SLTU-2014, St. Petersburg, Russia, 14-16 May 2014

CROSS-LANGUAGE MAPPING FOR SMALL-VOCABULARY ASR IN UNDER-RESOURCED LANGUAGES: INVESTIGATING THE IMPACT OF SOURCE LANGUAGE CHOICE

Anjana Vakil and Alexis Palmer

University of Saarland
Department of Computational Linguistics and Phonetics
Saarbrücken, Germany

ABSTRACT

For small-vocabulary applications, a mapped pronunciation lexicon can enable speech recognition in a target under-resourced language using an out-of-the-box recognition engine for a high-resource source language. Existing algorithms for cross-language phoneme mapping enable the fully automatic creation of such lexicons using just a few minutes of audio, making speech-driven applications in any language feasible. What such methods have not considered is whether careful selection of the source language based on the linguistic properties of the target language can improve recognition accuracy; this paper reports on a preliminary exploration of this question. Results from a first case study seem to indicate that phonetic similarity between target and source language does not significantly impact accuracy, underscoring the language-independence of such techniques.

Index Terms— under-resourced languages, speech recognition, lexicon building, phoneme mapping

1. INTRODUCTION

In recent years it has been demonstrated that speech recognition interfaces can be extremely beneficial for applications in the developing world, particularly in communities where literacy rates are low or where PCs and internet connections are not always available [1, 2, 3]. Typically, the languages spoken in such communities are under-resourced, such that the large audio corpora typically needed to train or adapt recognition engines are unavailable. However, in the absence of a recognition engine trained for the target under-resourced language (URL), an existing recognizer for a completely unrelated high-resource language (HRL), such as English, can be used to perform small-vocabulary recognition tasks in the URL. All that is needed is a pronunciation lexicon mapping each term in the target vocabulary to one or more sequences of phonemes in the HRL, i.e. phonemes which the recognizer can model.

While the mapped pronunciations could be hand-written by an expert linguist familiar with the two languages, algorithms such as the "Salaam" method [3, 4, 5] can create these pronunciations automatically from just a few minutes of data, and have been shown to yield higher recognition accuracy than is achieved with hand-coded pronunciations [3, 4]. The automatic technique also has the advantage of not depending on any expert knowledge of the source or target language or the relationship between them.

However, it is conceivable that the recognition accuracy for a given target URL will vary depending on the source HRL used, as the source/target combination will determine the degree to which the sound systems of the two languages differ, and thus the difficulty of the pronunciation mapping task. More specifically, we expect that by carefully selecting the source language such as to maximize the overlap between its phoneme inventory and that of the target (under-resourced) language, we can reduce the difficulty of phoneme mapping and thereby find better pronunciation sequences for the target terms, which should lead to increased accuracy in the recognition task.

We have begun to test this hypothesis by comparing recognition results for pronunciations generated for words in a target URL (Yoruba) using the Salaam method with two different source HRL recognizers (English and French). The aim of this paper is to present our experiment and findings, and discuss their implications for language-mapping techniques for language-independent small-vocabulary ASR.

2. BACKGROUND AND RELATED WORK

Many commercial speech recognition systems offer high-level Application Programming Interfaces (APIs) that make adding voice recognition capabilities to an application as simple as specifying (in text) the words/phrases that should be recognized; this requires very little general technical expertise, and virtually no knowledge of the inner workings of the recognition engine. If the target language is supported by the system – Microsoft's Speech Platform, for example, currently supports recognition and synthesis for 26 languages/dialects [6] – this makes it very easy for small-scale software developers (i.e. individuals or small organizations without much funding) to create new speech-driven applications.

While many such individuals or organizations in the developing world may be interested in using such platforms to create speech-driven applications for use in their communities, the under-resourced languages typically spoken in these areas are generally not supported by such commercial systems. And though many effective techniques for training or adapting recognizers for new languages exist (e.g. [7, 8]), these typically require hours of training audio to produce effective models, and even the highest-level tools for building new models still require a nontrivial amount of expertise with speech technologies; such data and expertise may not be available to the small-scale developers in question.

