Very low resource radio browsing for agile developmental and humanitarian monitoring

Armin Saeb[1], *Raghav Menon*[1], *Hugh Cameron*[2],
William Kibira[2], *John Quinn*[2], *Thomas Niesler*[1]

[1]Department of Electrical and Electronic Engineering, Stellenbosch University, South Africa
[2]UN Global Pulse, Kampala, Uganda

arsaeb@sun.ac.za, rmenon@sun.ac.za, hcameron@cit.ac.ug, williamkibira@gmail.com,
john.quinn@one.un.org, trn@sun.ac.za

Abstract

We present a radio browsing system developed on a very small corpus of annotated speech by using semi-supervised training of multilingual DNN/HMM acoustic models. This system is intended to support relief and developmental programmes by the United Nations (UN) in parts of Africa where the spoken languages are extremely under resourced. We assume the availability of 12 minutes of annotated speech in the target language, and show how this can best be used to develop an acoustic model. First, a multilingual DNN/HMM is trained using Acholi as the target language and Luganda, Ugandan English and South African English as source languages. We show that the lowest word error rates are achieved by using this model to label further untranscribed target language data and then developing SGMM acoustic model from the extended dataset. The performance of an ASR system trained in this way is sufficient for keyword detection that yields useful and actionable near real-time information to developmental organisations.

Index Terms: developmental and relief monitoring, radio browsing, multilingual deep neural network, semi-supervised training, Acholi, Luganda, Ugandan languages

1. Introduction

In societies with good internet connectivity, social media is a prevalent communication tool. When internet connectivity is insufficient, however, community radio stations hosting phone-in talk shows provide a popular way for citizens to communicate their news and challenges. This is the situation in rural Uganda, for example. The UN has piloted a radio browsing system in this region, with which it monitors such radio discussions to obtain information that can inform the organisation's relief and development programmes [1]. However, this system was developed using an approximately 9-hour long corpus of annotated audio per target language. Although this is regarded as a small corpus from the point of view of acoustic modelling, it nevertheless represents a key obstacle in terms of how quickly the radio browsing system can be deployed. In particular, when a crisis occurs, aid organisations must move very quickly to establish their relief and support strategy. The delay associated with compiling a 9-hour long corpus in the necessary language is unacceptable under these circumstances.

We therefore consider the question of what can best be achieved using resources that can be assembled in a very short time (1-2 weeks). From our experience in compiling the previous corpora in Luganda and Acholi, we estimate that it is reasonable to assume that within the space of one week a corpus of approximately 15 minutes of orthographically annotated speech can be compiled in the new target language. This timespan includes the recruitment and training of a proficient speaker of the target language as transcriber. We also assume that existing corpora in other languages are available to aid acoustic model development.

The combination of Deep Neural Networks (DNNs) and Hidden Markov Models (HMMs) is by now an established architecture in high performance speech recognition systems [2, 3, 4]. However, DNN/HMM acoustic models require many hours of transcribed speech for good performance. Thus the application of DNN/HMM models for under-resourced languages with severely limited training data leads to performance that is inferior to that achieved by more traditional acoustic models such as HMM/GMMs or Subspace Gaussian Mixture models (SGMMs) [5]. Two solutions have been proposed in the literature. The first consists of the use of transcribed data from other better-resourced languages to develop **multilingual acoustic models** that can be refined to the target language using a small training set [6, 7, 8, 9, 10, 11]. The other consists of the incorporation of untranscribed data from the target language into model training. This is generally referred to as **semi-supervised training** [7, 12, 13, 14, 15, 16].

In this paper, we combine multilingual acoustic modelling with DNNs and semi-supervised learning to develop an acoustic model for the radio browsing system using just 12 minutes of transcribed speech in the target language. We also describe the role of filters and human analysts as a final component of the radio browsing system which has been key to its active deployment.

The paper is organized as follows: Section 2 describes the radio browsing system. Sections 3 and 4 provide background on multilingual acoustic modelling using DNN/HMMs and on semi supervised learning. Section 5 describes the data we use and Section 6 the experimental setup. Sections 7 and 8 present experimental results and Section 9 concludes the paper.

