Improving automatically induced lexicons for highly agglutinating languages using data-driven morphological segmentation

Wiehan Agenbag, Thomas Niesler

Department of Electrical and Electronic Engineering
Stellenbosch University, South Africa
wagenbag@sun.ac.za, trn@sun.ac.za

Abstract

We present a method of improving the performance of automatically induced lexicons for highly agglutinating languages. Our previous work demonstrated the feasibility of using automatic sub-word unit discovery and lexicon induction to enable ASR for under-resourced languages. However, a particularly challenging case for such approaches is found in agglutinating languages, which have large vocabularies of infrequently used words. In this study, we address the unfavorable vocabulary distribution of such languages by performing data-driven morphological segmentation of the orthography prior to lexicon induction. We apply this novel step to a corpus of recorded radio broadcasts in Luganda, which is a highly agglutinating and severely under-resourced language. The intervention leads to a 10% (relative) reduction in WER, which puts the resulting ASR performance on par with an expert lexicon. When context is added to the morphological segments prior to lexicon induction, a further 1% WER reduction is achieved. This demonstrates that it is feasible to perform ASR in an under-resourced setting using an automatically induced lexicon even in the case of a highly agglutinating language.

Index Terms: unsupervised SWU discovery, automatic lexicon induction, ASR, under-resourced languages, morphological segmentation

1. Introduction

The manual development of pronunciation lexicons, particularly for the case of under-resourced languages, is generally a time consuming and expensive process which requires the availability of linguists familiar with the task language. If the steps of sub-word unit discovery (SWU) and pronunciation lexicon generation could be automated with satisfactory performance, it would greatly streamline the process of implementing automatic speech recognition (ASR) in an under-resourced setting. In some cases, such as where trained linguists are not available at all, it would enable the implemention of ASR where it would otherwise be infeasible.

In previous work we have demonstrated the feasibility of performing ASR with an automatically induced pronunciation lexicon in a truly under-resourced setting, using only recorded speech and orthographic transcriptions [1]. However, we also discovered that highly agglutinating languages (such as Luganda) pose a particular challenge for automatic lexicon induction. Specifically, the agglutinating nature of these languages leads to a massive expansion in the number of vocabulary types, many of which occur rarely. This makes it challenging to obtain a sufficient number of observations per word for robust pronunciation induction for a substantial portion of the vocabulary. Because of this, we observed a performance margin relative to expert lexicons which was significantly worse when compared to less agglutinating languages.

In this work, we attempt to address this shortcoming by inducing pronunciations for orthographic types that are shorter than words. Ideally, these would be constituent morphemes designed by expert linguists. However, since we operate in the under-resourced setting, our objective is to have fully automatic procedures that require minimal human input. We therefore investigate the use of a data-driven technique to subdivide words into morpheme-like chunks (hereafter referred to as morphological segments) for which more reliable pronunciations can potentially be induced. As far as we are aware, this is a novel intervention in the automatic lexicon induction task.

2. Automatic sub-word unit discovery and lexicon induction

2.1. Background

The task of automatically generating a pronunciation lexicon from a word-annotated speech corpus requires addressing a number of subtasks. First, a set of sub-word units needs to be established, if one (such as an expert defined phoneme set or suitable grapheme-based orthography) is not already available. Subsequently, pronunciations must be generated for each word. Due to the considerable acoustic variability of speech, this step produces a number of possible candidates, and necessitates some form of scoring and pruning in order to achieve an acceptable number of variants per word.

Many approaches to unsupervised SWU discovery rely on a two-step process where speech is first segmented [2–7] and the resulting segments are then clustered to form a compact set of units [3, 7–15]. Alternatively, joint approaches to segmentation and clustering can be taken, such as the estimation of a hierarchical Bayesian model (HBM) [16–21]. An alternative set of approaches to jointly discover SWU segmentations and clusters involves an iterative process where a speech corpus is tokenized using a fixed set of SWU templates, and thereafter the templates are updated while the tokenization is held fixed [7, 13, 14, 22].

Once a SWU inventory is defined, a pronunciation lexicon can be generated. If a seed lexicon and an alphabetic orthography is available, pronunciations for new words can be generated using grapheme-to-phoneme methods. Alternatively, candidates can be generated by means of Viterbi decoding, using acoustic models obtained from a seed lexicon [23–28] or from automatically discovered units [22]. There does not appear to be any evidence in the literature of previous work that attempts the generation of lexicons based on morphological segments instead of words.

