1st Workshop on Speech for Social Good (S4SG)
24-25 September 2022, Incheon, Korea

Towards an Automatic Speech Recognizer for the Choctaw language

Jacqueline Brixey[1], *David Traum*[1]

[1]USC Institute for Creative Technologies
Los Angeles, CA, USA

brixey@usc.edu, traum@ict.usc.edu

Abstract

In this work we describe ongoing development of the first automatic speech recognition (ASR) system for the American indigenous language, Choctaw (ISO 639-2: cho, endonym: Chahta). Choctaw is spoken by the Choctaw people, with an estimated 10,000 fluent speakers across three federally recognized Choctaw tribes. The Choctaw language is subject-object-verb order, and is highly inflectional, with prefixes, suffixes, and infixes possible on a single verb base. The language also has rhythmic lengthening, in which certain vowels are lengthened based on vowels in affixes. The motivation for developing an ASR system include: assisting in efforts to revitalize and reclaim the endangered language by aiding language learners; promoting additional contexts and scenarios for increased language use, such as conversations with automated dialogue systems; and supporting language documentation. We describe our collection of two-party conversational data and repetition of prepared phrases from a diverse set of speakers that was used to train the system. The ASR model was implemented using Kaldi. The model is currently trained and tested on a subset of the collected data, and achieves a WER of 49.35%.

Index Terms: speech recognition, American indigenous languages, low-resource languages

1. Introduction

In this work, we describe ongoing work towards development of an automatic speech recognition (ASR) system for the Choctaw language (ISO 639-2: cho, endonym: Chahta). Choctaw is an indigenous language spoken by the North American tribe of the same name. It is a low resource, endangered language[1].

The Choctaw people are working to revitalize and reclaim the language (described in more detail in Section 2). Developing an ASR system would assist in those efforts by aiding language learners as an ASR system can be a tool to practice and improve pronunciation. An ASR system would also promote additional contexts and scenarios for increased language use, such as in conversations with automated dialogue systems. Finally, an ASR system can support documentation efforts by improving transcription workflows [1]. Additionally, this work has personal significance for the first author, who is an enrolled member of the Choctaw Nation of Oklahoma.

2. Choctaw people and language

The Choctaw language is spoken by the Choctaws, an indigenous tribe that originally inhabited the southeastern United States in an area that covers the states Mississippi, Alabama, and Louisiana. In the Treaty of Dancing Rabbit Creek in 1830, the first of many removal treaties enacted by the US government's Indian Removal Act, Choctaws ceded their homelands in exchange for sovereign land in Oklahoma. Today, there are three federally recognized Choctaw tribes: the Jena Band of Choctaw Indians (based in Louisiana), the Choctaw Nation of Oklahoma, and the Mississippi Band of Choctaw Indians.

Consonants											
p	b	t	k	f	s	h	m	n	l	w	y [j]
[tʃ]	ch, č				[ʃ]	sh, š			[ɬ]	hl, lh, ł	
Vowels											
[a]	a, ʋ, v, ä				[i]	i			[o]	o, u	
[aː]	a, á, aa				[iː]	e, í, i, ii, ie			[oː]	o, ó, oo	
[ã]	a̲, an, am, ą				[ĩ]	i̲, in, im, į			[õ]	o̲, u̲, on, om, ǫ	

Figure 1: *Choctaw sounds and orthographic variants (from [7])*

Choctaws are the third most populous US tribal group, with approximately 195,000 people identifying as Choctaw in the 2010 US census[2]. In a 2015 census, there were approximately 9,600 fluent speakers of the language across the larger tribal group, and is designated as endangered according to Ethnologue[2]. However, the number of fluent speakers continues to decline, and the COVID pandemic has greatly impacted the Choctaw population in Oklahoma. As of January 2021, it is estimated that there are fewer than 1,000 fluent speakers in the Chocctaw Nation of Oklahoma [3]. Tribal revitalization efforts include language courses at local schools, online classes, and weekly community classes.

The Choctaw language is a part of the Muskogean language family [4]. The language is subject-object-verb order. Choctaw has a complex morphology. Affixes on the verb inflect for tense and argument agreement [5]. The language has fifteen consonants and nine vowels, consisting of short, long, and nasalized versions. A unique feature of the language is rhythmic lengthening, in which word base vowels can become lengthened due to vowels present in affixes [6].

