ASR for Documenting Acutely Under-Resourced Indigenous Languages

Robbie Jimerson, Emily Prud'hommeaux
Rochester Institute of Technology
Rochester, New York, USA
rcj2772,emilypx@rit.edu

Abstract

Despite its potential utility for facilitating the transcription of speech recordings, automatic speech recognition (ASR) has not been widely explored as a tool for documenting endangered languages. One obstacle to adopting ASR for this purpose is that the amount of data needed to build a reliable ASR system far exceeds what would typically be available in an endangered language. Languages with highly complex morphology present further data sparsity challenges. In this paper, we present a working ASR system for Seneca, an endangered indigenous language of North America, as a case study for the development of ASR for acutely low-resource languages in need of linguistic documentation. We explore methods of leveraging linguistic knowledge to improve the ASR language models for a polysynthetic language with few high-quality audio and text resources, and we propose a tool for using ASR output to bootstrap new data to iteratively improve the acoustic model. This work serves as a proof-of-concept for speech researchers interested helping field linguists and indigenous language community members engaged in the documentation and revitalization of endangered languages.

Keywords: low-resource languages, automatic speech recognition, indigenous languages

1. Introduction

By the end of this century, it is estimated that at least half and as many as 90% of the world's nearly seven thousand languages will be extinct (Krauss, 1992; Crystal, 2000). With each language that is lost, we lose insight not only into the culture of the the people who spoke that language, but also into the characteristics of that language that can shed light on the underlying structure of human language. Some communities on the verge of losing their language are engaged in preservation efforts, and many linguists carry out field work with native speakers to document endangered languages. Automatic speech recognition (ASR) has the potential to serve as a useful tool in these preservation and documentation efforts, but building models for these languages presents numerous challenges.

One particular challenge is a dearth of data, specifically, transcribed and labeled audio data to train the acoustic model and large amounts of text to train the language model. Languages that lack extensive data are known as *under-resourced* or *low-resource languages*, and all but a handful of the world's languages fall into this category. In fact, the set of languages considered to be low-resource for the purposes of ASR research includes many very widely spoken languages, including Bengali (spoken natively by 200 million people) and Vietnamese (with more native speakers than French) (Harper, 2014). Although researchers do not currently have access to large amounts of labeled data in these languages, it would be relatively easy to acquire more data with small investments of time and money to train speakers of the language to collect and label data.

Languages like Vietnamese and Bengali, which for political and economic reasons happen to have few ASR resources, stand in contrast to what we will call *acutely under-resourced languages*. Acutely under-resourced languages are typically spoken by very few people, are rarely written down, and may even lack a standardized writing system. Speakers of these languages, who sometimes live in remote parts of the world, might be reluctant to share their knowledge with outsiders or even to acknowledge that they speak the language. Linguists and community members routinely work to preserve and document these languages, often with the financial support of the government or non-governmental organizations. To date, however, there has been limited research performed in developing ASR systems for these languages, despite the potential benefit it would provide for language documentation and preservation.

In this paper, we present a case study in developing an ASR system for an acutely under-resourced language by focusing on Seneca, an endangered indigenous language of North America. We first provide an overview of the language and the obstacles to developing a robust ASR system for the language, given not only the poverty of the existing resources but also the unusually complex and productive morphology of the language. We propose approaches for leveraging linguistic knowledge and existing resources to increase the accuracy of our recognizer, as well as a tool for iteratively improving ASR performance and for optimizing the utility of ASR output for stakeholders engaged in transcription of audio data for language documentation purposes. A subset of the data we explore will be made available to other groups interested in developing ASR systems for low-resource polysynthetic languages. Our results demonstrate the potential for applying ASR to streamline and enhance the important task of documenting and preserving acutely under-resourced and endangered languages.

