Leveraging End-to-End ASR for Endangered Language Documentation: An Empirical Study on Yoloxóchitl Mixtec

Anonymous EACL submission

Abstract

"Transcription bottlenecks", created by a shortage of effective human transcribers, are one of the main challenges to endangered language (EL) documentation. Automatic speech recognition (ASR) has been suggested as a tool to overcome such bottlenecks. Following this suggestion, we investigated the effectiveness for EL documentation of end-to-end ASR, which unlike Hidden Markov Model ASR systems, eschews linguistic resources but is best in large-data settings. We use a recently available Yoloxóchitl Mixtec EL corpus. First, we review our method in building an end-to-end ASR system in a way that would be reproducible by the ASR community. We then propose a novice transcription correction task and demonstrate how ASR systems and novice transcribers can work together to improve EL documentation. We believe this combinatory methodology would mitigate the transcription bottleneck and transcriber shortage that hinders EL documentation.

1 Introduction

Grenoble et al. (2011) warned that half of the world's 7,000 languages would disappear by the end of the 21st century. Consequently, a concern with endangered language documentation has emerged from the convergence of interests of two major groups: (1) native speakers who wish to document their language and cultural knowledge for future generations; (2) linguists who wish to document endangered languages to explore linguistic structures that may soon disappear. Endangered language (EL) documentation aims to mitigate these concerns by developing and archiving corpora, lexicons, and grammars (Lehmann, 1999). There are two major challenges:

(a) Transcription Bottleneck: The creation of EL resources through documentation is extremely challenging, primarily because the traditional method to preserve such data is not merely with audio recordings but also through time-coded transcriptions. In a best-case scenario, the texts are presented in an interlinear format with aligned parses and glosses along with a free translation (Anastasopoulos and Chiang, 2017). But such (interlinear) transcriptions are difficult to produce in meaningful quantities: (1) ELs often lack a standardized orthography (if written at all); (2) invariably, few speakers can accurately transcribe recordings. Even a highly skilled native speaker or linguist will require approximately 30 to 50 hours to simply transcribe one hour of recording (Do et al., 2014; Zahrer et al., 2020). Additional time is needed for parse, gloss, and translation. This creates what is sometimes known as the "Transcription Bottleneck", where the expert transcribers cannot keep up with the amount of recorded material for documentation.

(b) Transcriber Shortage: It is generally understood that any viable solution to the transcription bottleneck must involve native speaker transcribers. Yet usually few, if any, native speakers have the skills (or time) to transcribe their language. Training new transcribers is one solution, but it is time-consuming, especially with languages that present complicated phonology and morphology. The situation is distinct regarding major languages, for which transcription can be crowd-sourced to speakers with little need for specialized training (Das and Hasegawa-Johnson, 2016). In Yoloxóchitl Mixtec (YM; Glottocode=yolo1241, ISO 639-3=xty), the focus of this study, training is time-consuming: after one-year part-time transcription training, a proficient native speaker, EG,[1] still has problems with certain phones, particularly tones and glottal stops. Documentation requires accurate transcriptions, a goal yet beyond even the capability of an

[1] To offer anonymity, a code is used.

enthusiastic speaker with many months of training.

As noted, ASR has been proposed to mitigate the Transcription Bottleneck and create increasingly extensive EL corpora. Previous studies first investigated HMM-based ASR for EL documentation. Along with HMM-based ASR, natural language processing and semi-supervised learning have been suggested as a way to produce morphological and syntactic analyses (Ćavar et al., 2016; Mitra et al., 2016; Jimerson and Prud'hommeaux, 2018; Cruz and Waring, 2019; Zahrer et al., 2020). As HMM-based systems have become more precise, they have been increasingly promoted as a mechanism to bypass the Transcription Bottleneck. However, ASR's context for ELs is quite distinct from that of major languages. Endangered languages seldom have sufficient extant language lexicons to train an HMM system and invariably suffer from a dearth of skilled transcribers to create these necessary resources (Gupta and Boulianne, 2020).

End-to-end ASR systems have shown comparable or better results over conventional HMM-based methods (Graves and Jaitly, 2014; Chiu et al., 2018; Pham et al., 2019; Karita et al., 2019a). As end-to-end systems directly predict textual units from acoustic information, they save much effort on lexicon construction. Nevertheless, end-to-end ASR systems still suffer from the limitation of training data. Attempts with resource-scarce languages have relatively high character (CER) or word (WER) error rates (Thai et al., 2020; Matsuura et al., 2020; Hjortnaes et al., 2020). It has nevertheless become possible to utilize ASR with ELs to reduce significantly, but not eliminate, the need for human input and annotation to create acceptable ("archival quality") transcriptions.