Once a set of candidate pronunciations has been generated, their number must be reduced to a compact set of "canon-

ical" variants. This can be achieved through pronunciation scoring and subsequent pruning. The most straightforward of these scoring schemes is simply to use the relative frequency of each candidate [8, 26, 27]. Alternative scoring approaches rely on graph structure representations of pronunciation variation, such as automata or lattices generated during acoustic decoding [22, 23, 28]. This enables scoring based on, for example, relative similarity to other hypotheses, even when no variant is observed more than once.

2.2. Lexicon induction in the under-resourced setting

The two-stage approach to SWU discovery and lexicon induction that we proposed in [1] requires as input only recorded speech and orthographic transcriptions. The first stage is an initial joint SWU and lexicon discovery step, which is performed on a limited high-occurrence subset of the training set vocabulary. Subsequently, a full lexicon extraction procedure is performed which refines the SWU inventory and produces a lexicon covering the entire training set vocabulary.

The initial SWU discovery step proceeds by assigning a number of HMM-GMM states to each word in the initial training set, and then allowing these states to self-organise. The resulting states are subsequently agglomeratively tied across words to form a compact set of SWUs and an initial lexicon.

The full lexicon extraction procedure uses a divide-and-conquer strategy of iteratively and successively updating either the lexicon or the SWU acoustic models while holding the other fixed. The lexicon update consists of a candidate generation step that simply performs an unconstained acoustic decode on speech features that have been segmented to the word level, followed by pruning. The pruning itself proceeds in two stages, both of which rely on the estimation of a pronunciation model for each word, based on an N-state left-to-right HMM with forward skips and emitting SWU symbols on each state. The first pruning stage reduces variability at the state-level, while the second discards entire pronunciations.

3. Morphological segmentation

The process of morphological segmentations usually consists of two stages: *training* a segmentation model $\boldsymbol{\theta}$, and subsequent *decoding* of a test set W using a tokenization function $\phi\left(W;\boldsymbol{\theta}\right)$ [29–31]. In our case, however, we tokenize the training corpus itself, so there is no secondary test set.

The training of the segmentation model $\boldsymbol{\theta}$ involves minimizing a cost function $L(\boldsymbol{D}_W, \boldsymbol{\theta})$, where $\boldsymbol{D}_W$ is the corpus of training set words. The cost function consists of a model likelihood $p(\boldsymbol{D}_W|\boldsymbol{\theta})$ and a model prior $p(\boldsymbol{\theta})$. The data likelihood assumes that the set of L_j morphological segments m_{ji} constituting a word w_j occur independently such that:

$$\log p(\boldsymbol{D}_W|\boldsymbol{\theta}) = \sum_{j=1}^{N}\sum_{i=1}^{L_j} \log p(m_{ji}|\boldsymbol{\theta}). \quad (1)$$

The maximum-likelihood estimate of the probability of a morphological segment is based on its usage count τ_i:

$$p(m_i|\boldsymbol{\theta}) = \frac{\tau_i}{N + \sum_i \tau_i}. \quad (2)$$

The prior probability of the model $p(\boldsymbol{\theta})$ assigns higher probabilities to segmentation models that consist of fewer and shorter segments. this is achieved by an implicit exponential prior on morpheme lengths.

In order to control the model's tendency to oversegment or undersegment, a weight parameter α is introduced into the cost function:

$$L(\boldsymbol{D}_W, \boldsymbol{\theta}) = -\log p(\boldsymbol{\theta}) - \alpha \log p(\boldsymbol{D}_W|\boldsymbol{\theta}). \quad (3)$$

By reducing α, it is possible to emphasise the cost function contribution due to the prior, which would favour a segmentation model consisting of shorter segments, and vice versa. We can therefore control the average segmentation length by tuning α.

The training algorithm implemented by Morfessor 2.0 proceeds with a greedy, local search [32]. It does this by considering one morphological segment at a time and evaluating each possible split of the segment into smaller segments to determine the one that minimizes the cost function with respect to the current model parameters. The model parameters themselves are then updated using the previously determined split as part of the model. This procedure of splitting and model updating is performed recursively on the segments resulting at each step. For the sake of efficiency, morphological segments are tied across all word types that contain many words.