However, many useful development-oriented applications (e.g. for accessing information or conducting basic transactions) require only small-vocabulary recognition tasks, by which we mean those requiring discrimination between a few dozen terms. For such tasks, an unmodified HRL recognizer can be used as-is to perform recognition of the URL terms; we simply need a suitable pronunciation lexicon.

This is the thinking behind the Speech-based Automated Learning of Accent and Articulation Mapping (Salaam) method [3, 4, 5], which provides the foundation for the original research presented in the following sections.

The basic idea is to discover the best pronunciation sequence for a given word in the target language by using the source language recognizer to perform phone decoding on one or more audio samples of the target word. However, the APIs for commercial recognizers such as Microsoft's are designed for word-decoding, and do not usually enable the use of the phone-decoding mode. The insight of the Salaam approach is to use a specially designed grammar to mimic this phone decoding [5, §3.2]. Specifically, Qiao et al. [4, 4.1] create a recognition grammar representing a phoneme "super-wildcard" that guides pronunciation discovery. This grammar allows the recognizer to treat an audio sample of the target word as a "phrase" made up of 0-10 "words", where each "word" can be matched to any possible sequence of 1, 2, or 3 source language phonemes [4, 4.1].

Given this super-wildcard grammar and one or more audio recordings of the target word, Qiao et al. [4, 4.1] use an iterative training algorithm to discover the best pronunciation(s) for that word, one phoneme at a time. In the first pass, the recognizer finds the best match(es) for the first phoneme, then for the first two phonemes in the second pass, and so on until a stopping criterion is met, e.g. the recognition confidence score assigned to the resulting "phrase" stops improving [4, p. 4].

Compared to expert-crafted pronunciations, using pronunciations generated automatically by this algorithm improves recognition accuracy substantially [4, 5.2]. By training on samples from two speakers instead of one, and by using a pronunciation lexicon containing multiple pronunciations for each word (i.e. the n-best results of the training algorithm instead of the single best result), Qiao et al. are able to further improve accuracy. In later work, Chan and Rosenfeld [5] achieve even higher accuracy by applying an iterative discriminative training algorithm, identifying and removing pronunciations that cause confusion between word types.

In sum, the Salaam method is fully automatic, requiring expertise neither in speech technology (to modify acoustic models) nor in linguistics (to manually generate seed pronunciations), and for each new target language it requires only a few minutes' worth of training data from one or two speakers, an amount that can be collected in a short time with little effort or expense. At the same time, it provides pronunciation lexicons that can help bring speech recognition applications to URLs. This has been demonstrated by at least two developing-world projects that have successfully used the Salaam method to add voice interfaces to real applications: an Urdu telephone-based health information system in Pakistan [3], and a text-free Hindi smartphone application to deliver agricultural information to farmers in rural India [2].

Our work takes the Salaam method as its foundation, and directly builds on this approach. All previous work using the Salaam method uses English as the high-resource language. In this paper, we begin investigating whether linguistically-motivated selection of the source high-resource language can improve recognition accuracy.

3. EXPERIMENT

Despite the proven success of the Salaam method, there is still room for improvement; Qiao et al. call for word recognition accuracy rates greater than 95% for real-world applications, yet are not quite able to achieve that, reporting e.g. less than 85% accuracy for a 30-word vocabulary in Yoruba [4, p. 5] (no improved Yoruba results are reported in [5]). We hypothesize that careful selection of the source language used for a given target language could yield additional improvement in accuracy, by potentially reducing the number of "mismatches" between phonemes of the two languages, and thus reducing the difficulty of the mapping task.

3.1. Target and Source Languages

For the target language in our research we use Yoruba, a language of the Niger-Congo family with approximately 20-30 million speakers in Nigeria, and additional speakers in neighboring countries such as Benin and Togo [9, 10].

