours and 2.5 hours for the CN- and lattice-based methods respectively. Another drawback of lattice-based search is that all possible n-grams need to be stored and indexed, which requires more disk space than the CN-based search (indexing by word slot).

An important advantage of the CN-based search is that decoding can output word/sub-word sequences that were never observed in the LM training texts and do not exist in the original lattice. This is useful for detecting rare keywords (all words are INV but the sequence is not in training) or OOV keywords using sub-word units as discussed in Section 4.4.

### 4.2. Combining phonetic and graphemic KWS systems

The different KWS are combined by taking a weighted average of the raw (unnormalized) scores of keyword hits with

Table 6: *MTWV results with/out* WB *in keywords.*

| Language | Unit | keep WB | | | remove WB | | |
|---|---|---|---|---|---|---|---|
| | | all | IV | OOV | all | INV | OOV |
| Assamese | 49 ph | 0.306 | 0.393 | 0 | 0.317 | 0.396 | 0.041 |
| | 29 ph | 0.311 | 0.399 | 0 | 0.321 | 0.401 | 0.041 |
| | 49 gr | 0.293 | 0.376 | 0 | 0.304 | 0.376 | 0.053 |
| Bengali | 55 ph | 0.302 | 0.396 | 0 | 0.312 | 0.396 | 0.045 |
| | 34 ph | 0.304 | 0.398 | 0 | 0.314 | 0.398 | 0.045 |
| | 48 gr | 0.284 | 0.373 | 0 | 0.298 | 0.373 | 0.062 |
| Creole | 35 ph | 0.448 | 0.513 | 0 | 0.427 | 0.475 | 0.103 |
| | 31 gr | 0.449 | 0.513 | 0 | 0.429 | 0.473 | 0.121 |
| Zulu | 40 ph | 0.186 | 0.343 | 0 | 0.226 | 0.379 | 0.048 |
| | 26 gr | 0.186 | 0.344 | 0 | 0.209 | 0.343 | 0.053 |

near or overlapping timecodes. The combined scores are then normalized and calibrated by KST normalization [18, 19]. The rightmost column of Table 5 gives 2-way (full and reduced phone sets with equal weight) and 3-way (all) systems for Bengali and Assamese. On the INV keywords an absolute MTWV improvement of about 2% with the 2-way system, with an additional gain of 0.8–1.3% combining this with the respective graphemic systems (with weight 0.3). Table 5 also shows combination results for phonetic and graphemic based systems for Haitian Creole and Zulu. For both languages combination improves the MTWV by about 3% absolute.

### 4.3. Influence of word boundaries

During the development of lattice-based search, we observed that removing word boundaries (WB) allows a significant portion of the OOV keywords to be detected with little effect on the INV keywords. The same observation is also observed for CN-based search [7]. Table 6 presents MTWV results keeping or ignoring the WB. Except for Creole, the MTWV results for INV keywords are nearly the same but removing the WB successfully detects some OOV keywords. For Haitian Creole, all clitics (frequent in the vocabulary) are separated by underscore from the word in the transcripts [2]. When removing WB, the underscore and white space are systematically removed, so the clitics are glued to the original words. This approach seems to increase the confusion for Creole more than for other languages, which may be why the performance of INV keyword detection is worse when WB are ignored. In contrast OOV keyword detection is much better than for the other languages.

### 4.4. KWS using sub-word units

Different types of sub-word units (Morfessor-based, syllable, phoneme) are investigated with the aim of detecting OOV keywords. The first set of units are derived using Morfessor, a tool for unsupervised morphological decomposition [20, 21]. Given the list of words in the training texts with their frequencies, Morfessor learns a set of morphological units that are then used to segment the training texts and the keyword list. The second unit type corresponds to syllables. The dictionaries provided by Appen include the syllable segmentation for the words in the training data. To determine segmentations for OOV words, a syllable-based trigram LM is used to determine the segmentation with the highest likelihood for each word. To ensure that a segmentation is possible, all individual characters are included in the LM. The training corpus and keyword list are segmented into syllables. For the phoneme units, the Sequitur G2P converter [22] is employed to learn and generate pronunciations for OOV words.