2. The radio browsing system

The radio browsing system currently in use includes an automatic speech recognition (ASR) system configured as a keyword spotter (KWS) as shown in Figure 1. Preprocessing includes the detection of segments of acceptable speech from a live audio stream and the rejection of non-speech such as music and singing. The ASR system processes the selected utterances, and the resulting transcriptions (represented as lattices) are searched for keywords of interest. This output is passed to human analysts who filter the information to obtain structured, categorised and searchable information appropriate for

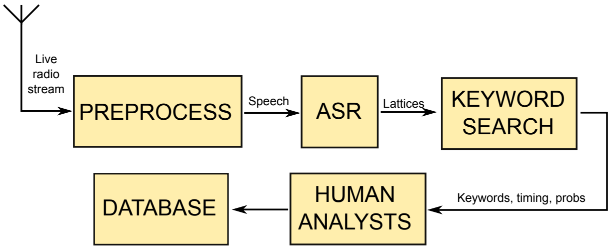

Figure 1: *The radio browsing system.*

humanitarian decision making and situational awareness. Despite high ASR word error rates, this strategy is successful because (a) when topics of relevance are discussed, speakers usually utter several words and phrases related to the topic, and (b) the human analysts can discard false detections. Using this semi-automated approach, a small team of analysts can process a large quantity of audio in near real-time while maintaining high confidence in the final output.

Before reaching the human analysts, a set of topic-level filters are defined to describe keyword and key phrase logic. For example, to identify discussions related to outbreaks of disease in humans, rules such as (‘disease’ $\wedge$ $\neg$ (‘crop’ $\vee$ ‘animal’)) are applied. The locations of matches are used to extract clusters where multiple rules succeed, and to assign relevance scores to each cluster. These relevance scores use Term Frequency/Inverse Document Frequency (TF/IDF) weights associated with each rule to reflect that some rules are more specific and therefore informative than others.

The human analysts assess the detected clusters in two stages. First, the analysts listen to an extract of the original recording and tag it as either a false match (keywords wrongly detected), as not relevant (keywords correctly detected but in a context that was not of interest, e.g. “. . .there has been a *flood* of accusation. . .”), or as relevant. For recordings identified as relevant, there is a second stage in which analysts provide detailed translations and category classifications (e.g. `disaster.flood`, `health.disease-outbreak.cholera`) and type of speaker (e.g. member of public, news-reader, high official, local official). This results in a structured, searchable database.

As an illustration, we now consider three topics analysed in this way: natural disaster, with a particular focus on small-scale disaster that goes unreported in other media; health and disease, particularly to assess experiences of healthcare service delivery; and refugees, to assess public perception of refugees, as during the time of the study Uganda experienced a major influx of refugees fleeing conflict in South Sudan. Some examples of discussion on these topics are shown in Table 1, illustrating the type of insight generated in practice with this system.[1]

3. Multilingual DNN/HMM acoustic models

Multilingual DNN/HMM acoustic models have been proposed as a means of addressing a scarcity of training data in an under-resourced target language [8, 9, 11]. Figure 2 illustrates one architecture that has been proposed to achieve this [6]. The DNN is trained with a relatively large training corpus from multiple languages to provide the class conditional posterior probabilities required by HMMs. In this architecture, the DNN shares hidden layers across all languages, but each language has its own softmax layer and its own HMM.

Table 1: *Examples of relevant discussion extracted by the radio browsing system.*

Topic	**Analyst translation**
natural-disaster, food-security	“Elephants that are suspected to have come from South Sudan went and attacked Abalo Kodi village and destroyed food [crops] about 20 acres.”
refugees.camps	“I stand with my two legs and say that staying in the camps is very very good [...] those days when people were not in the camps they used to keep money in anthills and under the beds, but after coming out of the camps they have knowledge about banking.”
health.service-delivery	“The road here is so bad that the ambulance got stuck in a ditch and could not reach the hospital. People came and had to collect the medicine and carry it on foot to the hospital.”
health.malaria-prevention	“People are using mosquito nets in the wrong way, for example scrubbing their bodies, washing dishes, making fences around chicken houses, some even turkey houses or pigsties.”

Assume that the inputs of the DNN are sequential acoustic feature vectors $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_t, \mathbf{x}_{t+1}, ...$ from all languages and that $\mathbf{x}_t$ belongs to a specific language with output classes c_k $k = 1, ..., K$. The outputs of the softmax layer for that language are the class conditional posterior probabilities:

$$p(c_k|\mathbf{x}_t) \quad k = 1...K \tag{1}$$

The class likelihood can be calculated using Bayes’ rule:

$$p(\mathbf{x}_t|c_k) = \frac{p(c_k|\mathbf{x}_t)p(\mathbf{x}_t)}{p(c_k)} \quad k = 1...K \tag{2}$$

These likelihoods are used by the respective HMM of each language as observation probabilities.