2. Background

2.1. Language documentation

Language documentation is the subfield of linguistics focused on producing a permanent and complete record of a language, which should include not only information about the grammar and lexicon but also labeled and annotated audio and textual data illustrating the information

contained in the grammar and lexicon and exemplifying how the language is used in everyday life (Himmelmann, 2006; Austin, 2014). Although social anthropologists have long collected language data from the communities they study, it is only in the last half-century that theoretical linguists have focused their efforts on systematically documenting languages in this way.

The importance of documentation for the communities of speakers of endangered languages is clear. Documenting an endangered language is one way to preserve important parts of a culture. Efforts by speech communities to maintain and revitalize extinct or endangered languages have also benefited from comprehensive language documentation. Theoretical linguists, while contributing to support for speech communities, are also motivated by the desire to find cross-linguistic evidence of language phenomena that can support or refute theoretical frameworks, with the goal of providing insight into the cognitive underpinnings of language.

The challenge in language documentation is that generating detailed transcriptions of recorded speech data and annotations of those transcriptions requires time, linguistic expertise, and technical knowledge. With this in mind, the primary goal of our work on developing ASR systems for acutely under-resourced languages is to provide an efficient and useful mechanism for helping linguists and community members working on endangered language documentation to produce complete and accurate annotated transcriptions of naturalistic, spontaneous speech data.

2.2. ASR for low-resource languages

The last several years have seen a surge in interest in developing robust ASR systems for low-resource languages (Besacier et al., 2014), fueled in part by the U.S. Intelligence Advanced Research Projects Activity (IARPA) Babel initiative (Harper, 2014). The IARPA Babel datasets consist of roughly 10 hours of transcribed speech for a number of relatively widely-spoken but low-resource languages, including Cantonese, Bengali, Turkish, Zulu, and Haitian Creole. The majority of the recent research on ASR for these languages has focused on optimizing the acoustic model in order to overcome the constraints imposed by having a limited amount of labeled audio training data. Researchers have explored modifications in approaches used to train the acoustic models (Grézl et al., 2014; Miao et al., 2013; Thomas et al., 2013); improvements in the features included in the models (Cui et al., 2014; Gales et al., 2014; Ghahremani et al., 2014; Tüske et al., 2014; Prabhavalkar et al., 2013); and supplementing the acoustic training data with data from other languages (Thomas et al., 2013; Gales et al., 2014; Grézl et al., 2014; Imseng et al., 2014; Tüske et al., 2014).

ASR has the potential to serve as a useful tool in language preservation and documentation efforts. To date, however, there has been little interest in building full ASR systems specifically for endangered language documentation. Much of the recent work specifically on developing ASR for low-resource languages has focused on tasks such as forced alignment of phonemes given manually generated transcriptions (DiCanio et al., 2012; DiCanio et al., 2013; Vetter et al., 2016), keyword spotting or spoken term detection (Prabhavalkar et al., 2013; Rosenberg et al., 2017; Metze et al., 2015), or pure phonetic transcription without word-level or utterance level information (Kong et al., 2016; Liu et al., 2016; Das et al., 2016; Hasegawa-Johnson et al., 2017).

The SPICE project's Rapid Language Acquisition Tool (Schultz et al., 2007), while offering promise as a means to collect data and exploit multilingual resources for building language technology systems for under-served languages, is geared toward languages with numerous speakers and large amounts of digitally available text data. In addition, most of the work stemming from this project has focused on TTS systems, rather than the development of ASR systems (Schultz et al., 2013; Schlippe et al., 2014). A more recent large-scale effort to develop language technologies for low-resource languages is the BULB project (Adda et al., 2016b; Adda et al., 2016a), which shares our goal of documenting endangered languages. The focus of the BULB project, however, is the development of a tablet-based interface for recording and transcribing languages lacking an established writing system. Smaller, language-specific efforts include the work of Mitra et al. (2016), who investigated using ASR for the documentation of Yoloxóchitl Mixtec, an endangered language with relatively abundant labeled audio data (125 hours). Although this work also included the use of hand-corrected ASR output to improve existing acoustic models, the authors did not incorporate synthetic data to improve either the acoustic or language models. We refer the reader to the excellent survey by Besacier et al. (2014) for a more complete discussion of the history of ASR for under-resourced languages.

