MORPH-TO-WORD TRANSDUCTION FOR ACCURATE AND EFFICIENT AUTOMATIC SPEECH RECOGNITION AND KEYWORD SEARCH

A. Ragni, D. Saunders, P. Zahemszky, J. Vasilakes, M. J. F. Gales, K. M. Knill

Department of Engineering, University of Cambridge
Trumpington Street, Cambridge CB2 1PZ, UK
{ar527,ds636,pz251,jav39,mjfg,kate.knill}@eng.cam.ac.uk

ABSTRACT

Word units are a popular choice in statistical language modelling. For inflective and agglutinative languages this choice may result in a high out of vocabulary rate. Subword units, such as morphs, provide an interesting alternative to words. These units can be derived in an unsupervised fashion and empirically show lower out of vocabulary rates. This paper proposes a morph-to-word transduction to convert morph sequences into word sequences. This enables powerful word language models to be applied. In addition, it is expected that techniques such as pruning, confusion network decoding, keyword search and many others may benefit from word rather than morph level decision making. However, word or morph systems alone may not achieve optimal performance in tasks such as keyword search so a combination is typically employed. This paper proposes a single index approach that enables word, morph and phone searches to be performed over a single morph index. Experiments are conducted on IARPA Babel program languages including the surprise languages of the OpenKWS 2015 and 2016 competitions.

Index Terms— morph-to-word transduction, speech recognition, keyword search, single index

1. INTRODUCTION

Accurate and efficient automatic speech recognition (ASR) and keyword search (KWS) have been a subject of extensive research for many years. Initiatives, such as the IARPA Babel program [1], have also looked at generalisation of those approaches beyond English [2], Mandarin [3] and Arabic [4] languages. Unlike English, many languages are highly inflective and agglutinative so a typical 60 hours speech corpus would yield a high out-of-vocabulary rate (OOV). This causes significant issues to downstream tasks such as KWS where missing a word occurrence is penalised orders of magnitude higher than predicting one if it does not in fact occur [5]. A number of approaches have been proposed to deal with the OOV problem such as web data and subword modelling. The use of web data [6, 7] relies heavily on the presence of a given language on the internet. For instance, there are about 1 million words of Dholuo texts available at the time of writing this paper. Subword modelling addresses the OOV problem by decomposing words into sequences of subword units such as morphs, syllables and phones. These units represent a different tradeoff between the number of units and the scope of modelling. Whereas syllable modelling requires language specific knowledge, morph modelling can be conducted in a completely unsupervised fashion [8, 9]. One standard issue with subword modelling is a rather poor word level performance which is believed to originate from the limited scope of language modelling [10] and, more generally, decision making. A number of approaches have been proposed to address this issue. These include the use of high order n-gram morph language models [10] and syllable transductions [11] for converting syllable sequences into word sequences upon which powerful word language models can be applied.

Although successful, subword systems are often used to complement word systems rather than being a stand alone approach in tasks such as KWS [12]. This significantly increases the cost of deployment. A number of approaches have been proposed to address this issue. One example is a parallel index combining indices generated by a word and different subword systems [13]. This approach requires multiple ASR runs and indices. Another example is the use of mixed word and subword units [14, 12, 15]. This approach is more advantageous, requiring single decoding, index and search only. The drawback of this approach is the large vocabulary, especially in the presence of web data, which would make decoding very slow.

This paper proposes a different approach to achieve a diverse single index approach. The main idea consists of picking a morph index and using various transductions to perform different unit searches on it. The key transduction in this work is between morph and word units. The previous work with syllable transduction proposed a two stage approach requiring a second alignment stage to propagate timing information [11]. This paper proposes a single pass approach utilising a lexicographic semiring [16] to propagate both acoustic scores and time information. In order to ensure that different unit searches can be applied on the morph index it should enable changing posterior probabilities depending on the unit to incorporate new information such as word language model probability. This paper shows that this is possible to achieve by keeping information such as acoustic and language model scores separate and pushing a modified form of index to compute posteriors.