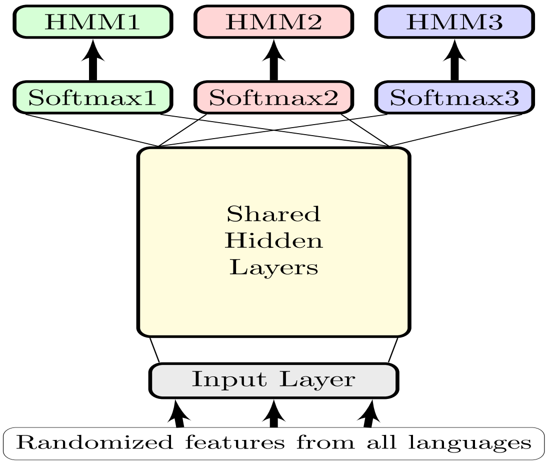

Figure 2: *Multilingual DNN/HMM acoustic model.*

[1] Examples can be accessed online at http://radio.unglobalpulse.net.

4. Semi-supervised learning

If $(\mathbf{x}_i, c_i)$ with $i = 1...L$ denotes the L labelled acoustic vectors $\mathbf{x}_i \in X$ from transcribed utterances with $c_i \in C$ labels and $\mathbf{x}_i$ with $i = L+1, ..., L+U$ denotes the U unlabelled vectors from untranscribed utterances (usually $L << U$), then semi-supervised learning attempts to make use of this combined information to increase the recognition accuracy in comparison with supervised learning using only the labelled acoustic vectors [17]. One straightforward approach to semi-supervised learning is self-training [18]. In this approach, the model is first trained with transcribed training data. This model is subsequently used to transcribe additional unlabelled data [19] . The resulting transcriptions should be filtered so that only those associated with high confidence are selected. This is to avoid the excessive injection of errors into the training set. The filtering may be implemented by only accepting utterances that were decoded with likelihoods above a threshold, or by giving a higher weight to such utterances. The selected labelled data is added to current training set and the process can be repeated until the error rate on a development set no longer improves.

5. Data

The datasets used in our experiments are described in Table 2. Radio broadcasts were recorded in Ugandan English, Luganda and Acholi and transcribed by mother-tongue speakers. This resulted in corpora with 6, 9.6 and 9.2 hours of speech respectively. The speech in these training sets as well as that in the live data presented to the system is fast, highly spontaneous, and often has a poor signal to noise ratio due to background noise and line quality. The Ugandan English was augmented with a 20 hour corpus of South African broadcast news [20]. Of the three languages, Acholi is the most under-resourced, and hence was chosen as the target language for our experiments, with Luganda and English as the source languages for multilingual modelling. From the Acholi data we selected a 12-minute subset containing speech from 4 speakers. The remaining Acholi data was considered untranscribed and used for semi-supervised training. Because of the limited availability of Acholi and Luganda phonetic experts, these pronunciation dictionaries were not as refined as that for English. The available text for developing the Acholi language model was also very limited. The trigram language model used for decoding was obtained with the SRILM toolkit [21].

Table 2: *Datasets used for experimentation*

	Acholi	**Luganda**	**English**
Transcribed train sentences	80	8773	14898
Transcribed train speakers	4	380	889
Transcribed train speech	12m	9.6h	26h
Untranscribed train sentences	4782	—	—
Untranscribed train speakers	199	—	—
Untranscribed train speech	9h	—	—
Test sentences	184	—	—
Test speech	18m	—	—
OOV rate	2.6%	—	—
Vocabulary words	15750	35098	77140
LM sentences	83831	—	—
LM words	1.3M	—	—

