Enhancing Documentation of Hupa with Automatic Speech Recognition

Zoey Liu
Boston College
zoey.liu@bc.edu

Justin Spence
University of California, Davis
jspence@ucdavis.edu

Emily Prud'hommeaux
Boston College
prudhome@bc.edu

Abstract

This study investigates applications of automatic speech recognition (ASR) techniques to Hupa, a critically endangered Native American language from the Dene (Athabaskan) language family. Using around 9h12m of spoken data produced by one elder who is a first-language Hupa speaker, we experimented with different evaluation schemes and training settings. On average a fully connected deep neural network reached a word error rate of 35.26%. Our overall results illustrate the utility of ASR for making Hupa language documentation more accessible and usable. In addition, we found that when training acoustic models, using recordings with transcripts that were not carefully verified did not necessarily have a negative effect on model performance. This shows promise for speech corpora of indigenous languages that commonly include transcriptions produced by second-language speakers or linguists who have advanced knowledge in the language of interest.

1 Introduction

The documentation of endangered and other less-studied languages typically involves the creation of high-quality audio and video recordings representing a variety of speech genres, with the long-term goal of generating general-purpose linguistic data that can be used by diverse audiences for different research and applied purposes (Himmelmann, 1998; Riesberg, 2018). With the advent of cheap, highly portable digital recording and storage technologies since the early 2000s, it is not uncommon for fieldwork projects to generate hundreds of hours of multimedia recordings.

While these collections of recordings are becoming increasingly accessible via web-based portals, in the sense that they can be downloaded, locating information of interest within them correctly and efficiently is another matter entirely. Coarse-grained catalog metadata describing the content of the recordings can provide users with some shallow guidance, but the identification of more specific information requires enormous investments of time and effort. Accordingly, it becomes essential to have adequate transcriptions of recordings for users to find the information they are interested in.

Transcribing recordings, however, is also an extremely time-consuming endeavor, leading to what is sometimes called the "transcription bottleneck" (Gupta and Boulianne, 2020; Zahrer et al., 2020; Ćavar et al., 2016; Shi et al., 2021), which refers to the situation where the language data is mostly in the form of (archival) recordings, and transcriptions of the data are not yet available.

Hupa (ISO 639-3 code: hup; Glottolog code: hupa1240), a critically endangered Native American language of northwestern California, provides a case in point. Since the early 2000s, Mrs. Verdena Parker, an elder from the Hoopa Valley Tribe, has generously shared her knowledge of the language with other community members and academic researchers. Recordings produced by and with Mrs. Parker include several hours of monolingual Hupa narratives and other texts, as well as over 800 hours of linguistic interviews that are a mixture of Hupa and English as the elicitation metalanguage. [1]

The sheer quantity of these Hupa recordings makes their transcription challenging, a situation that is exacerbated by other factors. First, the people who are considered first-language speakers of Hupa are older and tend not to be literate in the language. Therefore the pool of potential transcribers is limited to second-language speakers and linguists with advanced research knowledge. Second, while literacy is used as a tool for some pedagogical purposes in the contemporary Hupa community and there is a reasonably well-established practical orthography, many of the classes for learning Hupa

[1] Many of these recordings are now available through the California Language Archive web portal: https://cla.berkeley.edu/.

Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 187 - 192
May 26-27, 2022 ©2022 Association for Computational Linguistics

focus more on developing oral proficiency rather than on literacy skills per se. This means many of the younger people who have become second-language speakers of the language may not feel confident in their ability to produce accurate transcriptions of connected discourse.

In this work, we apply automatic speech recognition (ASR) technology to help address the transcription bottleneck for Hupa. In particular, we hope to develop effective techniques that would lend themselves to transcribing spoken Hupa. At this stage of the research, we are focusing primarily on monolingual narratives and other texts since these have the highest density of linguistic data and thus more value for research and language documentation.