The rest of this paper is organised as follows. Section 2 introduces the single pass morph-to-word transduction. A single index approach is discussed in Section 3. Experiments are presented in Section 4. Finally, conclusions are given in Section 5.

2. MORPH-TO-WORD TRANSDUCTION

In inflective and agglutinative languages morphological decomposition of words into constituent subword units called *morphemes* play

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense U. S. Army Research Laboratory (DoD/ARL) contract number W911NF-12-C-0012. 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.

an important role. Consider the following morphological decomposition of two English words

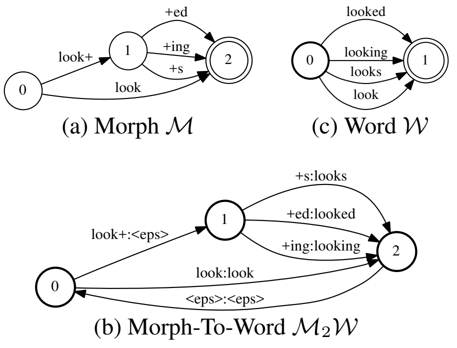
```
looking   look+ +ing
talked    talk+ +ed
```

where plus signs (+) are used to indicate how morphemes *can* and *cannot* be joined. A swap of the first morphemes creates two new valid English words. This simple example illustrates the power of morphological modelling. It is often the case that no morphological decomposition is available for a particular vocabulary or language. Unsupervised morphology induction approaches could be used to address this issue [8, 9]. These normally employ an automatic procedure embedding some general considerations such as the size of morph set and the cost of word decompositions.

It has been observed, however, that reducing entropy among morph units helps to improve both ASR and KWS performance. The simplest approach was illustrated earlier where a simple addition of plus signs ($+$) would cause a significant impact on the nature of decoded morph sequences. Another example would be to use an equivalent of syllable-to-word transduction to collapse morph lattices into word lattices where more constrained word language models could be used. In weighted finite state transducer [17] terminology, such morph-to-word transduction can be expressed as

$$\mathcal{W} = \mathcal{M} \circ \mathcal{M}_2\mathcal{W} \tag{1}$$

where $\mathcal{M}$ and $\mathcal{W}$ are morph and word lattices, $\mathcal{M}_2\mathcal{W}$ is a morph-to-word transducer which maps morph sequences in $\mathcal{M}$ into word sequences encoded by $\mathcal{W}$, epsilon removal and projection on the output label are not shown for simplicity. Figure 1 shows a toy example of these transducers. Not shown on the figure are acoustic log-likelihoods, language model log-probabilities and time information. Note that morph language model probabilities are a placeholder for word language model probabilities to be applied. In order to ensure these quantities are propagated from $\mathcal{M}$ to $\mathcal{W}$ all operations must be performed in a form of a lexicographic semiring [16] where multiplication (path continuation) is given by



(a) Morph $\mathcal{M}$

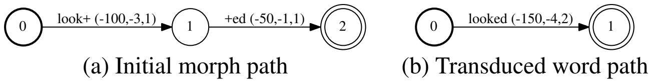
(c) Word $\mathcal{W}$

(b) Morph-To-Word $\mathcal{M}_2\mathcal{W}$

Fig. 1. Example of morph-to-word transduction

$$\langle a_1, l_1, t_1 \rangle \otimes \langle a_2, l_2, t_2 \rangle \quad = \quad \langle a_1 + a_2, l_1 + l_2, t_1 + t_2 \rangle \tag{2}$$

where a_1, a_2 and l_1, l_2 and t_1, t_2 are acoustic log-likelihoods, language model log-probabilities and durations of the first and second transition respectively. The example in Figure 2 shows morph-to-word transduction in the lexicographic semiring for a single path of the morph lattice in Figure 1 (a). Given a word lattice, it is possible to apply word language models to re-rank paths. An alternative approach [11] requires determinising word lattices prior to constrained decoding followed by the application of word language models.