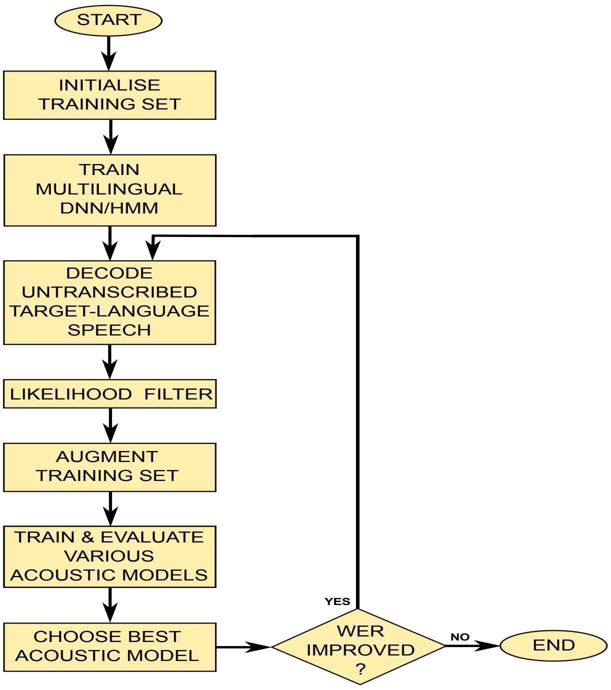
START
INITIALISE TRAINING SET
TRAIN MULTILINGUAL DNN/HMM
DECODE UNTRANSCRIBED TARGET-LANGUAGE SPEECH
LIKELIHOOD FILTER
AUGMENT TRAINING SET
TRAIN & EVALUATE VARIOUS ACOUSTIC MODELS
CHOOSE BEST ACOUSTIC MODEL
WER IMPROVED ?
YES
NO
END

Figure 3: *Acoustic model training strategy.*

6. Experimental setup

Figure 3 illustrates the strategy used to train multilingual DNN/HMM acoustic models with semi-supervised learning. The DNN/HMM architecture used was shown in Figure 2 and includes shared input and hidden layers and three separate softmax layers. The multilingual DNN/HMM model is first trained using transcribed data from non-target languages English and Luganda as well as 12 minutes of Acholi. This model is used to decode the 9 hours of untranscribed Acholi data. The utterances with per-frame log likelihoods above a threshold were selected to augment the training set for subsequent model retraining. The rejected utterances can be considered again in subsequent iterations. The updated labelled training set is then used to train new acoustic models. When the re-trained acoustic models do not deliver improved performance relative to the previous model, the training process is stopped. Due to time limitations, only a single iteration of this process has been performed in this paper. All experiments were conducted using the Kaldi speech recognition toolkit [22]. HMM/GMMs, SGMMs, DNN/HMMs and multilingual DNN/HMMs (MDNN/HMMs) were considered as acoustic modelling approaches. The acoustic model parameters have been optimized to achieve best results. For keyword spotting, the configuration described in [1] was used.

7. ASR results

Table 3 indicates the word error rates (WER) achieved with various acoustic models. The first three lines of the first column indicate the performance achieved by HMM/GMM, SGMM, DNN/HMM acoustic models when trained using only the 12 minutes of transcribed Acholi speech. It is clear that error rates are in all three cases extremely high. The last line of the first column shows the WER achieved by the MDNN/HMM acoustic model, which was trained on 26 hours of English, 9.6 hours

Table 3: *Acholi WER for various acoustic models*

Acoustic model	**%WER (12m)**	**%WER (12m+7h)**	**%WER (9h)**
HMM/GMM	96.06	64.32	48.63
SGMM	92.07	62.60	47.09
DNN/HMM	99.30	65.07	46.99
MDNN/HMM	77.92	64.91	43.56

Luganda and 12 minutes of Acholi. The inclusion of data from the other languages has led to an absolute drop in error rate between 14% and 21%. The MDNN/HMM model was subsequently used to decode the untranscribed 9 hours of Acholi speech. Of these 9 hours, 2 hours were excluded by the log-likelihood filtering, and 7 hours were added to the existing 12 minutes of Acholi training data. Acoustic models were retrained on this augmented training set and their performance is shown in the second column of Table 3. We see that the semi-supervised training has afforded an absolute WER improvement of between 13% and 34% relative to the first column. It is interesting to see that the SGMM offers the best performance in this case, and not the MDNN/HMM. Finally, the third column shows the performance achieved if all the transcriptions of the 9 hours of Acholi speech are used for training. The figures in this column can be regarded as an upper bound of what is achievable by semi-supervised training. Taking this interpretation, we see that between 38% (for MDNN/HMM) and 66% (for other 3 models) of the possible improvement has been achieved by semi-supervised training.