Given this target language, our selection of source language(s) for this research should be informed by an understanding of the Yoruba sound system and how it compares to the HRLs for which high-quality recognition engines are available. With respect to phonetics and phonology, Yoruba has one major difference from the high-resource European languages that are typically the focus of development of speech recognition systems: tone. Yoruba makes use of three

Table 1: Segmental phonemes of Yoruba compared with French and English

Shared with	Phoneme
French only	e*, a*, u*, o*, ɛ̃, ɔ̃/ã,
English only	h, ɾ (as intervocalic /t/)
Both	i, ɛ, ɔ, b, t, d, k, g, f, s, ʃ, m, l, j, w
Neither	ĩ, ũ, ɟ, k͡p, g͡b

* Realized in standard American English not as a pure vowel, but always with a strong offglide or as the first sound in a diphthong.

contrastive tones, low, medium and high, to distinguish many words and phrases that are otherwise identical [9, p. 869]. While this presents a difficult and interesting problem for speech recognition, it is beyond the scope of this paper; in the small vocabulary used for our work, there are no pairs of lexical items distinguished by tone alone.

However, there are also differences at the segmental level, and these will be our focus here (see Table 1 for an overview of Yoruba phonemes). In comparison with American English, which until now has been the only source (high-resource) language tested with the Salaam system [3, 4, 5, 2], Yoruba's phonemic inventory includes an additional class of segments, namely the nasal vowels. As nasal vowels do not occur in English, we chose another high-resource European language in which they do – French – as the basis for comparison.

French is traditionally said to have four nasal vowels: [ɛ̃], [ã], [ɔ̃], and [œ̃], though many speakers replace [œ̃] with [ɛ̃] [11]. All three of the main French nasal vowels occur in Yoruba, though [ã] and [ɔ̃] are variants of the same phoneme, /ɔ̃/ [9, p. 868]. Additionally, Yoruba has the nasal vowels [ĩ] and [ũ] (ibid.), which are not phonemic in French.

Yoruba's consonant inventory overlaps to a large extent with those of both English and French. It has a few segments which occur in neither language, namely the doubly articulated labial-velar stops [k͡p] and [g͡b] and the palatal stop [ɟ]. There are two Yoruba consonants, the glottal fricative [h] and the alveolar tap [ɾ] which occur in English, but not French. The English recognizer may thus be at an advantage when it comes to Yoruba consonants, while the French recognizer should have the advantage with vowels and overall, as Yoruba shares more vowels with French than consonants with English.

We therefore hypothesize that the Salaam method for pronunciation mapping will yield higher accuracy pronunciations for Yoruba words when using French as the source language. As the following sections describe, we test this hypothesis by reimplementing the Salaam method described by Qiao et al. [4], and comparing the word recognition accuracy using French and English recognizers.

3.2. Data

As training and testing data for our system we use a 25-word subset of the Yoruba data collected by Qiao et al. [4, 5.1]. For each term, we have five telephone-quality audio samples recorded by each of two speakers, one female and one male. Although some of these samples include noise or artifacts that might complicate the recognition process, such samples are not excluded from the training or testing data, as they reflect the type of recording errors that could reasonably be expected in real applications.

3.3. Method

In our implementation of the Salaam method we use English (US) and French (France) recognizers developed by Microsoft for server-side recognition of telephone-quality audio, accessible via the Microsoft Speech Platform SDK 11 [6]. This system was chosen for its robustness and to maintain comparability with the results obtained by Qiao et al. [4] and Chan and Rosenfeld [5], who also worked with Microsoft's server-side recognizers. In keeping with the overall objective of the Salaam approach, the recognizers are used as-is, with no modifications to underlying models.

In the training phase, pronunciations are generated for each Yoruba word in our vocabulary from a set of audio samples, using the Salaam algorithm [4, 4.1]. Following Chan and Rosenfeld [5, p. 2], we take all sequences returned in a given pass as input for the following pass, and we slightly modify the algorithm's stopping condition to terminate when the top-scoring phoneme sequence for a given word does not change for three consecutive iterations. As an alternative stopping condition, following Qiao et al. [4, p. 4], we also stop iterations if the best result from the i^{th} pass has a lower score than the best result of the $i - 1^{th}$ pass (with the $i - 1^{th}$ results returned as the best pronunciations). In both cases, at least three passes are required. To determine the best results for a given word from each pass, the set of results for all training samples of that word is sorted by the total confidence score assigned to each pronunciation (phoneme sequence). If a given pronunciation matches more than one sample, the overall score for that sequence in that pass is simply the sum of confidence scores for all samples it was associated with. In the work reported here, we do not perform discriminative training [5] on the discovered pronunciations.