3.1. Lexicon induction using morphological segments

We now describe how we extended our approach to lexicon induction (see Section 2.2) to use morphological segments. An overview of the steps involved is given in Figure 1. First, a morphological segmentation is performed using the approach described in Section 3. The training corpus for this segmentation is the set of unique vocabulary types in the training set orthographic transcriptions. A target average segment length is specified at this stage, and the parameter α of the segmentation cost function (Equation 3) is tuned to achieve this.

Subsequently, the resulting segmented orthographic transcriptions are aligned with the training set acoustic features, in order to extract pronunciations from the utterance-level SWU sequences that result from acoustic decoding. There are at least two ways to achieve this. The first is to regard the morphological segments as atomic, and train whole-segment acoustic models which can then be used for Viterbi alignment. The second option is to estimate acoustic models for grapheme SWUs and use those for acoustic alignment. The latter option is expected to yield more accurate alignments, since grapheme-level acoustic modeling is both more robust and more granular than word level modeling. A further advantage of using graphemes for alignment is that it decouples the performance of the resulting lexicons from the accuracy of the temporal segment alignments, which is otherwise expected to be correlated with the average segment lengths.

Using the procedure described in Section 2.2, but using morphological segment-level transcriptions instead of word-level transcriptions, a full pronunciation lexicon for the morphological segments can be extracted. Finally, in order to form a word-level pronunciation lexicon, we concatenate the newly induced pronunciations of the constituent morphological segments of each training set word.

4. Experiments and results

4.1. Dataset and experimental setup

The dataset used for experimentation is summarised in Table 1. This data has been compiled from recordings of Ugandan community radio stations broadcasting in Luganda, and have been orthographically annotated by mother-tongue speakers [33, 34].

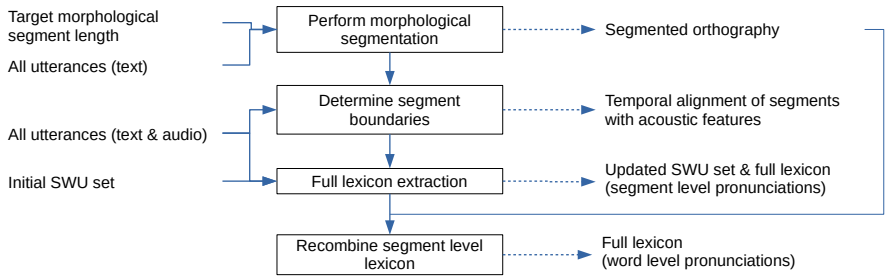

Figure 1: *Overview of the steps performed during lexicon induction using morphological segments.*

Table 1: *Summary of dataset used for experimental evaluation.*

Dataset	Hours	#Utts	#Word tokens
Train	9.59h	8774	18305
Test	0.29h	164	1150

Luganda is a highly agglutinating language, as can be seen from the disproportionately large vocabulary relative to its recording time. It is also a severely under-resourced language, with practically no resources available other than the dataset used in this study. It is worth investigating the effect of morphological segmentation on the other languages as well, but time constraints did not allow this. The dataset includes a pronunciation lexicon compiled by phonetic experts, which enabled the establishment of a baseline indicating what can be achieved using the conventional ASR system development approach.

In order to fairly compare the performance of performing lexicon induction with a morphologically segmented orthography to induction using the original orthography, we use the same initial SWU inventory and associated metaparameters (number of SWUs $N_p = 150$ and average SWU duration $R_p = 78$ms) in both cases. Further, in both cases we relied on accurate temporal alignment of the orthographic segments with the speech featurs.

We used the Morfessor 2.0 implementation of the approach to performing morphological segmentation described in Section 3 [32]. For the acoustic model estimation, decoding and alignment necessary during lexicon extraction, we used the HTK tools [35]. For the evaluation of ASR performance, we used the Kaldi toolkit, culminating in a system consisting of a combination of SGMM-MMI and DNN/HMM acoustic models (SGMM+DNN) with cross-word dependency [36] . Although newer and better models such as SGMM+TDNN are available, we found that they reflect similar performance trends with a relatively constant performance gain, while consuming significantly more processing time to train [1]. In addition to the acoustic models, we trained bigram language models from corpora consisting of only the corresponding training prompts.

