QUANTIFYING THE VALUE OF PRONUNCIATION LEXICONS FOR KEYWORD SEARCH IN LOW RESOURCE LANGUAGES

Guoguo Chen, Sanjeev Khudanpur, Daniel Povey, Jan Trmal, David Yarowsky and Oguz Yilmaz

Center for Language and Speech Processing, and Human Language Technology Center of Excellence
Johns Hopkins University, Baltimore, MD 21218, USA
radical@clsp.jhu.edu

ABSTRACT

This paper quantifies the value of pronunciation lexicons in large vocabulary continuous speech recognition (LVCSR) systems that support keyword search (KWS) in low resource languages. State-of-the-art LVCSR and KWS systems are developed for conversational telephone speech in Tagalog, and the baseline lexicon is augmented via three different grapheme-to-phoneme models that yield increasing coverage of a large Tagalog word-list. It is demonstrated that while the increased lexical coverage — or reduced out-of-vocabulary (OOV) rate — leads to only modest (ca 1%-4%) improvements in word error rate, the concomitant improvements in actual term weighted value are as much as 60%. It is also shown that incorporating the augmented lexicons into the LVCSR system before indexing speech is superior to using them *post facto*, e.g., for approximate phonetic matching of OOV keywords in pre-indexed lattices. These results underscore the disproportionate importance of automatic lexicon augmentation for KWS in morphologically rich languages, and advocate for using them early in the LVCSR stage.

Index Terms— Speech Recognition, Keyword Search, Information Retrieval, Morphology, Speech Synthesis

1. LOW-RESOURCE KEYWORD SEARCH

Thanks in part to the falling costs of storage and transmission, large volumes of speech such as oral history archives [1, 2] and on-line lectures [3, 4] are now easily accessible by large user populations via the world wide web. Unlike the text-web, however, searching speech using keywords continues to be a challenging problem. Manually transcribing the speech is often prohibitively expensive. Automatic keyword search (KWS) systems are able to address the problem in some cases, but not in others, because high performance KWS systems, in turn, rely on underlying large vocabulary continuous speech recognition (LVCSR) systems that are also expensive to develop. Good LVCSR systems utilize statistical acoustic- and language-models trained from large quantities of transcribed speech and "conversational" text in the search domain, and manually crafted *pronunciation lexicons* with good coverage of the collection.

We are interested in improving KWS performance in a low resource setting, i.e. where some resources are available to develop an LVCSR system — such as 10 hours of transcribed speech corresponding to about 100K words of transcribed text, and a pronunciation lexicon that covers the words in the training data — but accuracy is sufficiently low that considerable improvement in KWS performance is necessary before the system is usable for searching a speech collection.

A fair amount of past research has been devoted to improving the acoustic models from un-transcribed speech [5, 6, 7, 8, 9], and to adapt language models trained from out-of-domain text to the task at hand. Such methods of improving the LVCSR performance, which subsequently improve KWS performance, are not a focus of this paper. *We investigate the role of the pronunciation lexicon in KWS systems*.

The importance of pronunciation lexicons for LVCSR is not entirely underestimated. Several papers have addressed the problem of automatically generating pronunciations for out of vocabulary (OOV) words [10, 11] in order to improve LVCSR accuracy. But once a reasonably large lexicon is available, speech transcripts in most languages have a fairly small (1%-4%) OOV rate [12, 13]. Even when the OOV rate is reduced by lexicon augmentation, the former OOVs are often absent from the LVCSR transcript, due to poor triphone coverage or low LM probabilities. The impact of lexicon expansion on LVCSR accuracy, therefore, is usually very small.

Two notable exceptions to this conventional wisdom are (i) accuracy on infrequent, content-bearing words, which are more likely to be OOV, and (ii) accuracy in morphologically rich languages, e.g. Czech and Turkish. These exceptions come together in a detrimental fashion when developing KWS systems for a morphologically rich, low resource language such as Tagalog. *This is the setting in which we will quantify the impact of increasing lexical coverage on the performance of a KWS system*.