The relevant literature [8, 6, 9] indicates that there were historically at least three dialect variants in Mississippi. It is unclear if those three dialects were carried to Oklahoma following the forced relocation in the 1830s. However, one large source of difference is in orthography. Today multiple orthographic conventions are used, summarized in Figure 1(see [7] for more). In this work, we collected audio from speakers in Oklahoma and used the "traditional" orthography, which is used by the Choctaw Nation of Oklahoma.

[1] https://www.ethnologue.com/size-and-vitality/cho

[2] https://www.census.gov/population/www/cen2010/cph-t/t-6tables/TABLE%20(1).pdf

10.21437/S4SG.2022-2

3. Related work

Many world languages face the threat of becoming extinct, and with each language that dies, we lose unique culture, heritage, and history. However, there is currently a greater emphasis on developing technological resources for low resource and indigenous languages.

There has been limited work on developing ASR systems for American indigenous languages. One such work developed an ASR system for Cherokee [10], and developed it using Huggingface transformers [11]. A second work developed a speech recognizer for Mixtec [1] by utilizing ESPNet [12], an open-source platform for developing end-to-end ASR systems. The Cherokee ASR achieved a WER of 64%, while the Mixtec system had a best WER between 13 and 18%.

Other ASR systems implemented for specific endangered languages include Ainu [13], Samoyedic languages [14], Māori [15], and Komi [16].

4. Audio resources

In preparation for developing the ASR system, we reviewed existing audio materials in a multimodal corpus (Section 4.1), as well as collected novel data (Section 4.2).

4.1. Existing data

The Choctaw language, as noted in Section 2, has been documented over the years. A substantial portion of the documentation is text, since many of the older records predate audio and video recordings.

Our previous work included creating the corpus ChoCo [17], which is the only existing corpus of Choctaw data, formed from gathering historical records and novel materials. The multimodal corpus covers a range of topics and scenarios, and has a mix of text, video, and audio data. ChoCo contains examples of the Oklahoma Choctaw and Mississippi Choctaw variants of the language. We plan to use data from ChoCo to further train and test the ASR system in future work.

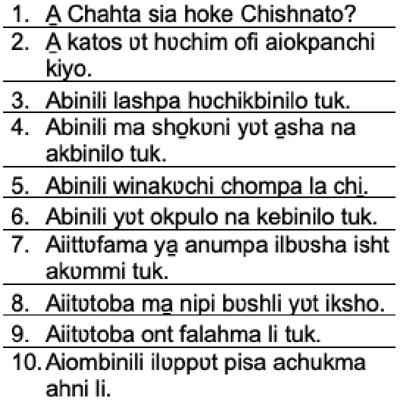
4.2. Novel data collection

Two novel sources of data were collected to train the ASR system. We designed two scenarios to record audio data described in detail below. To record Choctaw speakers in Oklahoma, we submitted an IRB protocol to our institutional IRB, as well as to the Choctaw Nation's IRB.

4.2.1. Repeated phrases

We recorded 48-speakers ranging in skill from beginner to fluent repeating aloud 200 prepared phrases. The phrases were selected to represent a diversity of sounds and words in Choctaw, and include 478 unique words. A sample of the phrases is shown in Figure 2. All phrases and related audio were drawn from the School of Choctaw Language's Lessons of the Day[3]. In the recording session, participants were shown each of the phrases one at a time, and could listen to a clip of the phrase being said. Participants were encouraged to say the phrase as they normally would, which could include different intonation, contractions, or dropping syllables.

Individuals who are learning the language were also recorded in order to train a version of the speech recognizer in future work that will recognize nonstandard Choctaw sounds common in language learner pronunciations. Such a model would be useful for learners hoping to practice the language in conversation where the goal is to be understood even if making some pronunciation errors.



1. A̱ Chahta sia hoke Chishnato?
2. A̱ katos ʋt hʋchim ofi aiokpanchi kiyo.
3. Abinili lashpa hʋchikbinilo tuk.
4. Abinili ma sho̱kʋni yʋt a̱sha na akbinilo tuk.
5. Abinili winakʋchi chompa la chi̱.
6. Abinili yʋt okpulo na kebinilo tuk.
7. Aiittʋfama ya̱ anumpa ilbʋsha isht akʋmmi tuk.
8. Aiitʋtoba ma̱ nipi bʋshli yʋt iksho.
9. Aiitʋtoba ont falahma li tuk.
10. Aiombinili ilʋppʋt pisa achukma ahni li.

Figure 2: *The first ten lines to be repeated by participants*

The total recorded audio from this collection is roughly 35 hours. Since the audio recorder was left running for the entirety of each participant's recording session, some parts of the recordings are not intended to be included for training the model, such as participant repeats due to mispronunciations, or interruptions during the recording session. As a result, clipping the audio is required, and is an ongoing process.

