Speech Recognition and Synthesis Technologies Applied to Preservation and Revitalization of the Ainu Language

Tatsuya Kawahara and Kohei Matsuura
School of Informatics, Kyoto University, Japan

Abstract

This paper gives an overview of our activities in developing automatic speech recognition (ASR) and text-to-speech (TTS) systems for the preservation and revitalization of the Ainu language, once spoken in the Hokkaido area of Japan, and listed as "severely endangered" of extinction. With a large pretrained model, a high-performing ASR system can be trained even with five hours of speech from a few speakers. It has been used to streamline the transcription and archiving of old recordings. A TTS system is also developed and used for revitalizing the speech of old folktales whose audio is missing. It is also used to provide a reference for speaking practice for new Ainu speakers. Speech technologies are important for endangered languages because their cultures have typically been passed down orally, and our efforts will be useful for passing them on to the future.

1. Introduction

While there are thousands of ethnic groups and languages in the world, the majority of them are minority groups and languages, and many of them are in danger of extinction. According to the UNESCO World Atlas of Languages, eight languages of that kind are listed in Japan. Among them, Ainu is classified as "critically endangered". Ainu people used to live in the northern part of Japan with their own culture and language, but were forced to assimilate into Japanese society after the late 19th century. As a result, there are only a few native speakers who become very old. Their culture and history have been passed down orally for a long time. Songs and lyrics are often added to dances and rituals as well.

Some might argue that they can be inherited by translating into Japanese or English with the language technology or AI. However, translation cannot convey the cultural background. For example, in the Ainu language, bears are called "kamuy", snakes are called "tannekamuy", orcas are called "repunkamuy", and owls are called "kamuy-cikappo"; this suggests that they are regarded as gods or their incarnations ("kamuy" means god in Ainu). Simply saying "I encountered an owl in the forest" does not convey that nuance. Moreover, the rhyming patterns in the lyrics are hard to keep in the translation. Simple use of translation technology may not lead to preservation of the culture, but to overreliance on English.

Given this background, movements aimed at preserving, passing on, and even reviving the endangered language are gaining momentum, involving both the private sector and the government. These efforts begin with recording and archiving oral traditions and include initiatives such as using these languages in museums and public spaces, as well as creating opportunities for younger generations to learn and speak the language. In the case of the Ainu, a large stock of recordings of oral folklore has been made since the 1970s. In 2020, the Japanese government opened the National Ainu Museum and Park, named "Upopoy," to preserve and exhibit the Ainu culture. The Ainu Language Archive is also set up to collect speech data and make it publicly available in a usable form. However, a large portion of the recordings have not been transcribed or aligned with audio because there are only a few experts in the Ainu language capable of this processing.

The authors' group has been engaged in the development of automatic speech recognition for Ainu to (semi-)automate this process, closely collaborating with museum staff and local communities. We also find the potential of speech synthesis technology for the revitalization of the language. Note that the current speech technology covers around 100 major languages, which have sufficient resources and market. The development of speech recognition and synthesis for minor and low-resource languages remains very challenging. This paper gives an overview of our activities, which are actually used (or being tested) for the preservation and revitalization of the Ainu language

2. Potentials of Speech Technologies for Endangered Languages

2.1 Archiving Oral Traditions

There are numerous audio recordings of stories from the past that were recorded while the native speakers were still alive. The Ainu Language Archive is one such example. However, only a portion of this data has been transcribed and aligned with the audio with timestamps. Therefore, speech recognition technology can be useful for performing these processes automatically. While high accuracy is required for speech recognition used in transcription, such high accuracy is not necessary for aligning the text with the audio once the transcription is available. Furthermore, there are still many unprocessed audio sources that have not yet been archived. Since many of these are conducted in an interview format and often contain sections spoken in languages other than the target language (such as Japanese), speaker recognition, language recognition, and resulting segmentation are necessary.

Proceedings of the Ninth Workshop on the Use of Computational Methods in the Study of Endangered Languages (ComputEL-9), pages 10–14

Figure 1: The Ainu Language Archive (https://ainugo.nam.go.jp/)
©National Ainu Museum & The Foundation for Ainu Culture

2.2 Generation of Speech Content

There is a growing need for audio narration in museum exhibition descriptions and educational materials, for which speech synthesis technology can be useful. In cases like the Ainu language, where there are very few native speakers, even experts may not know the correct intonation. This speech synthesis can provide a useful guide in this situation.