We assume a transcribed corpus of 10 hours of Tagalog conversational telephone speech [14], along with a 5.7K word pronunciation lexicon that covers all words seen in the transcripts, as our primary acoustic model (AM) training corpus. We assume that the language model (LM) training corpus is either just the transcripts (74K words), or a larger corpus of 595K words.

We first develop state-of-the-art LVCSR and KWS systems based on the given resources. We process and index a 10 hour search collection using the KWS system, and measure KWS performance using a set of 355 Tagalog queries.

We then explore three different methods for augmenting the 5.7K word lexicon to include additional words seen in the larger LM training corpus. The augmented lexicons are used to improve the KWS system in two different ways: reprocessing the speech with the larger lexicon, or using it during keyword search.

The efficacy of the augmented lexicons is measured in terms of

The authors, listed here in alphabetical order, were supported by DARPA BOLT contract No HR0011-12-C-0015, and IARPA BABEL contract No W911NF-12-C-0015. 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 DARPA, IARPA, DoD/ARL or the U.S. Government.

their impact on KWS performance, not just on LVCSR accuracy.

We find that even though lexicon augmentation results in only modest reductions in word error rate (WER), the concomitant improvement in actual term weighted value (ATWV) is often dramatically higher, particularly if the augmented lexicon is used in an early stage to generate the speech lattices used for indexing and search.

The remainder of the paper is organized as follows. We describe our core LVCSR and KWS systems in Section 2, and the three lexicon augmentation methods in Section 3. The impact of augmented lexicons on LVCSR is reported in Section 4 and on KWS in Section 5. The main claims are reiterated in Section 6.

2. BASELINE LVCSR AND KWS SYSTEMS

We conduct our investigations using the IARPA Babel Program Tagalog language collection release `babel106-v0.2f`. We use a 10 hour subset of the 80 hours of conversational telephone speech in this corpus, the transcriptions of this subset, and a pronunciation lexicon restricted to cover only these transcriptions, to simulate low resource conditions. The Babel Tagalog collection also sets aside 10 hours of conversational telephone speech for development-testing. We use a 1.5 hour subset of this development-test set for LVCSR system tuning, e.g. acoustic and language model selection, and refer to it as the "dev" set. The entire 10 hour development-test set, which we refer to as the "eval" set, is used for KWS evaluation[1]. We use a list of 355 keywords (actually, key phrases) created by and shared among the Babel program participants.

2.1. Kaldi-based LVCSR System Description

Our LVCSR system is built using the Kaldi tools [15]. We use standard PLP analysis to extract 13 dimensional acoustic features, and follow a typical maximum likelihood acoustic training recipe, beginning with a flat-start initialization of context-independent phonetic HMMs, and ending with speaker adaptive training (SAT) of state-clustered triphone HMMs with GMM output densities. This is followed by the training of a universal background model from speaker-transformed training data, which is then used to train a subspace Gaussian mixture model (SGMM) for the HMM emission probabilities. Finally, all the training speech is decoded using the SGMM system, and boosted maximum mutual information (BMMI) training of the SGMM parameters is performed.

Two different language models trained with the SRI LM tools [16] are used in the experiments reported below: a trigram LM estimated from the transcripts of the 10 hour acoustic training data (ca 74K words), and a larger trigram LM estimated from the transcripts of the entire 80 hour Babel corpus (ca 595K words). The LMs are estimated separately for each decoding lexicon, so that their vocabulary, the probability of unseen words, etc. match decoding conditions.

The Kaldi decoder generates word lattices [17] for the eval data using the GMM+SAT, SGMM and SGMM+BMMI models. The decoding lexicon is varied systematically, from the low resource lexicon of 5.7K words (8.9K pronunciations), through automatically augmented lexicons of three different sizes, to the full Babel reference lexicon of 23K words (35K pronunciations). Decoding is performed with the small as well as the large LM to create contrastive sets of lattices. A matrix of word error rates is thus measured on the dev set for 3 AMs $\times$ 5 lexicons $\times$ 2 LMs.