After the iterative training has completed, the 1-, 3-, or 5-best pronunciation sequences for each of the 25 Yoruba words are used to construct a pronunciation lexicon, which is in turn used to perform recognition in the testing phase. To compare the impact of the choice of source (high-resource) language, we analyze the same-speaker and cross-speaker word recognition accuracy for our vocabulary using the English and French recognizers. In each condition we test using a lexicon with 1, 3, or 5 pronunciations per word; when the lexicon contains multiple pronunciations per word, the recognizer will match

Table 2: Difference in accuracy using French and English recognizers.

	Pronunciations	Mean Accuracy (%)		Difference	p-value
		French	English		
Same-speaker	1	75.2	80.0	-4.8	0.1995
	3	77.2	80.0	-2.8	0.3434
	5	80.0	81.6	-1.6	0.5911
Cross-speaker	1	60.0	63.2	-3.2	0.4097
	3	64.8	71.6	-6.8	0.0374*
	5	61.6	73.6	-12.0	0.0424*

* Significant difference ($p < 0.05$).

any of the given pronunciations to that word, making no distinction or preference among them.

To determine same-speaker accuracy for each of the two speakers, we perform a leave-one-out evaluation on the five samples recorded per word per speaker (this amounts to a five-fold evaluation, reserving one sample per word per speaker for testing, and training on the other four). Cross-speaker accuracy is evaluated by training the system on all five samples of each word recorded by one speaker, and testing on all five samples from the other speaker. We use the R software environment [12] for statistical analysis and visualization of the results.

4. RESULTS AND DISCUSSION

The overall results are summarized in Table 2. The results of the same-speaker evaluation are outlined in Figure 1, while Figure 2 illustrates those of the cross-speaker evaluation.

As these figures show, the results of the comparison are the reverse of what we expect: in both same-speaker and cross-speaker evaluations, and for all numbers of pronunciations, the mean accuracy obtained with the French recognizer is lower than that of the English recognizer.

It should also be noted that the single-pronunciation same-speaker leave-one-out accuracy for the English recognizer is, at 80%, less than the corresponding accuracy of approximately 85% reported by Qiao et al. [4, p. 5]. We speculate that this may be due to differences in our implementation (such as the modification of the stopping condition), or perhaps to changes to the models underlying Microsoft's recognition engine made in the intervening years.

Paired two-tailed t-tests comparing the results from the two different source languages (see Table 2) reveal that the difference in same-speaker accuracy is not statistically significant (given a significance level of $p = 0.05$). The difference in mean cross-speaker accuracy, however, appears significant for the systems allowing 3 and 5 pronunciations per word, although we hesitate to conclude much from this given the very small sample size (only two speakers). Furthermore, it should be noted that the two speakers are of different genders, so it is possible that this factor may interact with source language choice to influence the cross-speaker accuracy.

Also of interest in the cross-speaker results is the fact that for the French system, the highest accuracy is observed using 3 pronunciations per word, while for the French same-speaker evaluation and both evaluations using the English recognizer, accuracy is highest using the maximum number of pronunciations per word (5). This indicates that in this case, the fourth and fifth pronunciations added introduce the type of "confusion" that can be reduced through discriminative training [5].

As explained in Section 3.1, in this work we select French as an alternative source language based on the hypothesis that since English does not have nasal vowels in its phonemic inventory, this class of Yoruba sounds might be better captured by a recognizer for French, which does have such phonemes. However, consider Table 3, which lists the recognition accuracy for each word type – i.e. the percentage of all test samples of that word type which are recognized correctly, over both same- and cross-speaker evaluations – with the highest-accuracy word types listed first. Also listed is the mean confidence score as reported by the recognizer for samples of each word type. Inspecting this data, we notice that of the words which contain nasal vowels (indicated by a vowel followed by "n" in the orthographic form), several are quite problematic for the French recognizer, indicated by their low accuracy and correspondingly low positions in the table (e.g. "mesan", "ookan", and "sun" with 51.67% accuracy). While the English recognizer also struggles with some of these words (e.g. "sun", 35%, and "ogorun", 53%), overall it does not seem to do significantly worse with these words than the French recognizer. In fact, the average rank of all 10 words in the dataset which contain a nasal vowel is 14.1 for the English recognizer, and for the French recognizer it is not much better at 13.4. If the French recognizer were truly better able to