4.2. Context independent pronunciation induction

In the first set of experiments involving morphological segmentation, we investigated the effect of various target average segment lengths on the resulting vocabulary distributions and resulting word error rates. The result of these experiments are shown in Table 2. The lexicons indicated by *Phoneme* and *Grapheme* represent the baseline results obtained using respectively an expert-compiled phonetic lexicon and a lexicon using graphemes as sub-word units. The lexicon indicated by *Auto* refers to an automatically induced lexicon with no morphological segmentation performed, i.e. using the unmodified orthography. The lexicons indicated by *Auto+segmentation* refer to lexicons that were automatically induced using the morphologically segmented orthography.

We observe from Table 2 that even modest levels morphological segmentation profoundly reduce the number of unique types in the vocabulary, with an approximate halving of the average segment length leading to a nine-fold reduction in segment types. The frequency of occurrence of segment types also becomes much more favourable.

These promising trends are also reflected in the word error rates. Without morphological segmentation, the automatically induced lexicon is substantially outperformed by the expert and grapheme lexicons. However, even the least aggressive segmentation yields error rates that are close to (or slightly better than) the expert baseline, while still retaining a significant proportion of segment types. A further reduction in error rates is observed as the segments become shorter.

4.3. Context-dependent pronuniation induction

It was observed in the previous section that shorter morphological segments yielded higher-performing lexicons. However, shorter segments also it lead to a significant reduction in the number of segment types. At some point the remaining segment types will become less able to capture all the acoustically distinct types that would be necessary for linguistic discrimination during speech recognition. In order to address this, we will attempt to increase the number of types associated with shorter morphological segments by taking the preceding and the following segments into account as contextual information. This should help capture segment-level pronunciation variation relevant for distinguishing words.

When an inventory of context-dependant morphological segments is established in such a way that each unique context becomes a new type, it is expected that there will be a large increase in types, many of which occur infrequently. This might counteract any performance gain achieved by introducing context-dependency. We address this by establishing a minimum occurrence threshold for each context-dependent instance of a segment. In cases where this threshold is not met, we pool infrequent contexts of a segment either by shared left or right context. If the specified threshold is still not met, then context is discarded altogether.

We investigated the application of this approach for various segment lengths and pooling thresholds, with the resulting word error rates shown in Table 3. In general, we can observe a performance improvement for the lexicons induced using context-

Table 2: *Vocabulary distributions for various levels of morphological segmentation (average segment length in graphemes), and the associated ASR performance for each system.*

Lexicon	**Average segment length**	**# Types**	$n > 3$		$n > 9$		**% WER**
			Types	Tokens	Types	Tokens	
Phoneme							54.94%
Grapheme							55.14%
Auto (unsegmented)	5.7	18305	14.4%	75.0%	5.6%	63.8%	60.91%
Auto + segmentation	3.0	2135	70.12%	99.31%	49.51%	97.72%	56.35%
	2.5	836	93.42%	99.94%	82.30%	99.63%	56.35%
	2.0	230	99.57%	100.00%	99.13%	100.00%	54.95%
	1.5	70	98.57%	100.00%	97.14%	100.00%	**54.90%**
	1.0	33	96.97%	100.00%	93.94%	100.00%	55.14%

Table 3: *The ASR performance of lexicons induced using context-dependant morphological segmentation for various pooling thresholds. A threshold of ∞ indicates context independent segments.*

Average segment length	**Threshold**	**# Types**	**% WER**
2	∞	230	54.95%
	250	434	**54.33%**
	125	720	54.95%
	62	1358	55.05%
1.5	∞	70	54.90%
	250	462	55.53%
	125	856	**54.81%**
	62	1573	56.83%
1	∞	33	55.15%
	250	655	54.52%
	125	992	**54.28%**
	62	1347	55.87%

dependant segments, relative to context-independant segments, at least for some pooling thresholds. It is likely that there is a trade-off involved with reducing the pooling threshold, which increases the number of types for which independant pronunciations are be obtained, but also reduces the number of observations per type, and thus the robustness of the estimated pronunciations.

Another important observation is that we have been able to improve on both the expert (phoneme) baseline as well as the grapheme baseline for Luganda (54.95% and 55.14% respectively), with the best-performing automatically induced system achieving an error rate of 54.28%. The system producing this error rate corresponds to a segment length of 1, which amounts to mapping the automatically discovered pool of acoustic SWUs to context-dependant graphemes. The improvement of such a system over a pure grapheme-based lexicon is interesting, especially because the ASR training pipeline used for performance evaluation already includes a context-dependant acoustic model expansion step (in the form of triphone creation). Since the ASR performance of the baseline grapheme lexicon already includes context-dependence, it can be deduced that the additional performance gain from using grapheme context during lexicon induction can be attributed to our SWU inventory.