4.2.2. Conversations

Code-switching, or the act of switching between two languages, is very common in spoken Choctaw. One reason is that many Choctaw speakers are English bilinguals. Additionally, some words do not have equivalents in the other language, such as for cultural items or vocabulary related to technology. As a result, speakers will "borrow" the word from the other language when used in conversation.

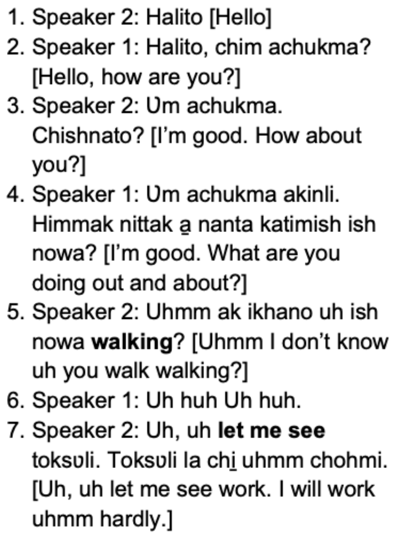
With these facts in mind, an ASR system that would best meet the way Choctaw is realistically spoken should be capable of recognizing code-switched utterances. In preparation for training an ASR system to recognize code-switched utterances, we designed a scenario in which code-switching might occur. In the scenario, a more fluent speaker and a less fluent speaker were paired to have a 15-minute long conversations in Choctaw. All participants were Choctaw-English bilinguals. Participants were instructed to talk on any topic of their choosing. Participants were not allowed to reference any outside materials to look up words, rather if they were unsure about how to say an item in Choctaw, they were instructed to ask their conversation partner or say it in English. Eight conversations in total were recorded, with fifteen unique participants (more fluent participants could participate twice, less fluent participants were restricted to only participating once). A portion of a conversation is shown in Figure 3, an example of code switching is in bold in the fifth line where the speaker switches to English to confirm the meaning of a previously said word.

The data collected is first invaluable for training the ASR system. Additionally, it provides insights into where and when a code-switch might occur. All conversations have been transcribed and have an English translation. Code-switching occurs at least once in all conversations. The less fluent speakers

[3] https://choctawschool.com/home-side-menu/lesson-of-the-day-subscribe-here.aspx

	Gold	Monophone model	Triphone model
1	Keyu sʋ nushkobo hʋt hottupa kiyo	Kiyo sʋ nushkoboka hottupa kiyo	Keyu sʋ nushkobo ʋt hottupa kiyo
2	Aka̱k nipi ilʋppʋt kalampi moma	Oka nipi ilʋppʋt kalampi moma	Aka̱k nipi ilʋppʋt kalampi moma
3	Chim ʋlla yʋt chim ofi aiokpanchi	Chim ʋlla yʋt chim ofi aiokpanchi	Chibbak ʋlla yʋt kiyo ofi aiokpanchi
4	Pi attoba chi̱	Ittonla chi̱	Kil
5	ʋllo̱si mʋt hochʋffo hʋt yaiya	Anusi mʋt hohchʋffo hʋpia yʋt	Anusi ilap hohchʋffo a yaiya
6	Holisso apisa cha ish antta hʋt shohbi tuk o̱	Holisso katos kucha ish ʋpʋt nukshobli tuk	Holisso apesʋchi la chahta shohbi tuk o̱

Table 1: *Examples of errors from the speech recognizer for both the monophone model (WER 49.35%) and triphone model (WER 61.64%). The gold (or expected) utterances are in the first column.*



1. Speaker 2: Halito [Hello]
2. Speaker 1: Halito, chim achukma? [Hello, how are you?]
3. Speaker 2: Ʋm achukma. Chishnato? [I'm good. How about you?]
4. Speaker 1: Ʋm achukma akinli. Himmak nittak a̱ nanta katimish ish nowa? [I'm good. What are you doing out and about?]
5. Speaker 2: Uhmm ak ikhano uh ish nowa **walking**? [Uhmm I don't know uh you walk walking?]
6. Speaker 1: Uh huh Uh huh.
7. Speaker 2: Uh, uh **let me see** toksʋli. Toksʋli la chi̱ uhmm chohmi. [Uh, uh let me see work. I will work uhmm hardly.]

Figure 3: *Transcription of one conversation recording between a fluent and student Choctaw speakers. Code-switching is shown in bold font and translations are in brackets.*

tended to code-switch more in the conversations, but in several conversations the more fluent speaker also code-switched.