(a) Initial morph path

(b) Transduced word path

Fig. 2. Example use of lexicographic semiring

3. SINGLE INDEX

The approach presented in the previous section enables a single morph decoding to yield two sets of lattices: a morph and word. If used directly two indices would be created. An additional phone index may also be needed to search for keywords not found in either of those indices. This significantly increases the footprint of a keyword search system. An interesting theoretical question is whether one index could be used. By manipulating the index accordingly morph, word and phone searches then could have been performed.

Consider an index [18] in Figure 3 constructed from the morph lattice shown in Figure 1 (a). Note that here weight tuples encode posterior probabilities (multiplied along paths), start time and end time (summed along paths). The output label of the final transitions encodes utterance identity. Such an index cannot be used to perform word search. One issue is morph posterior probabilities that cannot be split into individual acoustic model log-likelihoods and language model log-probabilities. An option would be to encode them separately. However, pushing such an index will not yield valid morph and word posterior probabilities since it has lost the original temporal arrangement of the morphs due to transitions related to the retrieval of individual morphs (1 → 9, 0 → 3, 0 → 4 and 0 → 5).

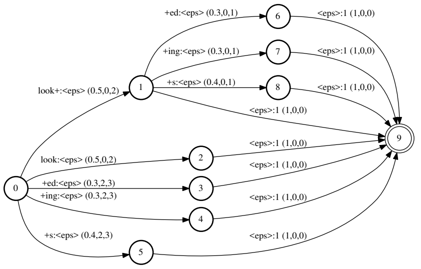


Fig. 3. Example of morph index transducer

Consider now an index in Figure 4 which encodes the same information as the index in Figure 3. In addition, it preserves the temporal arrangement of the morphs. This is accomplished by means

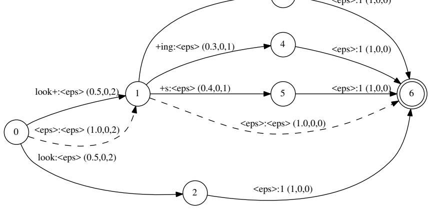


Fig. 4. Example of morph skip index transducer

of skip ϵ-transitions, shown dashed. Such a modification would allow computation of valid morph and word posterior probabilities by pushing the combined weight of acoustic log-likelihoods and language model log-probabilities (weighted appropriately) to the initial state if the skip transitions are omitted from propagation.

Morph search in the skip index $\mathcal{I}$ can be performed by

$$\mathcal{R} = (\mathcal{M}_2\mathcal{W} \circ \mathcal{Q}) \circ \texttt{push}(\mathcal{I}) \tag{3}$$

where $\mathcal{Q}$ is a word query and $\mathcal{R}$ is the result. In order to perform word search the index needs to be converted first. This is accomplished by composing it first with the inverse of $\mathcal{M}_2\mathcal{W}$ and then with a language model transducer $\mathcal{L}$. Note that $\mathcal{M}_2\mathcal{W}$ must encode ϵ-transitions and handle them correctly in composition. Search then can be performed by composing the outcome with the word query

$$\mathcal{R} = \mathcal{Q} \circ \texttt{push}(\mathcal{L} \circ \mathcal{M}_2\mathcal{W}^{-1} \circ \mathcal{I}) \tag{4}$$

There are several options how phone search can be performed. One option is to search the word index for keywords similar to the one requested and known as proxy keywords [19]. Similarity is usually defined in terms of phonetic confusability using a phone-to-phone confusion matrix $\mathcal{P}_2\mathcal{P}$. The search can then be expressed as

$$\mathcal{R} = ((\mathcal{P}_2\mathcal{W}\circ\mathcal{Q})\circ\mathcal{P}_2\mathcal{P}\circ\mathcal{P}_2\mathcal{W})^{-1}\circ\texttt{push}\left(\mathcal{L}\circ\mathcal{M}_2\mathcal{W}^{-1}\circ\mathcal{I}\right) \tag{5}$$