The original word lattices from the word-based STT systems are used, and the words in each lattice are simply decomposed into sub-word sequences, from which a sub-word CN is built. In fact, the same technique was employed for the syllable-based languages such as Chinese and Vietnamese where the STT results are usually measured in terms of character error rate (CER) or syllable error rate (SLER) instead of WER. Instead of using the word CN, the character- or syllable-based CN can improve the CER or SLER [23, 24].

Table 7: *Subword units for Bengali KWS with word-to-subword CN decoding. Results using dev keywords.*

| System | MTWV | | |
|---|---|---|---|
| | all | inv | oov |
| s0: word | 0.300 | 0.437 | 0 |
| s1: syllable | 0.297 | 0.400 | 0.076 |
| s2: morfessor | 0.326 | 0.434 | 0.094 |
| s3: phone | 0.214 | 0.292 | 0.045 |
| s1+s2+s3 | 0.322 | 0.417 | 0.118 |
| s0+s1+s2+s3 | 0.333 | 0.437 | 0.118 |

Table 7 provides the MTWV results for the Bengali system using the dev keyword list. All subword units lead to some of the OOV keywords are detected. The Morfessor-based system seems to work better than the others. Combining the hits from all systems (word and sub-word) results in the best KWS performance for both INV and OOV keywords.

## 5. Conclusions and Perspectives

The STT and KWS systems developed for four low resourced languages in the context of the IARPA Babel program have been described. Several novel research directions have been explored and successfully applied to the KWS task. On the STT side different acoustic units were considered and it was found that graphemic representations give comparable performance to phonemic ones. For KWS both word and subword units (morphologically decomposed, syllables, phonemes) were used, and KWS based on multi-hypotheses produced by a consensus network decoding or searching word lattices were investigates. Sub-word units were shown to be successful at locating some of the OOV keywords, and system combination improves KWS performance. The reported results on the development data will be completed with those on the evaluation data when available. The same techniques will be applied to the surprise language (currently unknown) as part of the NIST Open Keyword Search 2014 Evaluation [3] and results will be reported in the final version of this paper.

## 6. Acknowledgements

We would like to thank our Babelon partners for sharing resources (BUT for the bottle-neck features, BBN for VAD, auto-transcripts, score normalization tool and NWU for a Zulu pronunciation dictionary.

This research was 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.

# 7. References

[1] Harper, M., "IARPA Babel Program," http://www.iarpa.gov/Programs/ia/Babel/babel.html

[2] www.nist.gov/itl/iad/mig/upload/IARPA_Babel_Performer-Specification-08262013.pdf

[3] NIST Open Keyword Search 2014 Evaluation (OpenKWS14), www.nist.gov/itl/iad/mig/openkws14.cfm

[4] NIST, "The OpenKWS14 Evaluation Plan, v11," December 2013, http://www.nist.gov/itl/iad/mig/upload/KWS14-evalplan-v11.pdf

[5] Sainath, T., Kingsbury, B., Metze, F., Morgan, N., Tsakalidis, S., "An Overview of the Base Period of the Babel Program," SLTC Newsletter, November 2013. http://www.signalprocessingsociety.org/technical-committees/list/sl-tc/spl-nl/2013-11/BabelBaseOverview

[6] Hsiao, R., Ng, T., Grézl, F., Karakos, D., Tsakalidis, S., Nguyen, L., and Schwartz, R., "Discriminative semi-supervised training for keyword search in low resource languages," ASRU 2013, pp. 440–445, Olomouc, Czech Republic, December 2013.

[7] Hartmann, W., Le, V.B., Messaoudi, A., Lamel, L., and Gauvain, J.L., "Comparing Decoding Strategies for Subword-based Keyword Spotting in Low-Resourced Languages,' submitted to Interspeech 2014, Singapore, September 2014.