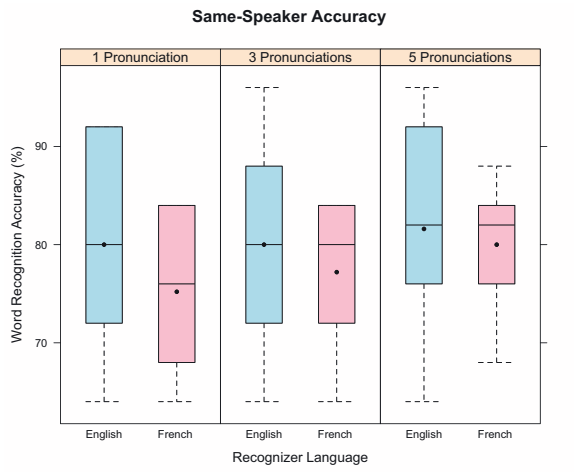

Fig. 1: Same-speaker leave-one-out word recognition accuracy for Yoruba. Mean values are represented by a dot, medians by a horizontal line.

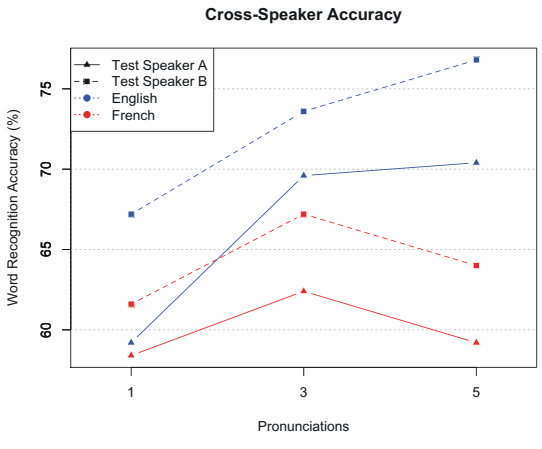

Fig. 2: Cross-speaker word recognition accuracy for two Yoruba speakers. Speaker A is the female speaker, speaker B the male.

model Yoruba nasal vowels than the English one, we would expect to observe a greater difference in accuracy for words containing these sounds.

The results would therefore seem to indicate that in the task of generating pronunciations for small-vocabulary recognition in Yoruba, the French recognizer at best performs no better than the English recognizer, and at worse yields significantly lower accuracy. However, before we completely reject our original hypothesis – that choosing a source language whose phoneme inventory overlaps more with that of the target language will result in higher recognition accuracy – we should consider several other plausible explanations.

First of all, it is possible that our choice of source languages to compare in this work is not an ideal one. As illustrated by Table 1, though French does seem to share more phonemes with Yoruba than English does, especially (nasal) vowels, most Yoruba phonemes are present in either both or neither of the two languages, and the number of phonemes that occur in one source language only is relatively small. So it is possible that the difference between English and French with respect to Yoruba is not a very significant one, and that this is reflected in the largely insignificant differences in recognition accuracy between the two systems.

Secondly, it might be the case that the two recognition systems we compared are not of the same quality to begin with, even though they were developed by the same organization and presumably with similar general techniques. It is conceivable that more data, time, and/or effort has gone into the development of Microsoft's English recognizer than the French one, and that the English system is therefore more robust, leading to better recognition performance in general. However, in the absence of data on the baseline accuracy of Microsoft's French and English recognizers, this explanation remains purely speculative.

Finally, it should be noted that as mentioned in Section 3.3 above, our implementation of the Salaam method does not make use of the discriminative training algorithm which Chan and Rosenfeld [5] demonstrate to be effective at improving word recognition accuracy. Inspecting the mean confidence scores for each word type, listed in Table 3, we observe that for the French recognizer, even word types which were recognized inaccurately were associated with high confidence scores: see e.g. "igba". A comparison of the confidence scores associated with correct and incorrect recognitions, given in Table 4, confirms that the English recognizer reports much higher confidence scores for the words it recognizes correctly than for those it recognizes incorrectly, as we might expect; with the French recognizer, however, there is a much smaller gap between the scores reported for accurate and inaccurate recognitions. It seems plausible that if discriminative training were applied to reduce the "confusion" between pronunciations for different word types as much as possible, a more noticeable difference in accuracy between source lan-

Table 4: Mean confidence scores reported by the recognition engine, for correct and incorrect recognition results.