where $\mathcal{P}_2\mathcal{W}$ is a phone-to-word transducer which maps phone sequences into words. Another option is to create a phone index using either the original morph index or created word index [20]. The word index is expected to give more accurate scores hence

$$\mathcal{R} = ((\mathcal{P}_2\mathcal{W} \circ \mathcal{Q}) \circ \mathcal{P}_2\mathcal{P}) \circ \texttt{push}(\mathcal{P}_2\mathcal{W} \circ \mathcal{L} \circ \mathcal{M}_2\mathcal{W}^{-1} \circ \mathcal{I}) \tag{6}$$

For a dense $\mathcal{P}_2\mathcal{P}$ matrix it may be important to restrict the search space of the phone search. For instance, only n highest scoring phone sequences could be selected in equation (6) prior to composing with the created phone index [20].

4. EXPERIMENTS

Experiments in this paper were conducted on 4 IARPA Babel program languages.[1] As shown in Table 1, these languages come with varying amounts of web data. A full language pack (FLP) comprising 60 hours of conversational telephone speech (CTS) data was used for training plus additional 10 hours is available for development. All systems described in this paper are graphemic and built using approaches described in [21]. Tandem and Hybrid acoustic models are used for each language [22]. These were built using features comprising perceptual linear prediction coefficients [23], pitch [24], probability of voicing [24] and bottleneck (BN) features. For Swahili the BN features were extracted from a feed-forward neural network (NN) trained on the FLP data. For the remaining languages these were extracted from multi-task feed-forward NNs trained on 24 Babel languages plus English, Arabic, Mandarin and Spanish CTS data provided by LDC. The multi-language NNs were trained by IBM and RWTH Aachen [25]. 4 acoustic models were built for these languages and 2 acoustic models for Swahili. For efficiency these multiple acoustic models were used in a single pass of joint decoding [26]. Language models (LM) are simple n-grams estimated on acoustic transcripts and web data where appropriate [6]. Unsupervised morphological decomposition was performed using the Morfessor toolkit [8]. This was estimated on the FLP data and then applied to the web data where appropriate. Morph LMs were then built in the same fashion as word based LMs. Table 2 summarises morph LM statistics for each language. Note that the number of characters is shown per vocabulary rather than per corpus which was set to 3. The resulting number of morphs extracted for each language varies a lot. The large number of Amharic morphs originates from the large number of graphs, 247, representing different consonant vowel sequences. These were split into constituent consonant and vowel graphemes for acoustic modelling. As can be seen from Table 2 morphological decompositions yield very low OOV rates for both ASR and KWS. Morph-to-word transduction experiments were performed using an internal version of the OpenFST toolkit [27]. Keyword search experiments were performed using proprietary IBM keyword search software [18]. About 2,000 keywords are available for each language [28]. These are split into in-vocabulary (IV) and out-of-vocabulary (OOV) keywords. The IV keywords are searched in word, morph and word indices for word, morph and morph-to-word systems respectively. The OOV as well as IV keywords not yielding any hits may be searched at the grapheme level by converting queries and lattices into their graphemic form and applying a grapheme-to-grapheme confusion matrix to yield top-n confusable phone sequences to search for, where n was set to 2000. The KWS performance is measured in terms of maximum term weighted value (MTWV) [5] which penalises misses higher than false alarms. The ASR performance is measured in terms of token error rate (TER).

Language	Data ($\times 10^3$)		Vocab ($\times 10^3$)		Char	OOV	
	FLP	Web	FLP	Web	(#)	ASR	KWS
Swahili	294	–	24.4	0	8.2	8.5	19.6
Dholuo	467	1,217	17.5	18.8	6.1	3.0	10.0
Amharic	388	13,911	35.0	223.6	5.1	5.7	9.2
Georgian	406	137,041	34.3	278.6	8.9	3.0	5.2