This Work: This work represents end-to-end ASR efforts on Yoloxóchitl Mixtec (YM), an endangered language from western Mexico. The YMC[2] corpus comprises two sub-corpora. The first ("YMC-EXP", expert transcribed, corpus) includes 100 hours of transcribed speech with carefully proofing. We built a recipe of the ESPNet (Watanabe et al., 2018) that shows the whole process of constructing an end-to-end ASR system using the YMC-EXP corpus. The second corpus, ("YMC-NT", native trainee, corpus) includes 8+ hours of additional recordings not included in the YMC-EXP corpus. This second corpus contains novice transcriptions with subsequent expert corrections. Both the YMC-EXP and YMC-NT corpora are publicly available under a CC BY-SA 3.0 License.[3]

The contributions of our research are:

- A new Yoloxóchitl Mixtec corpus to support ASR efforts in EL documentation.
- A reproducible workflow to build an end-to-end ASR system for EL documentation.
- A comparative study between HMM-based ASR and end-to-end ASR, demonstrating the feasibility of the latter. To test the framework's generalizability, we also experiment with another EL: Highland Puebla Nahuat (Glottocode=high1278; ISO 639-3=azz).
- An in-depth analysis of errors in novice transcription and ASR. Considering the discrepancies in error types, we propose Novice Transcription Correction (NTC) as a task for the EL documentation community. A rule-based method and a voting-based method are proposed.[4] In clean speech, the best system reduces word error rate in the novice transcription by 38.9% .

2 Corpus Description

In this section, we first introduce the linguistic specifics for YM and YMC. Then we discuss the recording settings. Since YM is a spoken language without textual format, we next explain the transcription style designed for this language. Finally, we offer the corpus partition and some statistics regarding corpora size.

2.1 Linguistic Specifics for Yoloxóchitl Mixtec

Yoloxóchitl Mixtec is an endangered, relatively low-resource Mixtecan language. It is mainly spoken in the municipality of San Luis Acatlán, state of Guerrero, Mexico. It is one of some 50 languages in the Mixtec language family, which is part of a larger unit, Otomanguean, that Suárez (1983) considers "a 'hyper-family' or 'stock'." Mixtec languages (spoken in Oaxaca, Guerrero, and Puebla)

[2]Specifically, we used material from the community of Yoloxóchitl (YMC), one of four in which YM is spoken.

[3]To follow the Anonymity rule for EACL, the link of the and the recipe will be published if accepted.

[4]A system combination method, Recognizer Output Voting Error Reduction (Fiscus, 1997))

are highly varied, resulting from approximately 2,000 years of diversification.

YM is spoken in four communities: Yoloxóchitl, Cuanacaxtitlan, Arroyo Cumiapa, and Buena Vista. Mutual intelligibility among the four YM communities is high despite significant differences in phonology, morphology, and syntax. All villages have a simple segmental inventory but fairly extensive tonal contrasts. YMC (refering only to the Mixtec of the community of Yoloxóchitl [16.81602, -98.68597]) manifests 28 tonal patterns on 1,451 identified bimoraic lexical stems. The tonal patterns carry a significant functional load in regards to the lexicon and inflection. For example, 25 distinct tonal patterns on the bimoraic segmental sequence [nama] yield 30 words (including five homophones). This ample tonal inventory presents challenges to both a native speaker learning to write and an ASR system learning to recognize. Notably, it also introduces difficulties in constructing a language lexicon for training of HMM-based systems.

2.2 Recording Settings

There are two corpora used in this study. The first (YMC-EXP) was used for ASR training. The second (YMC-NT) was used to train the novice speaker and for Novice Transcription Correction. The YMC-EXP corpus comprises expertly transcribed audio used as the gold-standard reference for ASR development. The YMC-NT corpus has paired novice-expert transcription as it was used to train and evaluate the novice writer.

The corpus used for ASR development comprises mostly two-channel recordings (split for training). Each of the two speakers was fitted with a separate head-worn mic (usually a Shure SM10a). Over two dozen speakers (mostly male) contributed to the corpus. The topics and their distribution were varied (plants, animals, hunting/fishing, food preparation, ritual speech). The YMC-NT corpus comprises single-channel field recordings made with a Zoom H4n at the moment plants were collected during ethnobotanical research. Speakers were interviewed one after another; there is no overlap. However, the recordings often registered background sounds (crickets, birds) that we expected would negatively impact ASR accuracy more than seems to have occurred. The topic was always a discussion of plant knowledge (a theme of only 9% of the YMC-EXP corpus). Expectedly, there were many out-of-vocabulary (OOV) words (e.g., plant names not elsewhere recorded) in this YMC-NT corpus.[5]