	English	French
Correctly recognized words	0.7378	0.8443
Incorrectly recognized words	0.4865	0.7402

Table 3: Overall recognition accuracy and mean confidence scores by word type.

(a) English recognizer

Rank	Word	% Correct	Mean confidence
1	duro	100.00	.8298
1	ogba	100.00	.7852
1	shii	100.00	.7758
1	ogoji	100.00	.7743
5	mesan	98.33	.8752
5	beeni	98.33	.7788
7	fisii	95.00	.7475
7	mejo	95.00	.7004
9	mewa	93.33	.6341
10	tunse	91.67	6767
11	bere	83.33	.7133
11	lehin	83.33	.6383
13	meje	80.00	.5927
14	ookan	75.00	.5277
15	ogun	70.00	.6074
16	tele	66.67	.4418
17	marun	63.33	.6863
18	merin	61.67	.7039
19	mefa	60.00	.6869
20	iba	56.67	.6405
20	igba	56.67	.6237
22	ogorun	53.33	.7142
23	meta	48.33	.6552
24	sun	35.00	.4716
25	meji	11.67	.5977

(b) French recognizer

Rank	Word	% Correct	Mean confidence
1	ogba	100.00	.8686
2	iba	93.33	.8727
2	mejo	93.33	.8067
4	ogoji	91.67	.8367
5	lehin	90.00	.8366
6	tunse	88.33	.8319
7	marun	86.67	.8161
8	duro	85.00	.8160
8	fisii	85.00	.7689
10	ogun	81.67	.8719
11	shii	76.67	.8614
11	tele	76.67	.7177
13	meje	75.00	.8407
14	ogorun	73.33	.8662
15	mefa	71.67	.8503
16	beeni	68.33	.8468
16	merin	68.33	.7906
18	mewa	61.67	.8777
19	meta	60.00	.8076
20	mesan	51.67	.7834
20	ookan	51.67	.7335
20	sun	51.67	.7134
23	meji	28.33	.8312
24	bere	25.00	.6535
25	igba	10.00	.8209

guages might emerge, since the remaining errors might have less to do with similarities between words in the data set and more to do with inefficient mapping between target language words and source language phonemes. We are undertaking this investigation in current work, as described in Section 5.

5. CONCLUSION AND ONGOING/FUTURE WORK

This paper has presented an extension of the Salaam method for pronunciation mapping [3, 4, 5], which enables the creation of pronunciation lexicons for small-vocabulary recognition tasks in a target under-resourced language using a pre-existing recognition engine for a high-resource source language. We have conducted a preliminary investigation of the impact of source language choice on recognition accuracy in a target under-resourced language, using Yoruba as the target language and comparing English and French as source languages. Our original hypothesis was that replacing English as the source language with a high-resource language whose phoneme inventory is more similar to that of the target language would increase accuracy; the results of this study do not support that hypothesis. However, further research involving different combinations of source and target language is necessary before drawing firm conclusions in either direction, and this investigation is an important part of our plans for future work. If the result holds, and swapping source languages never leads to increased performance on target language recognition, this would itself be an interesting result, as it would support the language-independent nature of mapping algorithms such as Salaam.

This paper therefore describes work in progress; some obvious next steps are currently being implemented, while others are planned for future research.

One of our current priorities is the application of discriminative training following Chan and Rosenfeld [5]. As discussed in the previous section, it is possible that eliminating errors in this way would reveal more significant differences between systems trained on different source languages.

The main reason we did not implement this discriminative training in the work reported here was the long running time

of the super-wildcard-grammar algorithm; In the reported implementation, generating the pronunciation lexicon for a single word takes at least several minutes, mainly due to the size of the long wildcard grammars used in the algorithm. Since the size of the grammars increases as more phonemes are discovered, this takes even more time for longer words/phrases. Therefore, we have been experimenting with modifying the algorithm to make it more time-efficient. Preliminary findings, to be reported in upcoming work, reveal that shortening the super-wildcard grammar can yield huge reductions in training time with no significant loss to accuracy.