Table 1. Summary of word-level language statistics

Language	Vocab	Char	OOV	
	($\times 10^3$)	(#)	ASR	KWS
Swahili	7.3	5.8	1.4	0.0
Dholuo	7.5	5.5	0.25	0.0
Amharic	25.6	4.3	0.0	0.8
Georgian	9.6	6.2	2.5	0.0

Table 2. Summary of morph language statistics

The first set of experiments was conducted to examine the impact of morph-to-word transduction on ASR and KWS performance. The first experiment was conducted on Swahili where both ASR and KWS OOV rate is the highest. Table 3 shows that the use of a word LM does not provide extra information to the morph-to-word ASR system once the language modelling context of the morph system is large enough (trigram). This suggests that the morph system has not

#	Unit	TER (%)			MTWV		
		BG	TG	CN	IV	OOV	Total
W	Word	47.6	46.7	44.7	0.5684	0.0000	0.4580
M	Morph	49.0	46.1	45.5	0.5145	0.4759	0.5077
	M2W	47.9	47.2	45.8	0.5448	0.0000	0.4388

Table 3. Swahili word, morph and morph-to-word transduction

[1] Swahili IARPA-babel202b-v1.0d, Dholuo IARPA-babel403b-v1.0b, Amharic IARPA-babel307b-v1.0b, Georgian IARPA-babel404b-v1.0a

generated all word sequences present in the word system. The KWS results show a large gain for IV keywords. Note that the IV/OOV split for the morph system is based on the word vocabulary. Not shown in Table 3 is the performance of morph system where plusses were removed from decompositions. This relaxes constraints on possible morph sequences but results in much poorer KWS performance totalling 0.4797. Table 4 shows Dholuo results. Here, the morph-to-word system shows better ASR results than the morph system. This indicates the usefulness of word level constraints imposed by the word language model for this language. Note that the use of a higher order morph LM does not show gains over the trigram LM. Amharic results in Table 5 show a mixture of the trends observed so far. Here, the morph-to-word system shows better ASR but worse KWS performance. Finally, Table 6 shows OpenKWS 2016 surprise, Georgian, language results which show improvements both in ASR and KWS. These results suggest that morph-to-word transduction can be an effective way to improve morph language modelling for ASR. However, the improvement largely depends on how effective the word language models are. Since TER and MTWV are not strongly correlated it is harder to predict the impact on MTWV. The results, however, confirm that large IV MTWV gains can be obtained. More research is needed to investigate why word level IV performance cannot be obtained with the current approach or why degradation is seen on Amharic. A summary of individual systems is shown in Figure 5. For 3 out of 4 languages combining single decoding based morph-to-word and morph systems, M2W⊗M, using posting list merging yields the best overall single system. For Georgian, where OOV is the smallest, the contribution of morph system is only sufficient to bridge the gap between word and morph-to-word system but not to gain significantly more overt that.

#	Unit	TER (%)			MTWV		
		BG	TG	CN	IV	OOV	Total
W	Word	39.6	39.3	38.3	0.6493	0.0000	0.5762
M	Morph	41.5	39.8	39.4	0.6241	0.5180	0.6132
	M2W	39.6	39.4	38.4	0.6375	0.0000	0.5656

Table 4. Dholuo word, morph and morph-to-word results

#	Unit	TER (%)			MTWV		
		BG	TG	CN	IV	OOV	Total
W	Word	41.5	41.2	40.8	0.6596	0.0000	0.6020
M	Morph	44.0	43.5	43.0	0.6242	0.4379	0.6087
	M2W	42.7	42.4	41.7	0.6090	0.0000	0.5563

Table 5. Amharic word, morph and word-to-morph results

#	Unit	TER (%)			MTWV		
		BG	TG	CN	IV	OOV	Total
W	Word	39.9	39.3	37.5	0.7363	0.0000	0.6988
M	Morph	44.5	41.1	40.9	0.6785	0.6463	0.6775
	M2W	40.5	40.4	39.3	0.6947	0.0000	0.6591