8. Keyword spotting results

The final set of experiments concern keyword search using some of the acoustic models in Table 3. Using the setup introduced in [1], the lattices generated by the speech recognizer are passed to the keyword search. The performance of the keyword spotter was evaluated using the NIST oracle measures, Actual Term Weighted Value (ATWV) and Maximum Term Weighted Value (MTWV). Table 4 shows the ATWV and MTWV values while Figure 4 shows the detection error tradeoff (DET) curves for various keyword spotting systems.

As expected, best keyword spotting performance is achieved by systems that were trained on all 9 hours of manually transcribed data (column 3, Table 3). All three systems trained in this way (SGMM, DNN/HMM and MDNN/HMM) show high ATWV and MTWV, with MDNN/HMM faring the best although only by a small margin. This is also reflected in the DET curves in Figure 4, where the MDNN/HMM shows the best performance. We also note from Table 4 that, when only 12m of data is available for training, the ATWV of the KWS system is negative and hence very poor. The MTWV, which reflects the best performance achievable when the decision threshold is calibrated, indicates that threshold optimisation can improve the system performance but that the performance remains poor. This is also reflected in the DET curve in Figure 4. Relative to the 12-minute SGMM and MDNN/HMM systems, the 12m+7h SGMM system affords a substantial improvement in terms of ATWV, MTWV and also the DET curve. We can also see that the performance of the KWS system is in line with the WERs in Table 3. Improvements in the WER also lead to improvements in the performance of the KWS system.

Table 4: *Keyword spotting performance for various systems*

Acoustic model	**ATWV**	**MTWV**
SGMM (9h)	0.5844	0.6484
DNN/HMM (9h)	0.5221	0.6857
MDNN/HMM (9h)	0.6400	0.6843
SGMM (12m)	-0.1872	-0.0356
MDNN/HMM (12m)	0.0415	0.0726
SGMM (12m+7h)	0.1430	0.2541

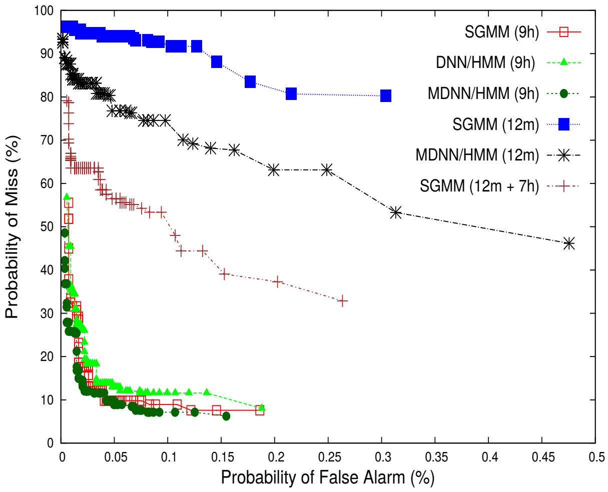

Figure 4: *DET curves for various systems.*

9. Discussion, summary and conclusion

We have shown that by combining semi-supervised learning and multilingual DNN/HMM acoustic modelling it is possible to achieve substantial improvements in performance over monolingual systems in a very low resource scenario. Using just 12 minutes of transcribed speech in the target language (Acholi), it was possible to train a multilingual acoustic model (Acholi, Luganda and English) that could be used supplement the Acholi training set for subsequent semi-supervised learning, leading to a further substantial performance gain. The best system exhibited a word error rate of 62.6%, which is a promising result in the light of the very small target language training corpus.

It remains to be seen in ongoing work by how much the performance can be further improved by increasing the pool of untranscribed target language data. The effect of multiple iterations of semi-supervised learning, as shown in Figure 3, must also still be assessed. Nevertheless, the framework presents a feasible way of implementing a radio browsing system in a very short space of time by requiring a minimal amount of annotated training material in the target language. Since the setup time is currently a key obstacle to the deployment of the radio browsing system in new crisis areas, this opens the door to more widespread incorporation into humanitarian relief efforts.

10. Acknowledgements

We thank Jonathan Mukiibi, Joseph Kambarage and Kirsten Van Camp for analysing the filter output and providing the examples. We gratefully acknowledge the support of NVIDIA corporation with the donation GPU equipment used for this research. We also gratefully acknowledge the support of Telkom South Africa.