2.3 Corpus Transcription

(a) Transcription Level: The YMC-EXP corpus presently has two levels of transcription: (1) a practical orthography that represents underlying forms; (2) surface forms. The underlying form marks prefixes (separated from the stem by a hyphen), enclitics (separated by an = sign), and tone elision (with the elided tones in parentheses). All these "breaks" and phonological processes disappear in the surface form. For example, the underlying be'^3e^3=an^4 (house=3sgFem; 'her house') surfaces as $be'^3\tilde{a}^4$. And $be'^3e^{(3)}$=^{2} ('my house') surfaces as be'^3e^2. Another example is the completive prefix ni^1-, which is separated from the stem as in ni^1-$xi^3xi^{(3)}$=^{2} (completive-eat-1sgS; 'I ate'). The surface form would be written $n\tilde{i}^1xi^3xi^2$. Again, processes such as nasalization, vowel harmony, palatalization, and labialization are not represented in the practical (underlying) orthography but are generated in the surface forms. The only phonological process encoded in the underlying orthography is tone elision, for which parentheses are used.

The practical, underlying orthography mentioned above was chosen as the default system for ASR training for three reasons: (1) it is easier than a surface representation for native speakers to write; (2) it represents morphological boundaries and thus serves to teach native speakers the morphology of their language; and (3) for a researcher interested in generating concordances for a corpus-based lexicographic project it is much easier to discover the root for 'house' in be'^3e^3=an^4 and $be'^3e^{(3)}$=^{2} than in the surface forms $be'^3\tilde{a}^4$ and be'^3e^2.

(b) "Code-Switching" in YMC: Endangered, colonialized Indigenous languages often manifest extensive lexical input from a dominant Western language, and speakers often talk with "code-switching" (for lack of a better term). Yoloxóchitl Mixtec is no exception. AU[6] considered how to write such forms best and decided that Spanish-origin words would be written in Spanish and without tone when their phonology and meaning are close to that of Spanish. So Spanish *docena* appears over a dozen times in the corpus and is written *tucena*; it always has the meaning of 'dozen'.

[5] After separating enclitics and prefixes as separate tokens, the OOV rate in YMC-NT is 4.84%.

[6] To follow the Anonymity rule for EACL, we use AU for authors of this paper during the reviewing session.

Corpus	Subset	UttNum	Dur (h)
EXP	Train	52763	92.46
	Validation	2470	4.01
	Test	1577	2.52
EXP(-CS)	Train	35144	58.60
	Validation	1301	2.16
	Test	2603	4.35
NT	Clean-Dev	2523	3.45
	Clean-Test	2346	3.31
	Noise-Test	1335	1.60

Table 1: YMC Corpus Partition for EXP (corpus with expert transcription), EXP(-CS) (subset of EXP without "code-switching"), NT (corpus with paired novice and expert transcription)

All month and day names are also written without tones. Note, however, that Spanish *camposanto* ('cemetery') is also found in the corpus and pronounced as $pa^3san^4tu^2$. The decision was made to write this with tone markings as it is significantly different in pronunciation from the Spanish origin word. In effect, words like $pa^3san^4tu^2$ are considered loans into YM and are treated orthographically as Mixtec. Words such as *tucena* are considered "code-switching" and written without tones.

(c) Transcription Process: The initial time-aligned transcriptions were made in Transcriber (Barras et al., 1998). However, given that Transcriber cannot handle multiple tiers (e.g., transcription and translation, or underlying and surface orthographies), the Transcriber transcriptions were then imported into ELAN (Wittenburg et al., 2006) for further processing (e.g., correction, surface-form generation, translation).

2.4 Corpus Size and Partition

Though endangered, YMC does not suffer from the same level of resource limitations that affect most ASR work with ELs (Ćavar et al., 2016; Jimerson et al., 2018; Thai et al., 2020). The YMC-EXP corpus, developed for over ten years, provided 100 hours for the ASR training, validation, and test corpora. There are 505 recordings from 34 speakers in the YMC-EXP corpus, and the transcription for the YMC-EXP are all carefully proofed by an expert native-speaker linguist. As shown in Table 1, we offer a train-valid-test split regarding the speakers. The partition considers the balance between speakers and relative size for each part.

As introduced in Section 2.2, the YMC-NT corpus has *both* expert and novice transcription. It includes only three speakers for a total of 8.36 hours. In the recordings of two consultants, the environment is relatively clean and free of background noise. The speech of the other individual, however, is frequently affected by background noise. This seems coincidental as all three were recorded together, one after the other in random order. But given this situation, we split the corpus into three sets: clean-dev (speaker EGS), clean-test (speaker CTB), and noise-test (speaker FEF; see Table 1).

The "code-switching" discussed in 2.3 (b) introduces different phonological representations and makes it difficult to train an HMM-based model using language lexicons. Therefore, previous work in (Mitra et al., 2016) using the HMM-based system for YMC did not consider sentences with "code-switching". To compare our model with their results, we have used the same experimental corpus in our evaluation. Their corpus (YMC-EXP(-CS)), shown in Table 1, is a subset of the YMC-EXP that does not contain "code-switching" utterances.