Another important step we are currently undertaking is the development of a GUI-based PC application based on the Salaam method, intended as a tool to enable non-expert users to create and evaluate lexicons quickly and simply, as Chan and Rosenfeld also suggest [5]. It is our hope that this open-source tool, *lex4all*,[1] will be of great use to individuals or small groups who wish to research or develop speech-driven applications in under-resourced languages.

Regarding our directions for future research, the most obvious next step is to evaluate a much wider range of source and target languages. While the results reported here serve as a good starting point for research into the relationship between source-language selection and recognition accuracy, broad conclusions can of course not be drawn until we have examined a larger, typologically diverse sample of high- and low-resource languages and their combinations. This research will be greatly facilitated by the aforementioned lexicon-building and evaluation tool.

Finally, to further improve recognition accuracy once discriminative training has been implemented, it might be worthwhile to find a better heuristic for combining the recognition results from multiple audio samples in each pass of the training algorithm. In our implementation (see Section 3.3), even if a given pronunciation is a match for multiple samples, it can be overshadowed by another sequence matching only one sample but with very high confidence; this could amount to overfitting on the training data. To make the system more robust, we should favor pronunciations which match as many different utterances of a word as possible, instead of those which match only a single utterance very well. Therefore, we intend to examine weighting schemes which prioritize pronunciation sequences matching multiple samples.

6. ACKNOWLEDGEMENTS

This research was partially supported by a Deutschlandstipendium scholarship, sponsored by IMC AG, granted to the first author. We are grateful to Roni Rosenfeld, Hao Yee Chan, and Mark Qiao for generously sharing their data, and for their valuable advice. We also thank Dietrich Klakow and the three anonymous reviewers for their feedback.

[1] http://lex4all.github.io/lex4all

7. REFERENCES

[1] J Sherwani and Roni Rosenfeld, "The case for speech technology for developing regions," in *Proc. HCI for Community and International Development, Florence, Italy*. 2008, ACM.

[2] Kalika Bali, Sunayana Sitaram, Sebastien Cuendet, and Indrani Medhi, "A Hindi speech recognizer for an agricultural video search application," in *Proceedings of the 3rd ACM Symposium on Computing for Development*, New York, NY, USA, 2013, ACM DEV '13, pp. 5:1–5:8, ACM.

[3] Jahanzeb Sherwani, *Speech interfaces for information access by low literate users*, Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA, USA, 2009.

[4] Fang Qiao, Jahanzeb Sherwani, and Roni Rosenfeld, "Small-vocabulary speech recognition for resource-scarce languages," in *Proceedings of the First ACM Symposium on Computing for Development*, New York, NY, USA, 2010, ACM DEV '10, pp. 3:1–3:8, ACM.

[5] Hao Yee Chan and Roni Rosenfeld, "Discriminative pronunciation learning for speech recognition for resource scarce languages," in *Proceedings of the 2nd ACM Symposium on Computing for Development*, New York, NY, USA, 2012, ACM DEV '12, pp. 12:1–12:6, ACM.

[6] Microsoft, "Language support," in *Microsoft Speech Platform SDK 11 Documentation*. 2012, http://msdn.microsoft.com/en-us/library/hh378476.

[7] "CMUSphinx: Open source toolkit for speech recognition," http://www.cmusphinx.org.

[8] Tanja Schultz and Alan W Black, "Multilingual Speech Processing – Rapid Language Adaptation Tools and Techniques," in *INTERSPEECH 2010*, 2010.

[9] Douglas Pulleybank and Olanike Ola Orie, "Yoruba," in *The World's Major Languages*, Bernard Comrie, Ed., chapter 51, pp. 866–882. Routledge, 2 edition, 2009.

[10] Felix A Fabunmi and Akeem Segun Salawu, "Is Yoruba an Endangered Language?," *Nordic Journal of African Studies*, vol. 14, no. 3, pp. 391–408, 2005.

[11] Cécile Fougeron and Caroline L. Smith, "French," in *Handbook of the International Phonetic Association*, International Phonetic Association, Ed. Cambridge University Press, 1999.

[12] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2013, http://www.R-project.org.