2.3 Language Learning Systems

A system similar to those used for learning major foreign languages is envisioned. In addition to vocabulary training, systems capable of pronunciation practice and even simple spoken interactions, such as everyday conversation, are also conceivable. Research has been conducted on Sami conversation (Jokinen 2018) and Maori pronunciation practice (Watson 2017). While this can be achieved through a combination of speech recognition and speech synthesis, it requires handling non-native speakers, such as Japanese people speaking English. Unlike major languages, data from native speakers is extremely scarce, making it difficult to build models for reliable pronunciation assessment.

3. Ainu Language and the Archive

3.1 Ainu Language

The Ainu people are the indigenous inhabitants of Hokkaido, southern Sakhalin, and the Kuril Islands, and their population was estimated at around 20,000 in the mid-19th century. Due to Japan's colonization of Hokkaido and its assimilation policies, the number of native speakers declined sharply, and in 2009, UNESCO designated the language as being in "critically endangered" status.

The Ainu language exhibits agglutinative and polysynthetic characteristics. Although it shares some similarities with Japanese and has borrowed vocabulary from it, it is a linguistically isolated language of unknown origin. The Ainu language is broadly classified into three groups: Hokkaido Ainu, Sakhalin Ainu, and Kuril Ainu, each with further dialect subdivisions. Our focus is primarily on the Hokkaido Saru dialect, for which the largest-scale data is available.

The Ainu language has both open and closed syllables, but a syllable may contain at most one consonant at the beginning or end. In other words, if we denote consonants as C and vowels as V, syllables take the form V, CV, VC, or CVC. The vowels V consist of five sounds {a, i, u, e, o}, and the consonants C consist of {k, s, t, n, h, m, y, r, w, c, p}. The symbol "_" is used to indicate elision, and "=" is used to indicate a personal connection. Words are generally separated by spaces, as in the following example.

hunak wa e=ek

(where) (from) (you) (come)

3.2 The Ainu Language Archive

The recording of Ainu oral traditions has been carried out since around 1970, ranging from individual efforts to municipal initiatives. In particular, the Ainu Museum in Shiraoi Town established the initial "Ainu Language Archive," which was transferred to the National Ainu Museum upon the opening of "Upopoy" in 2020. Its outlook is shown in Figure 1. The collected audio recordings total 670 hours, including Japanese segments; however, as of 2018 (when the authors began their research and development), only a few dozen hours had been made publicly available in the archive, and a large portion was unprocessed.

Ainu oral traditions can be broadly categorized into the following three types:

(1) Uwepekere (folktales, prose narratives): Stories told from a human perspective in prose style
(2) Yukar (heroic epics): Stories of heroes told in a rhythmic style
(3) Kamuy Yukar (divine songs): Stories told from a divine perspective in a rhythmic style with refrain phrases

The speech recognition research described below focuses on Uwepekere.

4. Application of Automatic Speech Recognition (ASR)

4.1 ASR Model Training and Evaluation

We have developed Ainu Automatic Speech Recognition (ASR) models using the dataset offered at the Ainu Language Archive. The training dataset (Matsuura 2020) consists of Uwepekere recordings by four speakers. They are all elderly female speakers of the Saru dialect. Though the total duration of the datasets is 32 hours, one speaker's recording accounts for about 60%. The scarcity and imbalance of speakers are typical problems in endangered languages, which often lead the model to overfit to these speakers; it performs well for them but significantly degrades for unseen speakers.

In the last several years, however, large pretrained models such as XLSR (Babu 2022) and Whisper (Radford 2022) have been developed and widely used. Some have targeted a large number of languages (Pratap 2023), but most of endangered languages such as Ainu are not included. These models are typically trained using a huge amount of multi-lingual datasets, where Japanese data accounts for less than 10%. We also developed a pretrained model (JP-90K) using 90,000 hours of Japanese data collected online, given that Ainu shares a majority of phones with Japanese and that Ainu speakers are also speakers of Japanese. In summary, we compared the following models.

(1) Conformer model (4-layer CNN + 12-layer encoder + 6-layer decoder) trained from scratch using the 32-hour Ainu dataset
(2) XLSR (300M) model finetuned with the Ainu dataset
(3) Whisper (small and large) models finetuned with the Ainu dataset
(4) Our JP-90K model finetuned with the Ainu dataset

A subword vocabulary of 500 tokens is defined by the SentencePiece algorithm (Kudo 2018). We prepared two test sets: one (Eval1) consists of 3-hour recordings of two different speakers of the same Saru dialect, and the other (Eval2) consists of a 12-hour recording of one speaker of a different Shizunai dialect.