3 ASR Experiments

3.1 End-to-End ASR

As ESPNet (Watanabe et al., 2018) is widely used in open-source end-to-end ASR research, our end-to-end ASR systems are all constructed using ESPNet[7]. For the encoder, we employed the conformer structure (Gulati et al., 2020), while for the decoder we used the transformer structure to condition the full context, following the work of Karita et al. (2019b). The conformer architecture is a state-of-the-art innovation developed from the previous transformer-based encoding methods (Karita et al., 2019a). A comparison between the conformer and transformer encoders shows the value of applying state-of-the-art end-to-end ASR to ELs.

3.2 Experiments and Results

As discussed above, our end-to-end model applied an encoder-decoder architecture with a conformer encoder and a transformer decoder. The architecture of the model follows Gulati et al. (2020) while its configuration follows the aishell conformer recipe from ESPNet (Watanabe et al., 2018).[8] The experiment is reproducible using ESPNet (Watanabe et al., 2018).

[7] To follow the Anonymity rule for EACL, we will upload our model construction process as a part to the ESPNet recipe

[8] See Appendices for details about the model configuration.

As the end-to-end system models are based on word pieces, we adopted CER and WER as evaluation metrics. They help demonstrate the system performances at different levels of graininess. But because the HMM-based systems were decoding with a word-based lexicon, for comparison to HMM we only use the WER metric. To thoroughly examine the model, we conducted several comparative experiments, as discussed in continuation.

(a) Comparison with HMM-based Methods: We first compared our end-to-end method with the Deep Neural Network-Hidden Markov Model (DNN-HMM) methods proposed in (Mitra et al., 2016). In Mitra et al. (2016)'s work, Gammatone Filterbanks (GFB), articulation, and pitch are configured for the DNN-HMM model. This baseline is a DNN-HMM model using Mel Filterbanks (MFB). In recent unpublished work, Kwon and Kathol develop a latest state-of-the-art CNN-HMM-based ASR model[9] for YMC based on on the lattice-free Maximum Mutual Information (LF-MMI) approach, also known as "chain model" (Povey et al., 2016). The experimental data of the above HMM-based models is YMC-EXP(-CS) discussed in Section 2.4. For the comparison, our end-to-end model adopted the same partition to ensure fair comparability with their results.

Table 2 shows the comparison between DNN-HMM systems and our end-to-end system on YMC-EXP(-CS). It indicates that the end-to-end system significantly outperforms the DNN-HMM baseline model. Moreover, without an external language lexicon it reaches a performance level comparable to that of the CNN-HMM-based state-of-the-art model.

Model	Feature	WER
DNN-HMM	MFB	36.9
DNN-HMM	GFB + Articu. + Pitch	31.1
CNN-HMM (Chain)	MFCC	**19.1**
E2E-Conformer	MFB + Pitch	20.6

Table 2: Comparison between HMM-based Models and the End-to-End Conformer (E2E-Conformer) Model on YMC-EXP(-CS) that is a subset of the YMC-EXP without "code-switching".

In Section 2.3 (b), we note that "code-switching" is invariably present in EL speech (e.g., YMC). Thus, ASR models built on "code-switching-free corpora (like YMC-EXP[-CS]) are not practical for real-world usage. However, a language lexicon is available only for the YMC-EXP(-CS) corpus so we cannot conduct HMM-based experiments either YMC-EXP or YMC-NT.

(b) Comparison with Different End-to-End ASR Architectures: We also conducted experiments comparing models with different encoders and decoders on the YMC-EXP corpus. For a Recurrent Neural Network-based (E2E-RNN) model, we followed the best hyper-parameter configuration, as discussed in Zeyer et al. (2018). For a Transformer-based (E2E-Transformer) model, the same configuration from Karita et al. (2019b) was adopted. Both models shared the same data preparation process as the E2E-Conformer model.

Table 3 compares different end-to-end ASR architectures on the YMC-EXP corpus.[10] The E2E-Conformer obtained the best results, obtaining 15% and 9% relative WER improvement to E2E-RNN and the E2E-Transformer model. The E2E-Conformer's WER on YMC-EXP(-CS) is slightly lower than the whole YMC-EXP, despite a significantly smaller training set in the YMC-EXP(-CS) corpus. Since the subset excludes Spanish words, "code-switching" may well be a problem to consider in ASR for endangered languages such as YM.

Model	CER dev/test	WER dev/test
E2E-RNN	11.7/11.7	24.8/24.8
E2E-Transformer	10.8/10.8	23.0/23.2
E2E-Conformer	**9.9/10.0**	**20.8/21.1**

Table 3: End-to-End ASR Results on YMC-EXP (corpus with "code-switching")

(c) Comparison with Different Transcription Levels: In addition to comparing model architectures, we compared the impact of transcription levels on the ASR model. E2E-Conformer models with the same configurations were trained using both the surface and the underlying transcription forms, which is introduced in Section 2.3. We also trained separate RNN language models for fusion and unigram language models to extract word pieces for different transcription levels.