2 Meet the Language Data

2.1 The Hupa Language

Hupa is the ancestral language of the Hoopa Valley Tribe in present-day Humboldt County, California. Since the mid-19th century, Hupa people have endured many hardships in the wake of the violent colonization of the region, including decades of educational policies that were designed to eradicate indigenous languages and other manifestations of traditional culture. As a result of this difficult history, by the mid-20th century most Hupa children grew up primarily speaking English as their first language, and today there are only a handful of elderly people (probably fewer than a dozen) who are considered first-language speakers of Hupa.

Nevertheless, at least since the 1970s, tribal members have been engaged in various kinds of language reclamation efforts (in the sense of Leonard (2011)), and today a number of people have developed a high degree of L2 proficiency in the language. Students at Hoopa Valley High School can take four years of Hupa language as part of their regular curriculum, and a practical orthography for the language developed in the 1980s and 1990s (Golla, 1996) is used in a number of pedagogically-oriented resources. Good descriptions of the linguistic features of Hupa are also obtainable from Golla (1970) and Sapir and Golla (2001) (see also Gordon (1996)), although there remains something of a disconnect between the highly technical descriptive materials produced by professional academics and the needs on the ground of language teachers and learners.

2.2 Audio data and transcriptions

The Hupa audio data in our experiments consists of a subset of audio recordings collected from fieldwork with Mrs. Verdena Parker (Table 1) that started in 2005 and is ongoing today. The majority of the recordings we use feature Mrs. Parker telling stories from different genres, including personal anecdotes from her life, oral-historical accounts of significant events in Hoopa Valley, and traditional stories that explain how the world came to be. Each recording has time-aligned transcriptions in the practical orthography of Golla (1996); the transcripts were produced by a human transcriber using annotation tools such as ELAN (Brugman and Russel, 2004).

Since the audio files had been transcribed gradually over a number of years by several researchers, each transcript was lightly edited and corrected by a linguist (an author of this paper), who has advanced research knowledge of the language. As of now, after removing utterances that are fully in English, the amount of spoken Hupa available for conducting ASR experiments totals 9h12m.

Although all transcriptions were checked in consultation with Mrs. Parker, each one typically goes through several stages of manual checking before being considered complete. As a result, some transcriptions have been subsequently examined more thoroughly than others. Based solely on transcription quality differences, we divided the audio data into two sets: the “verified" data (∼1h35m) vs. the “coarse" data (∼7h37m).

Overall, the transcriptions of the verified data are more accurate than those of the coarse data. That said, the verified transcriptions typically have undergone more orthographic normalization, which includes removing elements (e.g., word-final epenthetic vowels) that are audible in the recordings but are not part of the practical orthography (Golla, 1996). In a small number of instances, the verified transcriptions might have slight deviations from what was actually produced in the corresponding recording if Mrs. Parker felt strongly that she had misspoken. Therefore while the verified transcriptions tend to be more accurate, in some ways they are idealizations that are less faithful to the acoustic substance of their original recordings.

2.3 Digitized texts

In addition to the audio recordings and their transcriptions, we also included digitized texts for our

Data	N of words	N of types
verified transcriptions	9,265	2,024
coarse transcriptions	41,062	5,731
digitized written texts	41,381	8,205

Table 1: Descriptive statistics for the text data of Hupa applied in experiments.

experiments (Section 4); these texts were originally transcribed from dictation from Sapir and Golla (2001) and Goddard (1904) (Table 1).

3 Related Work

While research on ASR for endangered language documentation is still relatively rare, recently there has been growing efforts trying to mitigate this gap (Michaud et al., 2018; Prud'hommeaux et al., 2021). Shi et al. (2021) adopted end-to-end systems for Yoloxóchitl Mixtec, an endangered Mixtecan language. Using encoder-decoder architectures, they achieved the best word error rate (WER) ($\sim$16%) for over 55h of conversational speech from more than twenty speakers. Gupta and Boulianne (2020) applied neural ASR models for Cree, an indigenous language in Canada. Their data consists of 4h30m story retelling or reading from six speakers. Utilizing data from high-resource languages, Zahrer et al. (2020) performed cross-linguistic learning of phoneme recognition for the Muyu language. In a study of ASR for two tonal languages, Yongning Na and Eastern Chatino, Adams et al. (2018) proposed a neural architecture to jointly predict phonemes and tones without needing time-aligned transcripts and pronunciation dictionary.