[1]The LVCSR dev set is a part of the KWS eval set. We believe that any minor over-fitting that may result from this inclusion has negligible effect on KWS performance on the eval set.

2.2. OpenFST-based KWS System Description

Lattices generated by the BMMI models are processed using the lattice indexing technique described in [18]. The lattices of all the utterances in the eval set are converted from individual finite state transducers (FST) output by Kaldi to a single generalized factor transducer structure in which the start-time, end-time and lattice posterior probability of each word token in every lattice is stored as a 3-dimensional cost associated with that instance of the word. This factor transducer is, in essence, an inverted index of all word sequences seen in the collection of eval set lattices, and permits further manipulation easily using the Google OpenFST tools [19]. Interested readers are referred to [18] for details.

Given a keyword or phrase, one creates a simple finite state machine that accepts the keyword/phrase and composes it with the factor transducer to obtain all occurrences of the keyword/phrase in the eval set lattices, along with the conversation ID, start- and end-time and lattice posterior probability of each occurrence.

All putative instance of a keyword thus obtained are sorted according to their posterior probabilities. Furthermore, a YES/NO decision is assigned to each instance using the method proposed by [20]. Specifically, for each keyword, its expected count in the eval set is estimated by summing the posterior probabilities of all its putative hits, and a decision threshold that maximizes the expected term weighted value is computed for each keyword. All keyword instances with posterior probabilities above this keyword-specific threshold are marked YES.

Finally, the collection of all proposed keyword hits is evaluated against the ground truth using the NIST 2006 Spoken Term Detection evaluation protocol to compute the so called actual term weighted value (ATWV).

2.3. Utilizing Larger Lexicons for KWS

A limitation of the word based indexing scheme described above is that only words present in the LVCSR lexicon appear in the factor transducer. If a word in the query phrase is OOV relative to the lexicon, it will not be found by the FST composition step described above. And yet, the LVCSR vocabulary in low resource settings is often quite small, and the possibility that a query is OOV can be quite large. E.g. the Tagalog baseline vocabulary comprises only 5.7K words, and of the 355 phrasal queries provided for KWS system development, 25% contain at least one OOV relative to this vocabulary.

However, if a large word list is provided, over and above the acoustic training transcripts, a number of techniques are available to generate pronunciations for them, and mitigate the possibility that a keyword is OOV.

A key goal of this paper is to quantify the value of such lexicon augmentation to the KWS application, specifically to the improvement in ATWV from having a larger lexicon. Now, there are (at least) two ways in which one may utilize an augmented lexicon.

1. If the augmented lexicon is available before the speech is processed/indexed, one may incorporate it into the LVCSR stage. The lattices produced, and thus the factor transducer generated for search, will then contain the newly added words wherever there is sufficient evidence for them in the speech.
2. An alternative to decoding all the speech with an augmented lexicon, which is sometimes inconvenient or impossible, is to use it during keyword search.

   Specifically, if K represents a finite state acceptor for a keyword that is OOV relative to a baseline lexicon L_1, but

in-vocabulary relative to an augmented lexicon L_2, where both L_1 and L_2 are finite state transducers that accept phone sequences and output words, and if E is an "edit-distance" transducer that maps any phone sequence to any other phone sequence with a cost equal to their Levenshtein distance, then

$$K' = \text{Project}\left(\text{ShortestPath}\left((L_1^*)^{-1} \circ E \circ (L_2^*) \circ K\right)\right)$$

represents the in-vocabulary keyword/phrase K' that is closest to K. One may use K' as a *proxy* for K to search lattices generated using L_1.

We investigate these two methods of utilizing an augmented lexicon L_2 for handling keywords that are OOV relative to the decoding lexicon L_1 of the low resource KWS system. We demonstrate that the first way is the preferred way to use an augmented lexicon.