[9] See Appendices for details about the model configuration.

[10] The train set in YMC-EXP is significantly larger than that in YMC-EXP(-CS).

Transcription Level	CER dev/test	WER dev/test
Surface	10.2/**9.9**	21.6/21.2
Underlying	**9.9**/10.0	**20.8/21.1**

Table 4: E2E-Conformer Results for Two Transcription Levels (Underlying represents morphological divisions and underlying phonemes before the application of phonological rules; Surface is reflective of spoken forms and lacks morphological parsing)

Table 4 shows the E2E-Conformer results over different transcription levels. As introduced in Section 2.3, the surface form reduces several linguistic and phonological processes compared to the underlying practical form. The results indicate that the end-to-end system is able to automatically infer those morphological and phonological processes and maintain a consistent low error rate.

(d) Comparison with Different Corpus Size: As introduced in Section 1, most ELs are considered low-resources for the ASR system. Thus, we trained the E2E-Conformer model on 10, 20, and 50 hours subset of YMC-EXP to demonstrate the model performances over different sizes of resources.

Corpus	CER dev/test	WER dev/test
10h	31.9/31.9	59.9/59.9
20h	20.6/20.7	42.1/42.1
50h	11.6/11.6	24.4/24.5
Whole (92h)	**9.9/10.0**	**20.8/21.1**

Table 5: E2E-Conformer Results on Different Corpus Size

Table 5 shows the E2E-Conformer performances on different amounts of training data. It demonstrates how the model consumes data. As corpus size is incrementally increased, WER decreases significantly. It is apparent that the model still has the capacity to improve performance with more data. The result also indicates that our system can get reasonable performances from 50 hours of data. This would be an important guideline when we collect a new EL database.

(e) The Framework Generalizability: To test the end-to-end ASR systems' generalization ability, we conducted the same end-to-end training and test procedures on another endangered language: Highland Puebla Nahuatl (high1278; azz). The corpus is also open access.[11] It comprises 954 recordings that total 185 hours 22 minutes.[12]

Table 6 shows the performance of three different end-to-end ASR architectures on Highland Puebla Nahuatl. For this language the E2E-Conformer again offers better performances over the other models. These experiments indicate the general ability to consistently apply end-to-end ASR systems across ELs.

Model	CER dev/test	WER dev/test
E2E-RNN	11.0/10.3	27.6/25.7
E2E-Transformer	10.8/**10.0**	27.9/26.0
E2E-Conformer	**10.5/10.0**	**26.4/25.4**

Table 6: E2E-Conformer Results on another EL: Highland Puebla Nahuatl

4 Novice Transcription Correction

This paper presents novice transcription correction (NTC) as a task for EL documentation. We first analyze patterns manifested in novice transcriptions. Next, we introduce two baselines that fuse ASR hypotheses and novice transcription for the NTC task.

4.1 Novice Transcription Error

As mentioned in Section 1, transcriber shortages have been a severe challenge for EL documentation. Before 2019, only the native speaker linguist, AU, could accurately transcribe the segments and tones of YMC. To mitigate the YMC transcriber shortage, AU began to train another speaker, EG, in 2019. First, a computer course was designed to incrementally teach EG segmental and tonal phonology. In the next stage, he was given YMC-NT corpus recordings to transcribe. Compared to the paired expert transcription, the novice achieved a CER of 6.0% on clean-dev, defined in Table 1. However, it is not feasible to spend many months training speakers with no literacy skills to acquire the transcription proficiency achieved by EG in our project. Moreover, even with a 6.0% CER, there are still enough errors so as to require significant annotation/correction. The state-of-the-art ASR system (e.g., E2E-Conformer) shown in Table 3 gets an 8.2% CER on the clean-dev set, more errors than

[11] The corpus will be publicly available with YM.

[12] the recordings are almost all with two channels and two speakers in natural conversation

Error Types	Novice	ASR
Enclitics (=)	**96**	243
Prefixes (-)	141	**62**
Glottal Stop (’)	341	**209**
Parenthesis	1607	**302**
Tone	4144	**3241**
Stem-Nasal (n)	**0**	6
Others	**4263**	10175
Total	10592	14232

Table 7: Character Error-type Distribution of Novice and ASR (by number of errors)

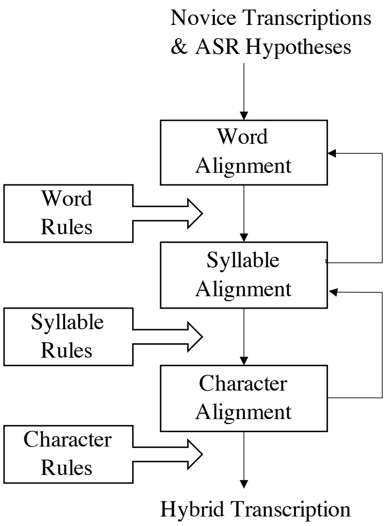

Figure 1: Novice-ASR Fusion Process

the novice CER. So for YMC, ASR is still not a good enough substitute for a proficient novice.