Table 6. Georgian word, morph and morph-to-word results

The second set of experiments was performed to investigate the single index approach. For this investigation a simplified approach was used. The Swahili morph system in Table 3 was used to produce a single set of lattices. These were then searched at the morph, word and phone level. The first block in Table 7 illustrates the accuracy of these searches. Due to the high OOV rate word units show the worst performance. However, combining (⊗) these units with morph and phone yields performance superior to any of the individual systems in Table 3. For comparison, the last block shows results obtained using cascaded search (⊕) where IV with no found examples as well as OOV keywords are searched at the phone level. The single index using 3 searches is more accurate than individual cascaded searches but loses to the combination of two cascaded searches requiring 2 decoding runs and 4 searches over 4 different indices.

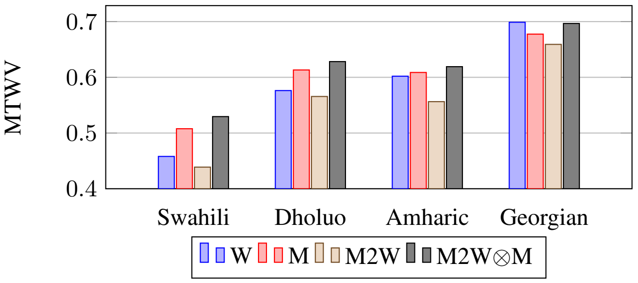


Fig. 5. Summary of overall MTWV performance for single decoding approaches

#	Search	MTWV		
		IV	OOV	Total
W	Word	0.5448	—	0.4388
M	Morph	0.5149	0.4759	0.5077
P	Phone	0.4633	0.4241	0.4568
	W⊗M	0.5420	0.4749	0.5295
	W⊗M⊗P	0.5522	0.5567	0.5554
	W⊕P	0.5706	0.4256	0.5433
	M⊕P	0.5316	0.5550	0.5362
	(W⊕P)⊗(M⊕P)	0.5707	0.5567	0.5687

Table 7. Multiple searches over single set of Swahili lattices

5. CONCLUSIONS

Finding an appropriate unit of language modelling is an open question for many applications and for many languages. Generally there are many options to choose from: words, morphs, syllables, phones, etc. This paper has looked at the morph units which provide a sufficiently wide scope of modelling yet empirically yield very low out-of-vocabulary rates. In order to enlarge the scope of modelling even further this paper has proposed a morph-to-word transduction that enables to convert morph sequence into word sequences. This has multiple benefits including possibility to apply powerful word language models as well as making decisions during pruning, confusion network decoding, keyword search at the word level. This paper has also looked at how word, morph and phone keyword searches can be efficiently performed on a single morph index. A modification to the standard index has been proposed to enable multiple searches. Experiments examining morph-to-word transduction were performed on 4 Babel program languages. These showed that it is possible to obtain speech recognition gains over morph systems as well as keyword search gains over individual word and morph systems.

6. REFERENCES

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

[2] G. Saon, T. Sercu, S. Rennie, and J. H.-K Kuo, "The IBM 2016 English conversational telephone speech recognition system," in *Interspeech*, 2016.

[3] K. Yu, M. J. F. Gales, and P. C. Woodland, "Unsupervised training with directed manual transcription for recognizing Mandarin broadcast audio," in *Interspeech*, 2007.

[4] M. Afify, L. Nguyen, B. Xiang, S. Abdou, and J. Makhoul, "Recent progress in arabic broadcast news transcription at BBN," in *Interspeech*, 2005.

[5] J. G. Fiscus, J. G. Ajot, J. Garofalo, and G. Doddington, "Results of the 2006 spoken term detection evaluation," in *SIGIR SSCS*, 2007.

[6] G. Mendels, E. Cooper, V. Soto, J. Hirschberg, M. Gales, K. Knill, A. Ragni, and H. Wang, "Improving speech recognition and keyword search for low resource languages using web data," in *Interspeech*, 2015.