3. THREE LEXICON EXPANSION METHODS

We next describe the five lexicons L_1-L_5 used for generating lattices for indexation and search. L_1 (5.7K words) and L_5 (23K words) were manually created, while L_2-L_4 use different grapheme-to-phoneme (G2P) methods to cover progressively larger subsets of the 17K words in L_5 that are OOV relative to L_1.

- L_1: The *5.7K reference lexicon* contains 5.7K words (8.9K pronunciations), and serves as our baseline lexicon.
- L_2: The *Povey lexicon*, developed for automatically augmenting English lexicons in WSJ-like settings, is able to cumulatively provide pronunciations for 6.6K of the 17K OOV words.
- L_3: The *Yarowsky lexicon* was developed with an explicit notion of morphology. It is able to automatically generate pronunciations for 7.3K of the 17K OOV words.
- L_4: The *Sequitur lexicon*, based on [21], was developed as a direct statistical grapheme-to-phoneme model. It is able to cover all 17K OOV words.
- L_5: The *23K reference lexicon* contains 23K words (35K pronunciations), and is our most accurate lexicon.

The three G2P methods and their accuracies are summarized below.

Key Remark: The methods vary in their ability to cover the same set of 17K OOV words, *naturally* yielding different-sized lexicons. But comparing different G2P methods *is not a goal* of this paper, only the value of larger lexicons. Therefore, we do not trim the lexicons to be of equal size. Details of the methods are similarly not germane to the paper, and are omitted due to page restrictions.

3.1. The Povey Lexicon

This lexicon augmentation method, originally designed for English, operates by splitting the OOV into potential prefixes and suffixes, finding the best possible match for the resulting stem-affix pair in the 5.7K reference lexicon, and stitching together a pronunciation from fragments of the matching lexicon entries. E.g., if the reference lexicon contains the entries `beat` ≡ `/b i t/`, `beatable` ≡ `/b i t 6 b l/` and `bear` ≡ `/b E r/`, and the word `bearable` is OOV, then it notes that `bearable` and `bear` differ in the suffix `-able`, just as `beatable` and `beat` do. Since the pronunciations of `beatable` and `beat` differ by the suffix `/6 b l/`, it generates `bearable` ≡ `/b E r 6 b l/`.

This lexicon covers 6.6K of the 17K OOVs (39%). Of the many pronunciations produced for each word, at least one exactly matches an *entire* reference pronunciation for 5.4K of those words (81%).

3.2. The Yarowsky Lexicon

Our second method for lexicon augmentation is based on a novel model of synchronous `word` ≡ `/pronunciation/` transduction which utilizes all existing entries in a pronunciation lexicon to generate new candidate word/pronunciation pairs. For example, in Tagalog, the method learns that the prefix transduction `mag-` ↔ `i-` of a word stem is accompanied — with probability 0.96 — by a synchronous prefix transduction `/m 6 g/-` ↔ `/? i/-` of its pronunciation. This facilitates generation of a pronunciation for an OOV word such as `magtuturo` from the pronunciation of the word `ituturo`, which is present in the 5.7K reference lexicon. Additional evidence for the pronunciation of `magtuturo` also obtains from the synchronous transduction of the word suffixes `-turo` ↔ `-ro` and the corresponding pronunciation suffixes `-/t u r o ?/` ↔ `-/r ?/`, and 76 other observed morphological phenomena, with a consensus probability of 0.98 for the correct pronunciation `magtuturo` ≡ `/m 6 g t u t u r o ?/`.

The algorithm requires as input only a reference lexicon, from which it infers a set of globally-optimized, performance-weighted set of *synchronous prefix and suffix transductions*. Post hoc inspection confirms that these transductions correspond to regular morphological affixations, allophonic substitutions, and variable-length prefixal and suffixal "rhymes."

This lexicon covers 7.3K of the 17K OOVs (44%). Of the many pronunciations produced for each word, at least one exactly matches an *entire* reference pronunciation for 6.4K of those words (88%).