ASR technologies have also been developed for some Dene languages (Littell et al., 2018), though in a limited way. For instance, speech recognition tools were incorporated into the Rosetta Stone language learning software for Diné Bizaad (Navajo). [2] The Persephone ASR software (Adams et al., 2018) was combined in ELAN (Brugman and Russel, 2004) for Tsuut'ina.

4 Experiments

4.1 Evaluation scheme

In (low-resource) ASR experiments [3], acoustic models are commonly evaluated with data from held-out speaker(s). This evaluation standard, however, is not applicable in our study here since all of the Hupa audio came from one speaker. Thus as alternatives, we designed two separate evaluation schemes for both the verified and the coarse data.

The first one utilized random splits, for which we randomly divided all the recordings into training and test sets at a 4:1 ratio for ten times. For the second scheme, taking into account the fact that the audio recordings were collected from distinct fieldwork dates (17 dates for the verified data and 34 dates for the coarse data), we used recordings from each held-out date as the test set and the rest of the data was employed as the training set. WER and character error rate (CER) were taken as evaluation metrics for model performance.

Note that the results obtained from these two evaluation methods are not directly comparable, given that the amount of training data and that of the test data for the two methods are different. On the other hand, the goal of employing separate evaluation schemes is to acquire more realistic estimates regarding the potential of the ASR systems in the case of Hupa.

4.2 Acoustic training data configuration

With the two evaluation schemes outlined above, we investigated different training settings with the goal of exploring: (1) the differences between the verified and coarse data; a(2) the utility of including all acoustic data, regardless of transcription quality.

In our first four experiments, we focused on the verified data, evaluating ASR performance with random splits then with held-out dates. We then included the coarse data for model training, keeping the test data the same in order to determine whether WER decreases with more training data, even when there is a mismatch in transcription quality between the test data and the training data. In our second set of experiments, we carried out the same model training procedures using the coarse data. Finally, we combined the coarse data and verified data to train and test acoustic models on random splits of this combined data.

4.3 Language and acoustic models

For each training/test set split of the audio data, we built one trigram language model with Witten-Bell discounting using the SRILM toolkit (Stolcke, 2002); the data used to train the language model also included the transcripts of the audio training data along with the digitized texts.

For acoustic modeling, we drew on the open-source Kaldi toolkit (Povey et al., 2011). The au-

[2] https://navajorenaissance.org/

[3] Code in quarantine at https://github.com/zoeyliu18/Hupa

Original utterance:	haya:ł	keh	do’ng	haya:	ch’in’	***	tehł
Model prediction:	haya:ł	***	do’ng	haya:	ch’in’	te:	niwhsing
Evaluation:		D				I	S

The original utterance has six words; compared to the original utterance; the utterance predicted by the ASR model contains one deletion (D), one insertion (I), and one substitution (S); therefore:

WER = 100 * $\frac{1+1+1}{6}$ = 50%

An example of WER calculation; I for insertion, D for deletion, and S for substitution.

Evaluation	Data	Training setting	WER (%)	CER (%)
random	train: 1h16m;	*just verified data*	53.23	24.58
splits	test: 19m	*add coarse data*	36.89	12.20
held-out	train: 1h30m;	*just verified data*	46.10	17.48
dates	test: 5m	*add coarse data*	37.96	13.57

Table 2: ASR evaluation results for the verified data.

Evaluation	Data	Training setting	WER (%)	CER (%)
random	train: 6h6m;	*just coarse data*	45.13	21.37
splits	test: 1h31m	*add verified data*	35.13	12.65
held-out	train: 7h24m;	*just coarse data*	37.70	12.58
dates	test: 13m	*add verified data*	35.60	12.37

Table 3: ASR evaluation results for the coarse data.