Our work stands in contrast to previous work on ASR for low-resource languages in several ways. First, unlike Turkish or Cantonese which have millions of native speakers, our language of interest, Seneca, in spoken natively by a handful of people and as a second language by only a few hundred more, many of whom are reluctant to allow their language to be recorded. Second, there is very little written data in Seneca available; we cannot simply crawl the web to collect additional training data for our language model in the way that researchers working on any of the IARPA Babel languages can. Third, very little previous effort has been directed at enhancing the language models, particularly on leveraging existing data and linguistic knowledge to produce synthetic text data to augment the language model training data. Finally, and perhaps most crucially, the objective of our work is not to develop a framework for quickly developing an ASR system for any arbitrary language with unknown linguistic properties; instead, our goal is to provide linguists and endangered language community members with data and tools for documenting a language whose linguistic properties are known by the stakeholders.

Although we will in future work investigate many of the methods for improving the acoustic model described in the literature on low-resource languages, the focus of the work presented here is on leveraging in-domain data and linguistic knowledge to improve the language model and to reduce the OOV rate, which is unusually high due to the extreme morphological complexity of the language. Using the output of our ASR system, we then generate data files that can be used by stakeholders to transcribe and annotate new audio data. This newly transcribed and annotated data can be used for documentation purposes and can be reincorporated into the ASR system as additional training data to improve the existing models.

3. Data

3.1. The Seneca language

Historically spoken primarily in the areas of North America now known as New York, Ontario, and Quebec, the Iroquoian language family includes Seneca, Cayuga, Onondaga, Oneida, Tuscarora, Mohawk, and Cherokee. All except Mohawk and Cherokee are considered severely endangered, and all are acutely under-resourced. Seneca, the language we will discuss, is spoken across three reservations in Western New York: the Cattaraugus, Allegany and Tonawanda Reservations. There are currently fewer than 50 native speakers of Seneca, most of them elderly, and a few hundred second-language speakers.

Iroquoian languages have polysynthetic morphological systems, in which words are composed of many morphemes. Unlike agglutinative languages such as Turkish or Hungarian, which are also highly inflected, polysynthetic languages often permit noun incorporation, a process by which fully inflected nouns can optionally be inserted between a verb and the morphemes that accompany that verb. As shown in Figure 3.1., the basic Iroquoian verb is made up of four morphemes: the prepronominal prefix indicating tense, the pronominal prefix indicating the subject, the verb root, and the aspect suffix. Every Iroquoian verb must have at least a pronominal prefix, verb root, and aspect suffix. Seneca has fifty-two possible pronominal prefixes (varying by person, number, gender, and other features), thirty prepronominal prefixes (including verb tense, case, and other grammatical features), and four aspect suffixes (including ongoing action, completed action, habitual action). Thus, for a given verb there can be as many as 4680 different forms – not including any potentially incorporated nouns – which stands in stark contrast to a morphologically poor language such as English, in which a regular verb can have up to only five possible forms. This very high degree of morphological complexity yields severe vocabulary sparsity problems.

3.2. ASR Training and Testing Data

The audio recordings used to the train the Seneca acoustic model consist of roughly 80 minutes of spontaneous, naturalistic speech produced by five adult speakers, three male and two female. All five are first-language Seneca speakers whose second language is English, and all five are middle-aged or elderly. Additional information about the acoustic training data is provided in Table 3.2.. Recordings were made over many years under a variety of conditions using various pieces of recording equipment, yielding a diverse set of audio data.

prepron. prefix	**pronom. prefix**	**verb root**	**aspect suffix**
ʌ	g	ɑdʌnod	ʌʔ
future	1sg	sing	punctual

'I will sing'

Figure 1: Morphological structure of the Seneca verb. [ʌgʌdeɪrɑnoʊdʌʔ], meaning *I will sing*.