3.3. The Sequitur Lexicon

The third method of lexicon augmentation may be formalized as finding the most probable sequence of phonemes (under a source-channel model) given the sequence of graphemes. This method is implemented in the Sequitur G2P software, and is well described in [21]. We recapitulate it briefly for completeness.

The method uses so-called joint-multigram models, i.e. alignments between consecutive n graphemes ($n \geq 0$) and m phonemes ($m \geq 0$). Contrary to the usual practice, where these alignments are hand-crafted, the Sequitur determines them automatically during the training phase from the input lexicon.

There are two hyper-parameters available to control the size and coverage of the augmented lexicon, namely V, the maximum number of pronunciation variants, and Q, the cumulative probability of all the generated pronunciations for a given OOV word. Multiple pronunciations are generated for a given OOV, in decreasing order of probability, until one of these targets is reached. To choose the best values of these two hyperparameters, we use the goodness criterion

$$\text{Goodness}(V, Q) = N_C - k|N_G - N_R|,$$

where N_G is the number of pronunciations variants generated at the given V and Q, N_C is the number of correct pronunciations among the N_G, and N_R is the number of reference pronunciations. The weight k controls over-generation. We set $k = 0.5$, and find the optimal hyperparameters to be $V = 2$ and $Q = 0.8$.

This lexicon covers all 17K OOVs (100%). Of the many pronunciations produced for each word, at least one exactly matches an *entire* reference pronunciation for 12K of those words (75%).

4. LVCSR IMPROVEMENT BY LEXICON EXPANSION

We perform LVCSR evaluations on the 1.5 hour dev set, and evaluate WERs for three sets of acoustic models, two language models and

Lexicon	L_1	L_2	L_3	L_4	L_5
Words	5.7K	12K	13K	23K	23K
Pronunciations	8.9K	21K	24K	39K	35K
GMM+SAT acoustic models					
3gram LM 74K	74.9	75.6	73.2	73.1	72.9
3gram LM 595K	73.4	74.8	72.0	71.3	71.0
SGMM acoustic models					
3gram LM 74K	71.6	70.4	70.1	69.4	68.8
3gram LM 595K	69.3	68.7	68.2	67.4	66.4
SGMM+BMMI acoustic models					
3gram LM 74K	71.1	70.1	69.8	68.9	68.5
3gram LM 595K	68.9	68.2	67.4	67.0	66.2

Table 1. WER (%) for 5 lexicons $\times$ 3 AMs $\times$ 2 LMs.

five different lexicons. They are reported in Table 1.

Note that small but consistent reductions in WER result from augmenting either the lexicon or the LM alone, and reductions from lexicon and LM augmentation are additive. Note also that the gains persist even as the acoustic models improve, demonstrating further complementarity of the three LVCSR components.

5. KWS IMPROVEMENT BY LEXICON EXPANSION

We perform KWS on the lattices produced by the BMMI acoustic models using the OpenFST-based technique summarized in Subsection 2.2 above.

To quantify the impact of using the augmented lexicons in the LVCSR stage of the KWS system, we index and search the lattices corresponding to all ten SGMM+BMMI systems in the bottom block of Table 1. For instance, the ATWV for the 5.7K reference lexicon and 74K word LM is obtained by using the lattices whose WER is 71.1%, ATWV for augmenting the lexicon via the Yarowsky method and using the 595K word LM is obtained by using the lattices whose WER is 67.4%, etc. The resulting ATWVs for 355 queries are reported in Table 2, where we also provide a breakdown of the ATWV between 87 OOV queries relative to the 5.7K reference lexicon, and 268 queries that are in-vocabulary.

Since OOV queries have no chance to be found in word lattices generated using the 5.7K lexicon, the gains in Table 2 may appear to be trivial to explain. To rule out this trivial explanation, we also investigate utilizing the augmented lexicon to generate in-vocabulary *proxies* for OOV queries, as described in Subsection 2.3. We construct the factor transducer from lattices generated using the 5.7K lexicon, but each time we encounter an OOV word in a query K, we use the method described in Subsection 2.3 to search the factor transducer[2] using proxy in-vocabulary queries K'. The resulting ATWVs, again broken down between in-vocabulary and OOV queries, are reported in Table 3.