11. References

[1] R. Menon, A. Saeb, H. Cameron, W. Kibira, J. Quinn, and T. Niesler, “Radio-browsing for developmental monitoring in uganda,” in *Proc. ICASSP*, 2017.

[2] G. E. Dahl, D. Yu, L. Deng, , and A. Acero, “Context-dependent pre-trained deep neural networks for large vocabulary speech recognition,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 20, no. 1, pp. 30–42, 2012.

[3] F. Seide, G. Li, and D. Yu, “Conversational speech transcription using context-dependent deep neural networks,” in *Proc. INTERSPEECH*, 2011, pp. 437–440.

[4] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, “Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups,” *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, 2012.

[5] D. Yu and L. Deng, *Automatic Speech Recognition: A Deep Learning Approach*. Springer, 2015.

[6] J. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, “Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers,” in *Proc. ICASSP*, 2013, pp. 7304–7308.

[7] S. Sitaram, S. Palkar, Y. N. Chen, A. Parlikar, and A. W. Black, “Bootstrapping text-to-speech for speech processing in languages without an orthography,” in *Proc. ICASSP*, 2013, pp. 7992–7996.

[8] F. Grezl, M. Karafiat, and K. Vesel, “Adaptation of multilingual stacked bottle-neck neural network structure for new language,” in *Proc. ICASSP*, 2014, pp. 7654–7658.

[9] A. Mohan and R. Rose, “Multi-lingual speech recognition with low-rank multi-task deep neural networks,” in *Proc. ICASSP*, 2015, pp. 4994–4998.

[10] J. Cui *et al.*, “Multilingual representations for low resource speech recognition and keyword search,” in *Proc. ASRU*, 2015, pp. 259–266.

[11] R. Sahraeian and D. Compernolle, “A study of rank-constrained multilingual dnns for low-resource asr,” in *Proc. ICASSP*, 2016, pp. 5420–5424.

[12] K. M. Knill, M. J. Gales, A. Ragni, and S. Rath, “Language independent and unsupervised acoustic models for speech recognition and keyword spotting,” in *Proc. INTERSPEECH*, 2014.

[13] V. Manohar, D. Povey, and S. Khudanpur, “Semi-supervised maximum mutual information training of deep neural network acoustic models,” in *Proc. INTERSPEECH*, 2015, pp. 2630–2634.

[14] H. Kamper, A. Jansen, and S. Goldwater, “Fully unsupervised small-vocabulary speech recognition using a segmental bayesian model,” in *Proc. INTERSPEECH*, 2015, pp. 678–682.

[15] F. Metze, A. Gandhe, Y. Miao, Z. Sheikh, Y. Wang, D. Xu, H. Zhang, J. Kim, I. Lane, W. K. Lee, S. Stuker, and M. Muller, “Semi-supervised training in low-resource asr and kws,” in *Proc. ICASSP*, 2015, pp. 4699–4703.

[16] J. T. Springenberg, “Unsupervised and semi-supervised learning with categorical generative adversarial networks,” in *Proc. International Conference on Learning Representations(ICLR)*, 2016.

[17] X. Zhu, “Semi-supervised learning literature survey,” Computer Sciences, University of Wisconsin-Madison, Tech. Rep. 1530, 2005.

[18] M. A. Hasegawa-Johnson, P. Jyothi, D. McCloy, M. Mirbagheri, G. M. di Liberto, A. Das, B. Ekin, C. Liu, V. Manohar, H. Tang, E. C. Lalor, N. F. Chen, P. Hager, T. Kekona, R. Sloan, and A. K. C. Lee, “ASR for under-resourced languages from probabilistic transcription,” *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 25, no. 1, 2017.

[19] I. Triguero, S. Garcia, and F. Herrera, “Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study,” *Knowledge and Information Systems*, vol. 42, no. 2, pp. 245–284, 2015.

[20] H. Kamper, F. D. Wet, T. Hain, and T. Niesler, “Capitalising on north american speech resources for the development of a south african english large vocabulary speech recognition system,” *Computer Speech and Language*, vol. 28, no. 6, pp. 1255–1268, 2014.

[21] A. Stolcke, “SRILM An extensible language modeling toolkit,” in *Proc. ICSLP*, Denever-Colorado, 2002, pp. 901–904.

[22] 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 *Proc. ASRU*, 2011.