Speaker A is from the Cattaraugus Seneca reservation located 30 miles south of Buffalo, NY. In his brief recording, he tells the story of his great-grandfather, who used to hunt bears without a gun. Speaker B is from the Allegany Seneca reservation located by Salamanca, NY. His brief recording is a description of his garden and the plants he usually includes in his garden each year. This data was recorded and transcribed by Wallace Chafe, an emeritus professor of linguistics at UC Santa Barbara.

Speaker C is also from the Cattaraugus Seneca reservation. In his brief recording, he discusses the habits of deer. Speaker D is from the Cold Spring portion of the Allegany Seneca reservation. Her audio data consists of 30 minutes of conversations in Seneca with the first author, a member of the Seneca nation who is a second-language speaker of Seneca. The topics in this recording are wide ranging and include the speaker's family and upbringing, various stories from her childhood, and current events. Speaker E is from the Allegany Seneca reservation. This data totals 41 minutes of conversations in which the speaker discusses with other Seneca speakers a wide range of topics, including personal narratives and Seneca culture and folklore. This data was recorded and transcribed by the first author.

In addition to the transcriptions of the audio data described above, we have access to two other sources of textual data for training the language model. The first is a collection of transcribed stories and narratives produced by a Seneca speaker from the Allegany Seneca reservation. The second source is the Seneca Topic Reference Guide, a pedagogically oriented resource created by various Seneca speakers from across both the Cattaraugus and Allegany reservations. The utterances in this document were designed to enable a learner to have a simple conversation with another speaker in a question-and-answer format.

The held-out audio data used to test the ASR system was produced by Speaker E and was 12 minutes in length, with 40 utterances and 672 words.

4. Methods

We use the Kaldi (Povey et al., 2011) toolkit to build and test our ASR models. The acoustic model was created

	Minutes	Words	Sentences
Speaker A	3	139	20
Speaker B	2	126	21
Speaker C	4	265	20
Speaker D	75	4375	474
Speaker E	60	6059	400
Total	**144**	**10964**	**1235**

Table 1: Breakdown of acoustic training data by speaker

	Sentences	Words	Types
Stories	572	3925	817
Topic Ref.	221	573	219
Total	**793**	**4498**	**1036**

Table 2: Additional language model data.

following the "Kaldi for Dummies" tutorial recipe, which uses the standard 13 dimensional cepstral mean-variance normalized MFCCs, plus their first and second derivatives, within a GMM framework. The recipe was extended to apply LDA transformation and Maximum Likelihood Linear Transform to the features. Other training techniques included boosted Maximum Mutual Information (bMMI) and Minimum Phone Error (MPE). Both bMMI and MPE were trained over 4 iterations and bMMI used a boost weight of 0.5.

As discussed above, our focus is on leveraging existing resources to improve the language model and to reduce what we expect to be a very high OOV rate, given the morphological complexity of the language. We will compare three ASR systems, each with a different lexicon and language model.

The baseline model was created using the transcriptions of the audio data used to train the acoustic model. In addition, a list of 1,992 words extracted from a Seneca-English dictionary (Chafe, 1967) and combined with other words from the transcriptions of the acoustic training data, resulting in a lexicon of 2156 words. The second language model was built using the data described above plus data described in Table 3.2., adding an additional 739 sentences and 4498 words to the training data for the language model and 329 new words to the lexicon.

The third model was built using all of the above data plus additional synthetic data created using a deterministic algorithm for generating a morphologically rich set of Seneca verb forms from verb roots given the phonological processes that apply across morpheme boundaries (Chafe, 2015). The most frequently occuring verb roots in the data used to train models 1 and 2, above, were identified. Each verb root was then processed by the algorithm to generate multiple other common but unseen forms of that verb. In all, about 5000 verb forms, synthetically generated in this way, were added to the lexicon. An overview of all three models is shown in Table 4.