5. Development

5.1. Kaldi

We implemented our ASR using Kaldi [18], an open-source toolkit for ASR development that is based on finite-state transducers. Kaldi provides a number of recipes for developing ASR models from scratch; in this work we followed the WSJ recipe[4]. In this recipe, a monophone model and triphone model are created. The ASR system is currently implemented with minimal fine tuning.

5.2. Lexicon

The lexicon is the pronunciation dictionary file in which all words are represented as phones. For our lexicon, we used a dictionary produced by the Choctaw Nation of Oklahoma [19]. In our lexicon, all words were standardized to match orthography found in the dictionary entries. Pronunciations and phones were derived from the same dictionary.

The lexicon currently contains 2,727 entries, which is a limitation and an important lesson to share with other language communities that may seek to develop ASR systems for their languages. It is difficult to develop the lexicon without expert knowledge. Other Choctaw dictionaries in Choctaw (such as [20]) contain more lexical entries, but do not include pronunciation. We expect our system to encounter many words not listed in the current lexicon as conversational data is added to the model, such as the many possible inflected forms of verbs. It will be necessary to consult with fluent speakers to add entries to our lexicon in order to develop a more robust ASR. For languages with few or no fluent speakers, developing a large lexicon could be a severe challenge. That said, a lexicon is a valuable and important contribution to language documentation.

5.3. Data subset

At the time of this submission, the current ASR model was trained and tested using repeated phrase audio from six of the twelve total fluent speakers, four for training and two for testing. The total training data used in the current model was roughly 45 minutes in duration, and the testing data was roughly 12 minutes. This is ongoing work, and we will add data to the ASR model once the it has been processed.

Although participants were given phrases to repeat, variations did occur. For example, one speaker said the shorter form *skʋlli* ("money") rather than the full form of the word, *iskʋlli*. Other changes occurred through participant errors, for example, one speaker said, *Aiitʋoba mʋt nipi bʋshli iksho tuk*, rather than the intended phrase, *Aiitʋoba ma̱ nipi bʋshli yʋt iksho* ("The store does not have a butcher.").

The language model built using the Kaldi recipe is an FST; the language model for the preliminary results described here only include the words that are in the selected phrases (see Section 4.2.1). There are 508 unique words in the phrases of the described subset of data.

5.4. Preliminary Results

We used Kaldi tools to train a monophone model and a triphone model. Our current system achieves a best word error rate (WER) of 49.35% for the monophone model, and a best WER of 61.64% on the triphone model.

Examples of errors are shown in Table 1. In the first example, both the monophone model and triphone model did not recognize the marker *hʋt*. In the second example, the monophone model made one error, producing *oka*("water") rather than *aka̱k*("chicken"). The triphone model made one error in example three with the first word by producing *chibbak*("your hands") rather than the possessive pronoun *chim*, and replaced the possessive pronoun later in the sentence with the negation marker *kiyo*. Both models made large errors in the fourth example, while both models were able to recognize at least *hochʋffo*("hungry") in the fifth example. In the final example in the table, both models made several errors but did recognize

[4] http://kaldi-asr.org/doc/kaldi_for_dummies.html

the first word *holisso*("book").

6. Discussion and Future Work

Our results indicate that building a functional ASR using Kaldi with the use case of repeating phrases can be accomplished with roughly one hour of audio data. This use case can be helpful for practicing oral fluency for language learners. It will also be a relevant finding for other indigenous languages that may have limited numbers of fluent speakers with whom to record and train a system but still wish to train an ASR system.

We plan to continue development on the ASR system in several directions. First, we will continue to add fluent speaker audio to the model, as well as our novel collection of conversational data and conversational data from the corpus ChoCo. The total amount of audio available with which to train the model is summarized in Table 2. We aim to analyze the model for its ability to recognize Choctaw's many infixes and rhythmic lengthening. This will form a fluent speaker only ASR, which we aim to share with the Choctaw Nation in support of their language revitalization efforts, and potentially publicly online.

Source	Total duration	Training	Testing
ChoCo audio	13:45	0	0
Dialogue collection	2:07	0	0
Repeated phrases collection	35:25	0:45	0:15

Table 2: *Number of hours and minutes of spoken Choctaw in ChoCo (Section 4.1) and from our novel data collection (Section 4.2.1 and Section4.2.2)*

Next, we will create a learning speaker ASR by adding the learning speaker audio data to the model. We hope to share this also with the Choctaw Nation and potentially publicly.