[7] L. Zhang, D. Karakos, W. Hartmann, R. Hsiao, R. Schwartz, and S. Tsakalidis, "Enhancing low resource keyword spotting with automatically retrieved web documents," in *Interspeech*, 2015.

[8] S. Virpioja, P. Smit, S.-A. Grönroos, and M. Kurimo, "Morfessor 2.0: Python implementation and extensions for morfessor baseline," Tech. Rep. SCIENCE + TECHNOLOGY, 25/2013, Aalto University, 2013, ISBN 978-952-60-5501-5.

[9] S.-A. Grönroos, S. Virpioja, P. Smit, and M. Kurimo, "Morfessor FlatCat: An HMM-based method for unsupervised and semi-supervised learning of morphology," in *COLING*, 2014.

[10] T. Hirsimäki, J. Pylkkonen, and M. Kurimo, "Importance of high-order N-gram models in morph-based speech recognition," *IEEE TASLP*, vol. 17, no. 4, pp. 724–732, 2009.

[11] H. Su, V. T. Pham, Y. He, and J. Hieronymus, "Improvements on transducing syllable lattice to word lattice for keyword search," in *ICASSP*, 2015.

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

[13] L. Mangu, "Keyword search using confusion networks and parallel search," in *IARPA Babel PI meeting*, 2014.

[14] I. Bulyko, J. Herrero, C. Mihelich, and O. Kimball, "Subword speech recognition for detection of unseen words," in *Interspeech*, 2012.

[15] Y. He, B. Hutchinson, P. Baumann, M. Ostendorf, E. Fosler-Lussier, and J. Pierrehumbert, "Subword-based modeling for handling OOV words in keyword spotting," in *ICASSP*, 2014.

[16] B. Roark, R. Sproat, and I. Shafran, "Lexicographic semirings for exact automata encoding of sequence models," in *ACL*, 2011.

[17] M. Mohri, F. Pereira, and M. Riley, "Weighted finite-state transducers in speech recognition," *Computer Speech and Language*, vol. 16, pp. 69–88, 2002.

[18] D. Can and M. Saraclar, "Lattice indexing for spoken term detection," *IEEE TASLP*, vol. 19, no. 8, pp. 2338–2347, 2001.

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

[20] L. Mangu, B. Kingsbury, H. Soltau, Kuo H.-K., and M. Picheny, "Efficient spoken term detection using confusion networks," in *ICASSP*, 2014.

[21] M. J. F. Gales, K. M. Knill, and A. Ragni, "Unicode-based graphemic systems for limited resource languages," in *ICASSP*, 2015.

[22] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, A. Ragni, V. Valtchev, P. C. Woodland, and C. Zhang, *The HTK Book (for HTK Version 3.5)*, University of Cambridge, `http://htk.eng.cam.ac.uk`, 2015.

[23] H. Hermansky, "Perceptual Linear Predictive (PLP) analysis of speech," *Journal of the Acoustic Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[24] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *ICASSP*, 2014.

[25] J. Cui, B. Kingsbury, B. Ramabhadran, A. Sethy, K. Audhkhasi, X. Cui, E. Kislal, L. Mangu, M. Nussbaum-Thom, M. Picheny, Z. Tüske, P. Golik, R. Schlüter, H. Ney, M. Gales, K. Knill, A. Ragni, H. Wang, and P. Woodland, "Multilingual representations for low resource speech recognition and keyword search," in *ASRU*, 2015.

[26] H. Wang, A. Ragni, M. J. F. Gales, K. M. Knill, P. C. Woodland, and C. Zhang, "Joint decoding of tandem and hybrid systems for improved keyword spotting on low resource languages," in *Interspeech*, 2015.

[27] C. Allauzen, M. Riley, J. Schalkwyk, W. Skut, and M. Mohri, "OpenFst: A general and efficient weighted finite-state transducer library," in *CIAA*, 2007.

[28] J. Cui, J. Mamou, B. Kingsbury, and B. Ramabhadran, "Automatic keyword selection for keyword search development and tuning," in *ICASSP*, 2014.