As AUs worked with the novice, they saw a repetition of types of errors that they worked to correct by giving the novice exercises focused on these transcription shortcomings. The end-to-end ASR, however, has demonstrated a different pattern of errors. For example, it developed a fair understanding of the rules for suppleting tones, marked by parentheses around the suppleted tones. Rather than over-specify the NTC correction algorithm, we first analyzed the error-type distribution using the Clean-dev from the YMC-NT corpus, as shown in Table 7.

4.2 Novice-ASR Fusion

Rapid comparison of the types of errors for each transcription (novice and ASR) demonstrated consistent patterns and has led us to hypothesize that a fusion system might automatically correct many of these errors. Two baseline methods are examined for the fusion: a voting-based system (Fiscus, 1997) and a rule-based system.

The voting-based system follows the definition in (Fiscus, 1997) that combines hypotheses from different ASR models with Novice transcription.

The framework of rule-based fusion is shown in Figure 1. The rules are defined in different linguistic units: words, syllables, and characters. They assume a hierarchical alignment between the novice transcription and ASR hypotheses. The rules are applied to the transcription from word to syllable to character level. The rules are developed based on interaction with a novice's progress. Thus they will be different but discoverable when applying to a new language. However, the general principle should be adaptable to other ELs: Novice trainees will learn certain transcription tasks easier than others. Below we explain the rules for YMC.

Word Rules: If a word from the novice transcription is Spanish (i.e., no tones and no linguistic indications [-, =, ’] that mark it as Mixtec), keep the novice transcription. If the novice has extra words, not in ASR, keep those extra words.

Syllable Rules: If a novice syllable is tone initial, use the corresponding ASR syllable. If the novice and the ASR have identical segments but different tones, use the ASR tones. When an ASR syllable has CVV or CV’V, and its corresponding novice syllable has CV,[13] use the ASR syllable (CVV or CV’V). If the tone from either transcription system follows a consonant (except a stem-final n), use the other system's transcription.

Character Rules: If the ASR has the following different linguistic symbols from the novice transcription: hyphens, equal signs, parentheses, glottal stops, then always trust the ASR.

We apply the edit distance (Wagner and Fischer, 1974) to find the alignment between the ASR model hypothesis $\{C_1, ..., C_n\}$ and the Novice transcription $\{C'_1, ..., C'_m\}$. The L_I, L_D, L_S are introduced in the dynamic function as the insertion, deletion, and substitution loss, respectively. In the naive setting, L_I, L_D are both set to 1. The L_S is set to 1 if C_i is different from C'_j and 0 otherwise. This setting is computation-efficient. However, it does not consider how the contents mismatch between the C_i and C_j. Therefore, we adopt a hierarchical dynamic alignment. In this method, the character

[13] A CV syllable can occur in a monomoraic word. But novice will often write a CV word when it should be CVV or CV’V. Stem-final syllables can be CV, CVV or CV’V. But novice tends to write CV in these cases.

Model	Clean-Dev		Clean-Test		Noise-Test		Overall	
	CER	WER	CER	WER	CER	WER	CER	WER
A. Novice	6.0	21.5	6.4	22.6	**8.4**	**26.6**	6.8	23.1
B. E2E-Transformer	9.8	23.1	8.8	21.2	24.3	47.0	12.9	28.1
C. E2E-Conformer	8.2	19.6	8.2	19.1	23.6	44.1	12.0	25.3
D. Fusion1 (A+C)	6.3	20.6	6.9	22.0	13.1	38.6	8.2	25.4
E. Fusion2 (A+C)	5.1	17.6	5.5	18.7	9.6	30.3	6.3	21.1
F. ROVER (A+B+C)	4.7	**14.6**	**4.6**	**13.8**	12.4	32.6	6.5	**18.5**
G.ROVER-Fusion2 (A+B+C+E)	**4.5**	16.1	4.7	16.7	9.0	28.3	**5.7**	19.3

Table 8: NTC Results on YMC-NT (the results are evaluated using the expert transcription in YMC-NT)

alignment follows the native setting. While the $L_S(C_i, C'_j)$ for syllable alignment is defined as the normalized character-level edit distance between C_i and C'_j as follows:

$$L_S(C_i, C'_j) = \frac{D[C_i, C'_j]}{|C_i|} \quad (1)$$

where the $|C_i|$ is the lengths of the syllable. Similarly, the $L_S(C_i, C'_j)$ for word alignment is defined based on syllable alignment.