The evaluation results in terms of character error rate (CER) are listed in Table 1. The pretrained models perform better, achieving 93% accuracy on unseen speakers in the same dialect, but degrade substantially in an unseen dialect. But the worse performance in Eval2 may be attributed to the noisy recording of the dataset. Our JP-90K model performs comparably to the Whisper models while being much smaller in size. It runs almost in real time on the CPU, while the Whisper large model takes 10 times real-time.

We also evaluated the effect of the training data size on the accuracy in Eval1. CER is plotted by changing the data size in Figure 2. It shows that the performance of the pretrained models almost converges with 5-to-10-hour speech, and our JP-90K model converges most rapidly. The result suggests that we can prepare a reasonable ASR model for a new language given a 5-hour speech by a few speakers. This is an important finding for developers of this kind of system for endangered languages.

Table 1: ASR evaluation results

	#params	Eval1 (CER)	Eval2 (CER)
Conformer(scratch)	29M	11.0	22.2
XLSR 300M	317M	10.4	19.5
Whisper small	201M	7.7	15.9
Whisper large	1570M	6.6	14.8
JP-90K	167M	7.3	15.6

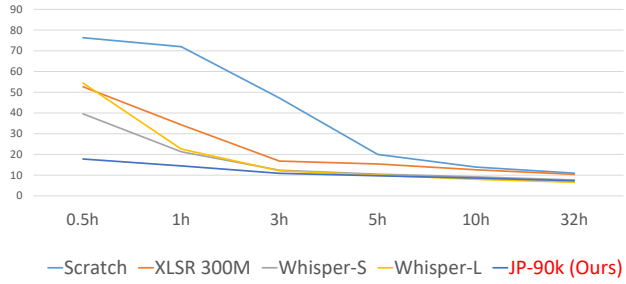

Figure 2: ASR performance (CER) according to training data size (hours)

4.2 Alignment of Speech and Text in the Archive

The first use case of the ASR system is the alignment of speech and its transcript. A large majority of the audio recordings of the Ainu Language Archive had been transcribed by human experts, but the transcripts need to be time-aligned with the audio for displaying, browsing, and searching the archive. Specifically, the text needs to be aligned at least by the phrase unit shown in a line on the right-hand side of Figure 1, and if possible, by the word unit for "karaoke-style" audio playing. This alignment is a tedious task, as a human expert takes one day for a one-hour speech, which had been the major bottleneck in making the archive open to the public.

The alignment can be done by text matching of the ASR output and the ground-truth transcript, and then the time information of the ASR output is copied to the transcript. This process is completely automated,

with ASR accuracy of 80-90%, removing the bottleneck. With this streamlining, all remaining speech data of 670 hours, in addition to the 32 hours used in the ASR model training, have been processed and are ready to be published.

4.3 Transcription of Speech Archive

There are many more audio recordings of Ainu speech collected and stored in the Hokkaido area, which are not transcribed. In 2025, we began a new project with the Ainu National Museum and some Ainu scholars to transcribe these materials using the ASR system. As the ASR output is error-prone, it needs to be proofread and corrected by human experts. To make this process more efficient, we designed and implemented a software editor that enables the human to correct the ASR output by referring to the corresponding speech segment.

Four audio recordings of approximately 12 minutes each were assigned to one person, who engaged in this task. The four workers are staff of the National Ainu Museum and learners of the Ainu language. Everyone completed the task based on the ASR-generated transcripts. Detailed analysis, such as ASR accuracy and post-edit time, is ongoing. The workers told us that the ASR output is very helpful because it is difficult to recognize many speech segments, and that this task is useful for enhancing their language proficiency. The comment is inspiring, as the ASR-assisted editing is useful not only for preserving old recordings but also for producing new, skilled Ainu speakers.

4.4 Assessment of Learners' Speech

We are also investigating the feasibility of the ASR system to assess Ainu learners' speech. They often write a speech to be presented in a classroom or a contest. We expect the system to provide effective feedback for self-practice. For preliminary investigation, we conducted ASR experiments on speech recordings by three Ainu learners of the museum staff. Although the ASR model was finetuned only with female elderly speakers, it works well for young speakers, including males. Since all learners are at an advanced level, their speech does not include apparent pronunciation errors, and ASR accuracy is almost perfect. We will extend the system for interactive speaking practice.