Evaluation	Data	WER (%)	CER (%)
random splits	train: 7h22m; test: 1h50m	35.26	12.38

Table 4: ASR evaluation results when combining all verified and coarse data together.

dio recordings were transformed to the standard 13 dimensional mel-frequency cepstral coefficients (MFCCs), as well as their delta- and delta-delta features. The delta- and delta-delta features are, respectively, numerical approximations of the first and second order derivatives of the MFCCs, both computed on a 25ms window with 10ms interval apart which enables modeling the trajectories of the audio signals. Linear Discriminant Analysis and Maximum Likelihood Linear Transform were then employed to reduce the dimensionality of the feature vectors.

The acoustic model architecture that we used is a fully connected deep neural network (DNN) (Miao et al., 2015), which has been demonstrated to have competitive performance when facing data limitation (Morris et al., 2021). The DNN had six hidden layers, each with 1024 hidden units. Sequence training was carried out with the default parameters in Kaldi using state-level minimum Bayes risk criterion and a per-utterance Stochastic Gradient Descent weight update. Decoding was performed with the finite state transducer-based decoder implemented in Kaldi.

5 Results

The average WER results for the verified data given each training setting and evaluation scheme are presented in Table 2. When only using the verified data for ASR training and evaluation, we obtained a WER of 53.23%; on the other hand, we see that combining coarse data with the training data of the verified set resulted in much lower WER values (and lower CER values as well), and accordingly better model performance; this pattern is consistent regardless of whether evaluating acoustic models with random splits or held-out dates. Similar observations hold when developing models for the coarse data with additional help of verified data (Table 3), which also led to lower WER values. These results indicate that including more training data, even when the transcription quality of the training data does not necessarily match that of the test data, is helpful to build better ASR models.

When combining all data from the verified set and the coarse set together, we reached a WER of 35.26% evaluated with random splits, which is comparable to the results of random splits for each data set separately.

6 Discussion & Ongoing Work

Leveraging ASR technologies, we investigate the possibility and effectiveness of automatically transcribing fieldwork recordings for Hupa. Through experimentation with different evaluation schemes and training settings, the acoustic models demonstrate reasonable WER results, showing promise for applying spoken language technology to document Hupa. Interestingly, training ASR models using recordings with transcripts that were not carefully verified did not negatively impact the performance, which bodes well for speech corpora of indigenous languages that include transcriptions produced by second-language speakers or linguists.

In ongoing work, we are extending our efforts in several directions. First, the transcripts of the coarse data are being manually checked periodically to improve transcription and gloss alignment quality. Second, as we are still in the preliminary stage of performing ASR for Hupa, the current study only used the DNN architecture from Kaldi. We plan to explore other more recent neural approaches (Watanabe et al., 2018) that have been found to be effective with limited amount of audio data (Shi et al., 2021); then apply the trained models to recordings that have not yet been transcribed in an iterative fashion to better combine ASR with documentation of Hupa. Even a WER as high as $\sim 35.26\%$ is expected to yield significant savings in the time required to make transcribed texts available.

Third, thus far our acoustic models are decoded with language models at the word level. However, given the complex morphological features of Hupa (Sapir and Golla, 2001), to reduce out-of-vocabulary rate in future experiments, we are working towards combining morphological segmentation or subword unit models Liu et al. (2019) into building ASR systems. Lastly, with better performing acoustic models and more transcriptions, we aim to develop a workflow to adapt these transcribed materials into pedagogically-oriented resources for use by members of the community.

Acknowledgements

We are grateful for the continuous support from the Hupa indigenous community. We would like to especially thank Mrs. Verdena Parker for her generous and valuable input for the documentation work of Hupa throughout the years. In addition, we thank the anonymous reviewers for their helpful feedback. This material is based upon work supported by the National Science Foundation under Grant #2127309 to the Computing Research Association for the CIFellows Project, and Grant #1761562. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation nor the Computing Research Association.