5 NTC Experiments

5.1 Experimental Settings

The novice transcription, the E2E-Transformer model, and the E2E-Conformer model were considered as baselines for the NTC task. For the end-to-end models, we adopted the trained model from Section 3 with the same decoding set-ups. To test the effectiveness of the hierarchical dynamic alignment, we tested the data with two fusion systems, namely Fusion1 and Fusion2. The Fusion1 system used the naive settings of edit distance, while the Fusion2 system adopted the hierarchical dynamic alignment. Both fusion systems adopt rules defined in Section 4. Two configurations for voting-based methods were tested. The first "ROVER" combined three hypotheses (i.e., the E2E-Transformer, the E2E-Conformer, and the Novice). In contrast, the "ROVER-Fusion2" combined the Fusion2 system with the above three.

5.2 Results

As shown in Table 8, voting-based methods and rule-based methods all significantly reduce the novice errors for clean speech. However, for the noise-test, the novice transcription is the most robust method. For overall results, the ROVER system has a lower WER, while the ROVER-Fusion2 system reaches a lower CER.

As we discussed in Section 4, novice and ASR transcriptions manifest different error patterns and they can be complementary. Table 8 shows that our proposed rule-based and voting-based fusion methods can potentially eliminate the errors come from the novice transcriber, and it can mitigate the transcriber shortage problems based on this fusion methods. However, we should note that the noisy recording condition would be harmful for the fusion, and we should rely on the novice transcriber in such a condition for a practical use case.

6 Conclusion and Future Work

This work presents an open-source endangered language corpus in Yoloxóchitl Mixtec and a comparative study towards its end-to-end ASR systems in a reproducible manner. We demonstrate that end-to-end approaches are feasible and present comparable results over conventional HMM ASR approaches that require resources such as language lexicons. Additionally, we propose novice transcription correction as a potential task for ASR in EL documentation. We examine two methods for this task. First, a rule-based approach uses hierarchical dynamic alignment and linguistic rules to perform novice-ASR hybridization. Second, a voting-based method combines hypotheses from the novice and end-to-end ASR systems. Empirical studies on the YMC-NT corpus indicate that both methods significantly reduce the CER/WER of the novice transcription for clean speech.

The above discussion suggests that a useful approach to EL documentation using both human and computational (ASR) resources might focus on training each for particular transcription tasks. If we know from the start that ASR will be used to correct novice transcriptions in areas of difficulty, we could train an ASR system to maximize accuracy for those areas that challenge novice learning.

References

Antonios Anastasopoulos and David Chiang. 2017. A case study on using speech-to-translation alignments for language documentation. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 170–178.

Claude Barras, Edouard Geoffrois, Zhibiao Wu, and Mark Liberman. 1998. Transcriber: a free tool for segmenting, labeling and transcribing speech. In *Preceedings of the First international conference on language resources and evaluation (LREC)*, pages 1373–1376.

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)*, pages 4004–4011.

Chung-Cheng Chiu, Tara N Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J Weiss, Kanishka Rao, Ekaterina Gonina, et al. 2018. State-of-the-art speech recognition with sequence-to-sequence models. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4774–4778. IEEE.

Hilaria Cruz and Joseph Waring. 2019. Deploying technology to save endangered languages. *arXiv preprint arXiv:1908.08971*.

Amit Das and Mark Hasegawa-Johnson. 2016. An investigation on training deep neural networks using probabilistic transcriptions. In *Interspeech*, pages 3858–3862.

Thi-Ngoc-Diep Do, Alexis Michaud, and Eric Castelli. 2014. Towards the automatic processing of yongning na (sino-tibetan): developing a 'light'acoustic model of the target language and testing 'heavyweight'models from five national languages. In *Spoken Language Technologies for Under-Resourced Languages*.

Jonathan G Fiscus. 1997. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover). In *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, pages 347–354. IEEE.

Pegah Ghahremani, Bagher BabaAli, Daniel Povey, Korbinian Riedhammer, Jan Trmal, and Sanjeev Khudanpur. 2014. A pitch extraction algorithm tuned for automatic speech recognition. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 2494–2498. IEEE.

Alex Graves and Navdeep Jaitly. 2014. Towards end-to-end speech recognition with recurrent neural networks. In *International conference on machine learning*, pages 1764–1772.

Lenore A Grenoble, Peter K Austin, and Julia Sallabank. 2011. Handbook of endangered languages.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

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.

Nils Hjortnaes, Niko Partanen, Michael Rießler, and Francis M Tyers. 2020. 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*, pages 31–37.

Robbie Jimerson and Emily Prud'hommeaux. 2018. Asr for documenting acutely under-resourced indigenous languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Robert Jimerson, Kruthika Simha, Ray Ptucha, and Emily Prud'hommeaux. 2018. Improving asr output for endangered language documentation. In *The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*.

Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson Enrique Yalta Soplin, Ryuichi Yamamoto, Xiaofei Wang, et al. 2019a. A comparative study on transformer vs rnn in speech applications. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 449–456. IEEE.

Shigeki Karita, Nelson Enrique Yalta Soplin, Shinji Watanabe, Marc Delcroix, Atsunori Ogawa, and Tomohiro Nakatani. 2019b. Improving transformer-based end-to-end speech recognition with connectionist temporal classification and language model integration. *Proceedings of Interspeech 2019*, pages 1408–1412.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.

Christian Lehmann. 1999. Documentation of endangered languages. In *A priority task for linguistics. ASSIDUE Arbeitspapiere des Seminars für Sprachwissenschaft der Universität Erfurt*, 1.

Kohei Matsuura, Sei Ueno, Masato Mimura, Shinsuke Sakai, and Tatsuya Kawahara. 2020. Speech corpus of ainu folklore and end-to-end speech recognition for ainu language. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2622–2628.

Vikramjit Mitra, Andreas Kathol, Jonathan D Amith, and Rey Castillo García. 2016. Automatic speech transcription for low-resource languages-the case of yoloxóchitl mixtec (mexico). In *Proceedings of Intespeech 2016*, pages 3076–3080.

Ngoc-Quan Pham, Thai-Son Nguyen, Jan Niehues, Markus Müller, and Alex Waibel. 2019. Very deep self-attention networks for end-to-end speech recognition. *Proceedings of Interspeech 2019*, pages 66–70.

Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur. 2016. Purely sequence-trained neural networks for asr based on lattice-free mmi. In *Interspeech*, pages 2751–2755.

Jorge A Suárez. 1983. *The mesoamerican indian languages*. Cambridge University Press.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 4278–4284.

Bao Thai, Robert Jimerson, Raymond Ptucha, and Emily Prud'hommeaux. 2020. Fully convolutional asr for less-resourced endangered languages. 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 126–130.

Robert A Wagner and Michael J Fischer. 1974. The string-to-string correction problem. *Journal of the ACM (JACM)*, 21(1):168–173.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson-Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, et al. 2018. Espnet: End-to-end speech processing toolkit. *Proceedings of Interspeech 2018*, pages 2207–2211.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. Elan: a professional framework for multimodality research. In *5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1556–1559.

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.

Albert Zeyer, Kazuki Irie, Ralf Schlüter, and Hermann Ney. 2018. Improved training of end-to-end attention models for speech recognition. *Proceedings of Interspeech 2018*, pages 7–11.

A Appendices

Experimental Settings for End-to-End ASR: All the end-to-end ASR systems adopted the hybrid CTC/Attention architecture integrated with an RNN language model. It selected the best model by performance on the development set. The input acoustic features were 83-dimensional log-mel filterbanks features with pitch features (Ghahremani et al., 2014). The window length and the frameshift were set to 25ms and 10ms. The prediction targets were the word pieces trained using the unigram language modeling (Kudo and Richardson, 2018). The CTC ratio for Hybrid CTC/Attention was set to 0.3. The decoding beam size was 20. Training and Testing are based on Pytorch.

E2E-Conformer Configuration: The E2E-Conformer used 12 encoder blocks and 6 decoder blocks. All the blocks adopted 2048 dimension feed-forward layer and four-head multi-head-attention with 256 dimensions. Kernel size in Conformer block was set to 15. For training, batch-size was set 32. Adam optimizer with 1.0 learning rate and Noam scheduler with 25000 warmup-steps were used in the training. We trained for a max epoch of 50.

E2E-RNN Configuration: The E2E-RNN used 3 encoder blocks and 2 decoder blocks. All the blocks adopts 1024 hidden units. Location-based attention adopted a 1024-dim attention. Adadelta was chosen as the optimizer and we trained for a max epoch of 15.

E2E-Transformer Configuration: The E2E-Transformer used 12 encoder blocks and 6 decoder blocks. All the blocks adopted 2048 dimension feed-forward layer and four-head multi-head-attention with 256 dimensions. Adam optimizer with 1.0 learning rate and Noam scheduler with 25000 warmup-steps were used in the training. We trained for a max epoch of 100.

Experimental Settings for HMM-based ASR: Acoustic feature input for this model are 40 dimensional Mel Frequency Cesptral Coefficients (MFCC). The chain model is trained with a sequence-level objective function and operates with an output frame rate of 30 ms, which is three times longer than the previous standard. The longer frame rate increases decoding speed, which in turn makes it possible to operate with a significantly deeper DNN architecture for acoustic modeling. The best results were achieved with a neural network based on the ResNet architecture (Szegedy et al., 2017). This consists of an initial layer for Linear Discriminative Analysis (LDA) transformation and subsequent alternating 160-dimensional bottleneck layers, adding up to 45 layers in total. The DNN acoustic model is then compiled with a 4-gram language model into a weighted finite state transducer for word sequence decoding.