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

[9] Lamel, L., Gauvain, J.L., Le, V.B., Oparin, I., and Meng, S., "Improved models for Mandarin speech-to-text transcription," ICASSP 2011, pp. 4660–4663, Prague, Czech Prague, May 2011.

[10] Lamel, L., Courcinous, S., Despres, J., Gauvain, J.L., Josse, Y., Kilgour, K., Kraft, F., Le, V.B., Ney, H., Nußbaum-Thom, M., Oparin, I., Schlüter, R., Schultz, T., Fraga Da Silva, T., Stüker, S., Sundermeyer, M., Vieru, B., Vu. N.T., Waibel, A., and Woehrling, C., "Speech Recognition for Machine Translation in Quaero," IWSLT 2011, pp. 121–128, San Francisco, CA, USA, December 2011.

[11] Grézl, F., and Karafiát, M., "Semi-Supervised Bootstrapping Approach For Neural Network Feature Extractor Training," ASRU 2013, pp. 470–475, Olomouc, Czech Republic, December 2013.

[12] Ng, T., Zhang, B., Nguyen, L., Matsoukas, S., Zhou, X., Mesgarani, N., Veselý, K., and Matejka, P., "Developing a Speech Activity Detection System for the DARPA RATS Program," Interspeech 2012, pp. 1969–1972, Portland, OR, USA, September 2012.

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

[14] Karafiát, M., Grézl, F., Hannemann, M., and Černocký, J., "BUT Neural Network Features for Spontaneous Vietnamese in Babel," ICASSP 2014, Florence, May 2014, to appear.

[15] Fiscus, J., "A Post-Processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER)," ASRU 1997, pp. 347–354, Santa Barbara, CA, USA, December 1997.

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

[17] Mamou, J., Ramabhadran, B., and Siohan, O., "Vocabulary independent spoken term detection," ACM SIGIR 2007, pp. 615–622, Amsterdam, The Netherlands, July 2007.

[18] Karakos, D., Schwartz, R., Tsakalidis, S., Zhang, L., Ranjan, S., Ng, T., Hsiao, R., Nguyen, L., Grézl, F., Hannemann, M., Karafiát, M., Szöke, I., Veselý, K., Lamel, L., and Le, V.B., "Score Normalization and System Combination for Improved Keyword Spotting," ASRU 2013, pp. 210–215, Olomouc, Czech Republic, December 2013.

[19] Karakos, D., Bulyko, I. Schwartz, R., Tsakalidis, S., Nguyen, and L., Makhoul, J., "Normalization of Phonetic Keyword Search Scores," to appear in ICASSP 2014, Florence, May 2014.

[20] Creutz, M., and Lagus, K., "Unsupervised Morpheme Segmentation and Morphology Induction from Text Corpora using Morfessor 1.0," Computer and Information Science, Technical Report A81, 27 pages, Helsinki University of Technology, 2005.

[21] Creutz, M., Hirsimäki, T., Kurimo, M., Puurula, A., Pylkkönen, J., Siivola, Varjokallio, M.,Arisoy, E., Saraclar, M., and Stolcke, A., "Analysis of Morph-Based Speech Recognition and the Modeling of Out-of-Vocabulary Words Across Languages," *ACM Transactions on Speech and Language Processing* **5**(1.3), December 2007.

[22] Bisani, M., and Ney, H., "Joint-Sequence Models for Grapheme-to-Phoneme Conversion," Speech Communication, vol. 50, no. 5, pp. 434–451, May 2008.

[23] Fu, Y.S., Pan, Y.C., and Lee, L.S., "Improved Large Vocabulary Continuous Chinese Speech Recognition by Character-Based Consensus Networks," ISCSLP 2006, pp. 422–434, Singapore, December, 2006.

[24] Le, V.B., Seng, S., Besacier, L., and Bigi, B., "Word/sub-word lattices decomposition and combination for speech recognition," ICASSP 2008, pp. 4321–4324, Las Vegas, NV, USA, March–April, 2008