6. DISCUSSION AND CONCLUSION

Several interesting conclusions may be drawn from the three tables presented above.

Begin by comparing the last line in Table 1, where relative WER improvement is 4% (68.9% $\rightarrow$ 66.2%), with the last line in Table 2,

Lexicon	L_1	L_2	L_3	L_4	L_5
Words	5.7K	12K	13K	23K	23K
SGMM+BMMI acoustic models $\times$ 3gram LM 74K					
In-Voc queries	0.253	0.271	0.269	0.276	0.273
OOV queries	0.000	0.163	0.262	0.373	0.388
All queries	0.191	0.244	0.267	0.300	0.301
SGMM+BMMI acoustic models $\times$ 3gram LM 595K					
In-Voc queries	0.277	0.287	0.304	0.320	0.320
OOV queries	0.000	0.138	0.294	0.405	0.416
All queries	0.209	0.250	0.302	0.341	0.343

Table 2. ATWV for in-vocabulary (268), OOV (87) and all (355) queries, when the augmented lexicon is used in LVCSR.

Lexicon	L_1	L_2	L_3	L_4	L_5
Words	5.7K	12K	13K	23K	23K
SGMM+BMMI acoustic models $\times$ 3gram LM 74K					
In-Voc queries	0.253	0.253	0.253	0.253	0.253
OOV queries	0.000	0.010	0.063	0.045	0.065
All queries	0.191	0.194	0.206	0.202	0.207
SGMM+BMMI acoustic models $\times$ 3gram LM 595K					
In-Voc queries	0.277	0.277	0.277	0.277	0.277
OOV queries	0.000	0.025	0.035	0.046	0.036
All queries	0.209	0.215	0.217	0.220	0.218

Table 3. ATWV for in-vocabulary (268), OOV (87) and all (355) queries, when the augmented lexicon is used in a pre-indexed KWS system to create proxy queries K' for OOV queries K.

where ATWV improves 64% (0.209 $\rightarrow$ 0.343). This demonstrates that lexicon augmentation has significantly greater impact on KWS performance than LVCSR.

Next, compare the first column for the BMMI acoustic models in Table 1 with the first column of the two "All queries" lines in Table 2. The WER improves by 3% relative (71.1% $\rightarrow$ 68.9%) in Table 1, but the ATWV improves by only 9% (0.191 $\rightarrow$ 0.209) in Table 2. This demonstrates that not all WER reductions are equal: errors reduced by lexicon augmentation matter more for KWS than errors reduced by improving the LM.

Next, compare the two "All queries" lines in Table 2, and note that ATWV improves due lexicon augmentation from 0.191 to 0.301 (58%) for the small LM, compared to 0.209 to 0.343 (64%) for the larger LM. This demonstrates that the KWS improvements from lexicon augmentation are persistent even after LM improvements.

Next, compare the last lines in Tables 2 and 3, and note that ATWV improves in the former by 64%, but only 4% in the latter. This demonstrates that utilizing the augmented lexicon in the KWS stage via approximate phonetic matching (Table 3) is much less effective than utilizing them in the LVCSR stage (Table 2).

Finally, compare the ATWVs for in-vocabulary and OOV queries in Table 2 to note that while much of the improvement from lexicon augmentation is on keywords that were previously OOV, there is significant (10%-15%) collateral improvement in detecting in-vocabulary keywords as well.

We hope that these results convince readers that this paper not only quantifies the significant benefits of lexicon augmentation for KWS, but also provides meaningful insights into the way in which KWS performance is improved.

[2]We have noticed that when a proxy K' returns an unusually large number of YES's for an OOV keyword K, they are predominantly false alarms, and hurt KWS performance, perhaps because K' is a frequent phrase. So we simply discard all hits due to *high-yield* proxies, be they true or false alarms.