References

Oliver Adams, Trevor Cohn, Graham Neubig, Hilaria Cruz, Steven Bird, and Alexis Michaud. 2018. Evaluation Phonemic Transcription of Low-Resource Tonal Languages for Language Documentation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Hennie Brugman and Albert Russel. 2004. Annotating multi-media/multi-modal resources with ELAN. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Malgorzata Ćavar, Damir Ćavar, and Hilaria Cruz. 2016. Endangered Language Documentation: Bootstrapping a Chatino Speech Corpus, Forced Aligner, ASR. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4004–4011, Portorož, Slovenia. European Language Resources Association (ELRA).

Pliny Earle Goddard. 1904. *Hupa texts*, volume 1. The University Press.

Victor Golla. 1996. Hupa Language Dictionary Second Edition.

Victor Karl Golla. 1970. *Hupa grammar*. Ph.D. thesis, University of California, Berkeley.

Matthew Gordon. 1996. The phonetic structures of Hupa. *UCLA Working Papers in Phonetics*, pages 164–187.

Vishwa Gupta and Gilles Boulianne. 2020. Speech Transcription Challenges for Resource Constrained Indigenous Language Cree. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 362–367, Marseille, France. European Language Resources association.

Nikolaus P Himmelmann. 1998. Documentary and descriptive linguistics. *Linguistics*, pages 161–195.

Wesley Leonard. 2011. Challenging "extinction" through modern Miami language practices. *American Indian Culture and Research Journal*, 35(2):135–160.

Patrick Littell, Anna Kazantseva, Roland Kuhn, Aidan Pine, Antti Arppe, Christopher Cox, and Marie-Odile Junker. 2018. Indigenous language technologies in Canada: Assessment, challenges, and successes. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2620–2632, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Chang Liu, Zhen Zhang, Pengyuan Zhang, and Yonghong Yan. 2019. Character-Aware Sub-Word Level Language Modeling for Uyghur and Turkish ASR. In *The Annual Conference of the International Speech Communication Association (Interspeech)*, pages 3495–3499.

Yajie Miao, Hao Zhang, and Florian Metze. 2015. Speaker adaptive training of deep neural network acoustic models using i-vectors. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(11):1938–1949.

Alexis Michaud, Oliver Adams, Trevor Cohn, Graham Neubig, and Séverine Guillaume. 2018. Integrating automatic transcription into the language documentation workflow: Experiments with Na data and the Persephone toolkit. *Language Documentation & Conservation*, 12:393–429.

Ethan Morris, Robert Jimerson, and Emily Prud'hommeaux. 2021. One size does not fit all in resource-constrained ASR. In *The Annual Conference of the International Speech Communication Association (Interspeech)*, pages 4354–4358.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, CONF. IEEE Signal Processing Society.

Emily Prud'hommeaux, Robbie Jimerson, Richard Hatcher, and Karin Michelson. 2021. Automatic Speech Recognition for Supporting Endangered Language Documentation. *Language Documentation & Conservation*, 15:491–513.

Sonja Riesberg. 2018. Reflections on descriptive and documentary adequacy. In Bradley McDonnell, Andrea L. Berez-Kroeker, and Gary Holton, editors, *SP15: Reflections on Language Documentation 20 Years after Himmelmann 1998*, chapter 15, pages 151–156. University of Hawai'i Press.

Edward Sapir and Victor Golla. 2001. Hupa texts, with notes and lexicon. *The Collected Works of Edward Sapir, ed. by Victor Golla & Sean O'Neill*, 14:19–1011.

Jiatong Shi, Jonathan D. Amith, Rey Castillo García, Esteban Guadalupe Sierra, Kevin Duh, and Shinji Watanabe. 2021. 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*, pages 1134–1145, Online. Association for Computational Linguistics.

Andreas Stolcke. 2002. SRILM-an extensible language modeling toolkit. In *Seventh international conference on spoken language processing*.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. 2018. ESPnet: End-to-End Speech Processing Toolkit. In *Proceedings of Interspeech*, pages 2207–2211.

Alexander Zahrer, Andrej Zgank, and Barbara Schuppler. 2020. Towards Building an Automatic Transcription System for Language Documentation: Experiences from Muyu. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2893–2900, Marseille, France. European Language Resources Association.