5. Summary and conclusions

In this paper, we have presented an approach for improving the performance of automatic lexicon induction in the case of a highly agglutinating language (Luganda). This involved performing data-driven morphological segmentation on the training set orthography before lexicon induction. As summarised in Table 4, the resulting ASR system exhibited an approximately 10% relative reduction in word error rate compared to the best lexicon induced on an unsegmented orthography. We also demonstrated that further performance gains can be achieved by making the segments context dependent, resulting in a system that performs 1% better than one trained using an expert lexicon. This demonstrates that it is feasible to use automatically induced lexicons to facilitate ASR in an under-resourced setting even for highly agglutinating languages.

Table 4: *Summary of the best ASR performance for each of the various systems evaluated in this study.*

System	**% WER**
Phoneme	54.94%
Grapheme	55.14%
Auto	60.91%
Auto + segmentation	54.90%
Auto + segmentation + context	54.28%

6. Acknowledgements

The presented study was supported by Telkom South Africa. All experiments were performed using the University of Stellenboschs Rhasatsha HPC or the facilities at the Centre for High Performance Computing (CHPC).

7. References

[1] W. Agenbag and T. Niesler, "Automatic sub-word unit discovery and pronunciation lexicon induction for ASR with application to under-resourced languages," *Computer Speech and Language*, vol. 57, pp. 20 – 40, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0885230818301980

[2] T. Svendsen and F. Soong, "On the automatic segmentation of speech signals," in *Proceedings of ICASSP '87. IEEE Interna-*

tional Conference on Acoustics, Speech, and Signal Processing, vol. 12, Apr 1987, pp. 77–80.

[3] L. ten Bosch and B. Cranen, "A computational model for unsupervised word discovery." in *Proceedings of Interspeech*, 2007, pp. 1481–1484.

[4] M. Sharma and R. Mammone, "Blind speech segmentation: automatic segmentation of speech without linguistic knowledge," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 2, Oct 1996, pp. 1237–1240 vol.2.

[5] M. Bacchiani and M. Ostendorf, "Joint lexicon, acoustic unit inventory and model design," *Speech Communication*, vol. 29, no. 2?4, pp. 99 – 114, 1999.

[6] V. Z. van Vuuren, L. ten Bosch, and T. Niesler, "Automatic segmentation of TIMIT by dynamic programming," in *Proceedings of the Annual Symposium of the Pattern Recognition Society of South Africa (PRASA)*, 2012.

[7] M. hung Siu, H. Gish, A. Chan, W. Belfield, and S. Lowe, "Unsupervised training of an hmm-based self-organizing unit recognizer with applications to topic classification and keyword discovery," *Computer Speech & Language*, vol. 28, no. 1, pp. 210 – 223, 2014.

[8] G. Goussard and T. R. Niesler, "Automatic discovery of subword units and pronunciations for automatic speech recognition using TIMIT," in *Proceedings of the Annual Symposium of the Pattern Recognition Society of South Africa (PRASA)*, 2010.

[9] A. Jansen and K. Church, "Towards unsupervised training of speaker independent acoustic models." in *Proceedings of Interspeech*, 2011, pp. 1693–1692.

[10] M. Razavi *et al.*, "An HMM-Based Formalism for Automatic Subword Unit Derivation and Pronunciation Generation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[11] H. Wang, T. Lee, C.-C. Leung, B. Ma, and H. Li, "Unsupervised mining of acoustic subword units with segment-level gaussian posteriorgrams," in *Proceedings of Interspeech*, 2013, pp. 2297–2301.

[12] B. Varadarajan and S. Khudanpur, "Automatically learning speaker-independent acoustic subword units." in *Proceedings of Interspeech*, 2008, pp. 1333–1336.

[13] W. Agenbag and T. R. Niesler, "Automatic segmentation and clustering of speech using sparse coding and metaheuristic search," in *Proceedings of Interspeech*, 2015.

[14] ——, "Refining sparse coding sub-word unit inventories with lattice-constrained viterbi training," in *Proceedings of the Workshop on Spoken Language Technologies for Under-resourced languages*, 2016.