	Sentences in Corpus	Words in Lexicon
LM 1	778	2156
LM 2	1571	2485
LM 3	1571	7549

Table 3: Number of utterances and words in each of the three language models.

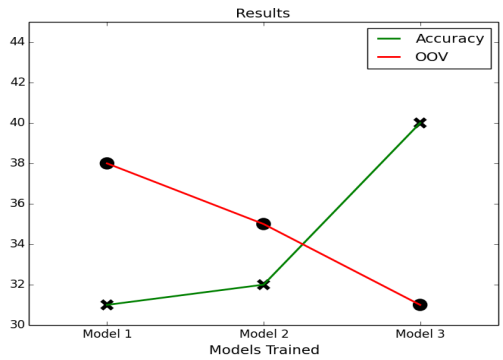

Figure 2: OOV rates and accuracy of three ASR models.

5. Results

Figure 2 plots the ASR accuracy against the OOV rate for the three models. As expected, given Seneca's complex morphology and the small amount of available language model training data, the OOV rate for all three models is very high but decreases with each addition of data to the language model, from 38% to 35% to 31%. The largest reduction in the OOV rate came from the introduction of the synthetic data, which included only the most frequent verbs in the original text data. We anticipate further reductions in the number of OOVs with more extensive use of the algorithm to generate more possible verb forms.

Despite the large number of OOVs, the recognizer performs adequately given the small amount of training data, with WER decreasing from 69% to 68% to 65% with each addition of data to the language model. The accuracy of our systems compare favorably with that reported by research groups working on low-resource languages with much simpler morphology (e.g., 60-70% WER on four of the IARPA Babel languages in Cui et al. (2014)) or in artificially low-resource scenarios (e.g., 60% WER in Thomas et al. (2013)).

Recall that one goal of our work is to provide a tool that linguists and endangered language speakers can use to more efficiently transcribe and annotate recorded language data. To that end, we have created a tool that works in conjunction with Kaldi to speed the process of labeling new training data. Using Kaldi's online wav decoder (online-wav-gmm-decode-faster), an unlabeled .wav file can be decoded using one of the trained Seneca models. The decoder produces a text file with the beginning and end timestamps of the spoken Seneca utterances. We convert this file using custom tools to a TextGrid file containing the aligned Seneca utterances. Using Praat (Boersma, 2001), a linguist or speaker

of the language can, with relatively little training, quickly review words and utterances, listen to the associated audio, easily correct the transcription produced by the ASR system, and adjust the boundaries between words and utterances. The corrected transcripts and annotations can then be saved out to simple text files for use by other linguists and community members. In addition, as more audio data is collected, the audio along with corrected ASR transcripts and timestamps can be incorporated into the acoustic model, resulting in improved ASR performance.

6. Conclusions and future work

In this paper, we used the Iroquoian language, Seneca, as a case study for exploring how to develop an ASR system for an acutely under-resourced and endangered language, with the goal of creating a tool for facilitating language documentation and preservation. Our methods, which included generating synthetic linguistically-informed data in order to lower the OOV rate and improve the language model, demonstrate the feasibility of this project. A subset of the data will be made available to other researchers interested in developing robust ASR systems for under-resourced highly inflected languages.

Our future work will concentrate on exploring two avenues to further reduce the word error rate of our recognizer. We will first apply methods similar to those described in the literature to build more robust acoustic models using DNNs. We are particularly interested in adapting our acoustic model training to include data from Mohawk and Oneida, two Iroquoian languages with very similar phonetic inventories but much more substantial audio resources. In addition, we plan to continue our research using automated morphological parsing tools, such as Morfessor (Smit et al., 2014), to reduce the OOV rate in our data. Our preliminary work using these tools has been disappointing, with very low morphological parsing accuracy, but we anticipate that training the supervised version of the parser with sufficient synthetic verb forms will result in meaningful accuracy improvements.