5. Application of Text-to-Speech (TTS)

5.1 Development of TTS Model

We have also developed a Text-to-Speech (TTS) system using the dataset of the Ainu Language Archive, in particular, the dataset of a single speaker with 20-hour speech. The amount is sufficient for training a state-of-the-art TTS model, such as VITS (Kim 2021). Due to poor audio quality, however, noise reduction and speech enhancement processing were necessary. Although it is difficult to conduct standard subjective evaluations because we cannot recruit native Ainu speakers, the quality of the generated speech is impressive to Ainu scholars and museum staff. It is often difficult to distinguish generated speech from real speech.

5.2 Reference for Speaking Practice

The first use case of the TTS system is to provide a reference for speaking practice. The director of the National Ainu Museum occasionally gives a speech in Ainu, for example, at the opening of a new special exhibition. He can write a speech by himself, but finds it difficult to speak in the proper delivery, as Ainu is very different from Japanese. Thus, he asks us to prepare a synthesized speech for reference in his speaking practice. So far, we have prepared speech material for him four times.

5.3 Revitalization of Old Speech Content

The second use case of the TTS system is to generate speech of Uwepekere folktales, which have transcripts but no audio. In the old days, once transcripts of interviews were made, recording tapes were often recycled, and the audio data was lost. There are several well-known folktales without audio. The examples include "God of Thunder's Sister", interviewed and transcribed in 1958, and "Tale of Bear", scripted in 1950-60s. We generated speech materials for these folktales and provided the audio files to the museum.

5.4 Ethical Issues and Challenges

In many domains of generative AI, copyrights and portrait rights of the source data have become a major issue. The Ainu community is particularly sensitive to this issue because they were afraid of the generation of fake speech apparently told by dead people. They do not allow any use of their speech data without explicit permission, even for academic purposes. The National Ainu Museum obtains consent from the family members of the deceased before making the speech data public in its Archive.

On the other hand, there would be no problem in generating new voice characters for virtual (anime) characters. It would be useful for making new speech content used for public announcements, movies, and educational materials.

In the case of the Ainu language archive, all the speakers are elderly, and the majority are women. This skew in age and gender is generally considered typical of languages in danger of extinction. When we consider applications for educational and recreational content, it is desirable to have a diverse range of speakers. Therefore, we are exploring methods to generate a variety of voices.

6. Conclusions

This paper addresses our work on ASR and TTS applications for the preservation and revitalization of the Ainu language. It was made possible with close collaboration with the staff of the National Ainu Museums and people in the local Ainu community (local autonomy and NPOs). It was crucial for us to listen to them on what is needed and to build human relationships for conducting many trials.

Acknowledgments

The project has been conducted in a collaboration with the National Ainu Museum. We are grateful for many staffs in the Museum and the Ainu community for this collabration. We are also grateful for Prof. Osami Okuda for his kind advice on the Ainu Language.

References

Jokinen, K. (2018). Researching Less-Resourced Languages – the DigiSami Corpus, Proc. LREC.

Watson, C., Keegan, P., Maclagan, M., Harlow, R., and King, J. (2017). The motivation and development of MPAi, a Maori Pronunciation Aid. Proc. Interspeech.

Matsuura, K., Ueno, S., Mimura, M., Sakai, S., and Kawahara, T. (2020). Speech corpus of Ainu folklore and end-to-end speech recognition for Ainu language. Proc. LREC, pp.2622—2628.

Babu. A., Wang, C., Tjandra, A., Lakhotia, K., Xu Q., Goyal N., Singh, K., von Platen, P., Saraf, Y., Pino, A., Baevski, J., Conneau, A., and Auli, M. (2022). XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. Proc. Interspeech

Radford, A., Kim, J-W., Xu, T., Brockman, G, McLeavey, C., and Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision, arXiv:2212.04356.

Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., Elkahky, A., Ni, Z., Vyas, A., Fazel-Zarandi, M., Baevski, A., Adi, Y., Zhang, X., Hsu, W-N., Conneau, A., Auli, M., Scaling Speech Technology to 1,000+ Languages, arXiv:2305.13516.

Kudo, T., and Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing, Proc. EMNLP (Demo Paper).

Kim, J., Kong, J., and Son, J. (2021). Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech. Proc. ICML.