7. REFERENCES

[1] The USC Shoah Foundation. Visited 11/30/2012. [Online]. Available: http://sfi.usc.edu/

[2] StoryCorps. Visited 11/30/2012. [Online]. Available: http://www.storycorps.org/

[3] TED: The ideas worth spreading. Visited 30/11/2011. [Online]. Available: http://www.ted.com/

[4] MITVideo. Visited 11/30/2012. [Online]. Available: http://video.mit.edu/

[5] J. Ma and R. Schwartz, "Unsupervised versus supervised training of acoustic models," in *Proceedings of Interspeech 2008*. ISCA, 2008.

[6] L. Wang, M. Gales, and P. Woodland, "Unsupervised training for mandarin broadcast news and conversation transcription," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4, april 2007, pp. IV–353 –IV–356.

[7] F. Wessel and H. Ney, "Unsupervised training of acoustic models for large vocabulary continuous speech recognition," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 1, pp. 23 – 31, jan. 2005.

[8] L. Lamel, J. Gauvain, and G. Adda, "Investigating lightly supervised acoustic model training," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on*, vol. 1, 2001, pp. 477 –480 vol.1.

[9] T. Kemp and A. Waibel, "Unsupervised training of a speech recognizer: Recent experiments," in *Proc. of EUROSPEECH'99*, 1999, pp. 2725–2728.

[10] P. Vozila, J. Adams, Y. Lobacheva, and R. Thomas, "Grapheme to phoneme conversion and dictionary verification using graphonemes," in *8th European Conference on Speech Communication and Technology, EUROSPEECH 2003*. ISCA, 2003.

[11] L. Galescu and J. Allen, "Bi-directional Conversion Between Graphemes and Phonemes Using a Joint N-gram Model," 2001. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.22.5445

[12] R. Rosenfeld, "Optimizing lexical and n-gram coverage via judicious use of linguistic data," in *In Proc. European Conf. on Speech Technology (EUROSPEECH)*, 1995, pp. 1763–1766.

[13] J. Psutka, P. Ircing, J. Psutka, J. Hajič, W. Byrne, and J. Mírovský, "Automatic transcription of czech, russian, and slovak spontaneous speech in the MALACH project," in *Proceedings of Eurospeech 2005*, ISCA. Lisboa, Portugal: ISCA, 2005, pp. 1349–1352.

[14] IARPA-BAA-11-02. Babel – adressing the language deluge. Office of the Director of National Intelligence.

[15] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motliček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, "The Kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011.

[16] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *7th International Conference on Spoken Language Processing, ICSLP2002*, J. H. L. Hansen and B. L. Pellom, Eds. ISCA, 2002.

[17] D. Povey, M. Hannemann, G. Boulianne, L. Burget, A. Ghoshal, M. Janda, M. Karafiát, Štefan Kombrink, P. M. Y. Qian, K. Riedhammer, K. Veselý, and N. T. Vu, "Generating exact lattices in the wfst framework," in *Proceedings of 2012 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE Signal Processing Society, 2012, pp. 4213–4216. [Online]. Available: http://www.fit.vutbr.cz/research/view_pub.php?id=9914

[18] D. Can and M. Saraclar, "Lattice indexing for spoken term detection," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 8, pp. 2338 –2347, nov. 2011.

[19] C. Allauzen, M. Riley, J. Schalkwyk, W. Skut, and M. Mohri, "OpenFst: a general and efficient weighted finite-state transducer library," in *Proceedings of the 12th international conference on Implementation and application of automata*, ser. CIAA'07. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 11–23.

[20] D. R. H. Miller, M. Kleber, C. Kao, O. Kimball, T. Colthurst, S. Lowe, R. M. Schwartz, and H. Gish, "Rapid and accurate spoken term detection," in *Proc. of Interspeech 2007*, vol. 7, 2007, pp. 314–317.

[21] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Communication*, vol. 50, no. 5, pp. 434 – 451, 2008.