With the continued rise of globalization and the corresponding decreasing isolation of many indigenous communities, the need to document endangered languages grows more urgent. Automatic speech recognition and other computational linguistic technologies have the potential to transform the way linguists and community members preserve and revitalize their languages, and in turn, the culture they encompass and the insight into human cognition that they provide.

7. References

Adda, G., Adda-Decker, M., Ambouroue, O., Besacier, L., Blachon, D., Bonneau-Maynard, H. E., Gauthier, E., Godard, P., Hamlaoui, F., Idiatov, D., Kouaratab, G.-N., Lamela, L., Makassoe, E.-M., Riallandb, A., Stukerf, S., Van de Veldec, M., Yvona, F., and Zerbiang, S. (2016a). Innovative technologies for under-resourced language documentation: The bulb project. In *Workshop CCURL 2016-Collaboration and Computing for Under-Resourced Languages-LREC*.

Adda, G., StÃ$\frac{1}{4}$ker, S., Adda-Decker, M., Ambouroue, O., Besacier, L., Blachon, D., Bonneau-Maynard, H., Godard, P., Hamlaoui, F., Idiatov, D., Kouarata, G.-N., Lamel, L., Makasso, E.-M., Rialland, A., de Velde, M. V., Yvon, F., and Zerbian, S. (2016b). Breaking the unwritten language barrier: The BULB project. In *Proceedings of the SLTU-2016 5th Workshop on Spoken Language Technologies for Under-resourced languages*, pages 8–14.

Austin, P. K. (2014). Language documentation in the 21st century. *JournaLIPP*, 3:57–71.

Besacier, L., Barnard, E., Karpov, A., and Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85–100.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, 5(9-10):341–45.

Chafe, W. L. (1967). *Seneca morphology and dictionary*, volume 4. Smithsonian Press.

Chafe, W. (2015). *A Grammar of the Seneca Language*. Univ of California Press.

Crystal, D. (2000). *Language death*. Ernst Klett Sprachen.

Cui, X., Kingsbury, B., Cui, J., Ramabhadran, B., Rosenberg, A., Rasooli, M. S., Rambow, O., Habash, N., and Goel, V. (2014). Improving deep neural network acoustic modeling for audio corpus indexing under the IARPA Babel program. In *Fifteenth Annual Conference of the International Speech Communication Association*.

Das, A., Jyothi, P., and Hasegawa-Johnson, M. (2016). Automatic speech recognition using probabilistic transcriptions in swahili, amharic, and dinka. In *INTERSPEECH*, pages 3524–3528.

DiCanio, C., Nam, H., Whalen, D. H., Bunnell, H. T., Amith, J. D., and Castillo Garcia, R. (2012). Assessing agreement level between forced alignment models with data from endangered language documentation corpora. In *INTERSPEECH*.

DiCanio, C., Nam, H., Whalen, D. H., Timothy Bunnell, H., Amith, J. D., and García, R. C. (2013). Using automatic alignment to analyze endangered language data: Testing the viability of untrained alignment. *The Journal of the Acoustical Society of America*, 134(3):2235–2246.

Gales, M. J., Knill, K. M., Ragni, A., and Rath, S. P. (2014). Speech recognition and keyword spotting for low-resource languages: Babel project research at CUED. In *SLTU*, pages 16–23.

Ghahremani, P., BabaAli, B., Povey, D., Riedhammer, K., Trmal, J., and Khudanpur, S. (2014). A pitch extraction algorithm tuned for automatic speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 2494–2498. IEEE.

Grézl, F., Karafiát, M., and Vesely, K. (2014). Adaptation of multilingual stacked bottle-neck neural network structure for new language. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 7654–7658. IEEE.