Additionally, we are planning to work towards building a code-switching ASR. This will present a number of challenges, including developing an appropriate language model, however it will be an important step towards building conversational computer systems that address the bilingualism present in modern day spoken Choctaw.

Finally, the data collected in 4.2.1 and 4.2.2 will be released at the archives held at the Sam Noble Museum once the ASR system development is completed.

7. Acknowledgements

Data collection was funded by a USC Research Enhancement Fellowship. This research was sponsored in part by the Army Research Office and was accomplished under Cooperative Agreement Number W911NF-20-2-0053. We thank the anonymous reviewers for their helpful comments and feedback. We also thank the Choctaw Nation IRB for their helpful comments and feedback. We thank the many participants of the Choctaw Nation for giving their time, wisdom, and expertise. We honor the participants who died from COVID, we aspire to be good stewards of their final recordings.

8. References

[1] J. Shi, J. D. Amith, R. C. García, E. G. Sierra, K. Duh, and S. Watanabe, "Leveraging end-to-end asr for endangered language documentation: An empirical study on Yolóxochitl Mixtec," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2021, pp. 1134–1145.

[2] G. F. Simons and C. D. Fennig, Eds., *Ethnologue: Languages of the World*, twenty-first ed. Dallas, Texas: SIL International, 2018. [Online]. Available: https://www.ethnologue.com/language/cho

[3] M. Rogers, "Choctaw Nation members talk about impact of losing native speakers to COVID-19," *News 12*, Jan 2021. [Online]. Available: https://bit.ly/3MlEzFO

[4] M. R. Haas, "Southeastern languages," in *The Languages of Native America: Historical and Comparative Assessment*, L. Campbell and M. Mithun, Eds. University of Texas Press, 1979, pp. 299–326.

[5] G. A. Broadwell, "Parallel affix blocks in Choctaw," in *Proceedings of the 24th International Conference on Head-Driven Phrase Structure Grammar, University of Kentucky, Lexington*, S. Müller, Ed. Stanford, California: CSLI Publications, 2017, pp. 103–119.

[6] Broadwell, George Aaron, *A Choctaw Reference Grammar*. U of Nebraska Press, 2006.

[7] J. Brixey, E. Pincus, and R. Artstein, "Chahta anumpa: A multimodal corpus of the Choctaw language," in *Proceedings of LREC 2018*, Miyazaki, Japan, 2018.

[8] G. A. Broadwell, "Choctaw," in *Native Languages of the Southeastern United States*, H. K. Hardy and J. Scancarelli, Eds. U of Nebraska Press, 2005, pp. 157–199.

[9] T. D. Nicklas, "The elements of Choctaw," Ph.D. dissertation, University of Michigan, 1972.

[10] S. Zhang, B. Frey, and M. Bansal, "How can NLP help revitalize endangered languages? A case study and roadmap for the Cherokee language," in *ACL 2022*, 2022.

[11] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 2020, pp. 38–45.

[12] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen *et al.*, "Espnet: End-to-end speech processing toolkit," *arXiv preprint arXiv:1804.00015*, 2018.

[13] K. Matsuura, S. Ueno, M. Mimura, S. Sakai, and T. Kawahara, "Speech corpus of Ainu folklore and end-to-end speech recognition for Ainu language," *arXiv preprint arXiv:2002.06675*, 2020.

[14] N. Partanen, M. Hämäläinen, and T. Klooster, "Speech recognition for endangered and extinct samoyedic languages," *arXiv preprint arXiv:2012.05331*, 2020.

[15] R. C. Solano, S. A. Nicholas, and S. Wray, "Development of natural language processing tools for cook islands māori," in *Proceedings of the Australasian Language Technology Association Workshop 2018*, 2018, pp. 26–33.

[16] N. Hjortnaes, N. Partanen, M. Rießler, and F. M. Tyers, "Towards a speech recognizer for Komi, an endangered and low-resource Uralic language," in *Proceedings of the Sixth International Workshop on Computational Linguistics of Uralic Languages*, 2020, pp. 31–37.

[17] J. Brixey and R. Artstein, "Choco: a multimodal corpus of the Choctaw language," *Language Resources and Evaluation*, pp. 1–17, 2020.

[18] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[19] The Choctaw Nation of Oklahoma Dictionary Committee, *Chahta Anumpa Tosholi Himona: New Choctaw Dictionary*, 1st ed. Choctaw Print Services, 2016.

[20] C. Byington, *A Dictionary of the Choctaw Language*. US Government Printing Office, 1915, edited by John R. Swanton and Henry S. Halbert. Smithsonian Institution Bureau of American Ethnology Bulletin 46.