[15] M. Razavi, R. Rasipuram, and M. Magimai.-Doss, "Towards weakly supervised acoustic subword unit discovery and lexicon development using hidden Markov models," *Speech Communication*, vol. 96, pp. 168 – 183, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S016763931730105X

[16] C.-y. Lee and J. Glass, "A nonparametric Bayesian approach to acoustic model discovery," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, 2012, pp. 40–49.

[17] L. Ondel, L. Burget, and J. Černocký, "Variational inference for acoustic unit discovery," *Procedia Computer Science*, vol. 81, no. Supplement C, pp. 80 – 86, 2016, sLTU-2016 5th Workshop on Spoken Language Technologies for Under-resourced languages 09-12 May 2016 Yogyakarta, Indonesia.

[18] C. Liu, J. Yang, M. Sun, S. Kesiraju, A. Rott, L. Ondel, P. Ghahremani, N. Dehak, L. Burget, and S. Khudanpur, "An empirical evaluation of zero resource acoustic unit discovery," in *Proceedings of Interspeech*, 2017.

[19] A. H. H. N. Torbati, J. Picone, and M. Sobel, "Speech acoustic unit segmentation using hierarchical Dirichlet processes." in *Proceedings of Interspeech*, 2013, pp. 637–641.

[20] C.-y. Lee, Y. Zhang, and J. R. Glass, "Joint learning of phonetic units and word pronunciations for ASR." in *Proceedings of Empirical Methods on Natural Language Processing (EMNLP)*, 2013, pp. 182–192.

[21] C.-y. Lee, T. J. O'Donnell, and J. Glass, "Unsupervised lexicon discovery from acoustic input," *Transactions of the Association for Computational Linguistics*, vol. 3, pp. 389–403, 2015.

[22] R. Singh, B. Raj, and R. Stern, "Automatic generation of subword units for speech recognition systems," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 2, pp. 89–99, Feb 2002.

[23] X. Zhang, V. Manohar, D. Povey, and S. Khudanpur, "Acoustic data-driven lexicon learning based on a greedy pronunciation selection framework," in *Proceedings of Interspeech*, 2017.

[24] D. Harwath and J. Glass, "Speech recognition without a lexicon - bridging the gap between graphemic and phonetic systems," in *Proceedings of Interspeech*, 2014.

[25] M. Ravishankar and M. Eskenazi, "Automatic generation of context-dependent pronunciations," in *Fifth European Conference on Speech Communication and Technology*, 1997.

[26] G. Chen, D. Povey, and S. Khudanpur, "Acoustic data-driven pronunciation lexicon generation for logographic languages," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 5350–5354.

[27] N. Goel, S. Thomas, M. Agarwal, P. Akyazi, L. Burget, K. Feng, A. Ghoshal, O. Glembek, M. Karafit, D. Povey, A. Rastrow, R. C. Rose, and P. Schwarz, "Approaches to automatic lexicon learning with limited training examples," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, March 2010, pp. 5094–5097.

[28] L. Lu, A. Ghoshal, and S. Renals, "Acoustic data-driven pronunciation lexicon for large vocabulary speech recognition," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 374–379.

[29] M. Creutz and K. Lagus, *Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0*. Helsinki University of Technology Helsinki, 2005.

[30] O. Kohonen, S. Virpioja, and K. Lagus, "Semi-supervised learning of concatenative morphology," in *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*. Association for Computational Linguistics, 2010, pp. 78–86.

[31] S. Virpioja, O. Kohonen, and K. Lagus, "Evaluating the effect of word frequencies in a probabilistic generative model of morphology," in *Proceedings of the 18th Nordic conference of computational linguistics (NODALIDA 2011)*, 2011, pp. 230–237.

[32] S. Virpioja, P. Smit, S.-A. Grönroos, M. Kurimo *et al.*, "Morfessor 2.0: Python implementation and extensions for morfessor baseline," 2013.

[33] A. Saeb, R. Menon, H. Cameron, W. Kibira, J. Quinn, and T. Niesler, "Very low resource radio browsing for agile developmental and humanitarian monitoring," in *Proceedings of Interspeech*, 08 2017, pp. 2118–2122.

[34] R. Menon, A. Saeb, H. Cameron, W. Kibira, J. Quinn, and T. Niesler, "Radio-browsing for developmental monitoring in uganda," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 5795–5799.

[35] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, "The HTK book, version 3.4," 2006.

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