Harper, M. (2014). IARPA Babel program. *https://www.iarpa.gov/index.php/research-programs/babel*.

Hasegawa-Johnson, M. A., Jyothi, P., McCloy, D., Mirbagheri, M., di Liberto, G., Das, A., Ekin, B., Liu, C., Manohar, V., Tang, H., et al. (2017). Asr for under-resourced languages from probabilistic transcription. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 25(1):50–63.

Himmelmann, N. P. (2006). Language documentation: What is it and what is it good for. *Essentials of language documentation*, 178(1).

Imseng, D., Motlicek, P., Bourlard, H., and Garner, P. N. (2014). Using out-of-language data to improve an under-resourced speech recognizer. *Speech communication*, 56:142–151.

Kong, X., Jyothi, P., and Hasegawa-Johnson, M. (2016). Performance improvement of probabilistic transcriptions with language-specific constraints. *Procedia Computer Science*, 81:30–36.

Krauss, M. (1992). The world's languages in crisis. *Language*, 68(1):4–10.

Liu, C., Jyothi, P., Tang, H., Manohar, V., Sloan, R., Kekona, T., Hasegawa-Johnson, M., and Khudanpur, S. (2016). Adapting asr for under-resourced languages using mismatched transcriptions. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 5840–5844. IEEE.

Metze, F., Gandhe, A., Miao, Y., Sheikh, Z., Wang, Y., Xu, D., Zhang, H., Kim, J., Lane, I., Lee, W. K., et al. (2015). Semi-supervised training in low-resource asr and kws. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 4699–4703. IEEE.

Miao, Y., Metze, F., and Rawat, S. (2013). Deep maxout networks for low-resource speech recognition. In *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, pages 398–403. IEEE.

Mitra, V., Kathol, A., Amith, J. D., and García, R. C. (2016). Automatic speech transcription for low-resource languages-the case of yoloxóchitl mixtec (mexico). In *INTERSPEECH*, pages 3076–3080.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al. (2011). The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society.

Prabhavalkar, R., Livescu, K., Fosler-Lussier, E., and Keshet, J. (2013). Discriminative articulatory models for spoken term detection in low-resource conversational settings. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8287–8291. IEEE.

Rosenberg, A., Audhkhasi, K., Sethy, A., Ramabhadran, B., and Picheny, M. (2017). End-to-end speech recognition and keyword search on low-resource languages. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 5280–5284. IEEE.

Schlippe, T., Ochs, S., and Schultz, T. (2014). Web-based tools and methods for rapid pronunciation dictionary creation. *Speech Communication*, 56:101–118.

Schultz, T., Black, A. W., Badaskar, S., Hornyak, M., and Kominek, J. (2007). Spice: Web-based tools for rapid language adaptation in speech processing systems. In *Eighth Annual Conference of the International Speech Communication Association*.

Schultz, T., Vu, N. T., and Schlippe, T. (2013). Globalphone: A multilingual text & speech database in 20 languages. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8126–8130. IEEE.

Smit, P., Virpioja, S., Grönroos, S.-A., Kurimo, M., et al. (2014). Morfessor 2.0: Toolkit for statistical morphological segmentation. In *The 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Gothenburg, Sweden, April 26-30, 2014*. Aalto University.

Thomas, S., Seltzer, M. L., Church, K., and Hermansky, H. (2013). Deep neural network features and semi-supervised training for low resource speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 6704–6708. IEEE.

Tüske, Z., Golik, P., Nolden, D., Schlüter, R., and Ney, H. (2014). Data augmentation, feature combination, and multilingual neural networks to improve asr and kws performance for low-resource languages. In *Fifteenth Annual Conference of the International Speech Communication Association*.

Vetter, M., Müller, M., Hamlaoui, F., Neubig, G., Nakamura, S., Stüker, S., and Waibel, A. (2016). Unsupervised phoneme segmentation of previously unseen languages. In *INTERSPEECH*, pages 3544–3548.