Comparing efficacy of IPA vs Pinyin romanisation transcriptions for complex tonal languages: A case study in Baima

Katia Chirkova
CNRS-CRLAO
katia.chirkova@gmail.com

Rolando Coto-Solano
Dartmouth College
rolando.a.coto.solano@dartmouth.edu

Rachael Griffiths
EPHE-PSL
rachael.griffiths@ephe.psl.eu

Marieke Meelen
University of Cambridge
mm986@cam.ac.uk

Abstract

How is automated tone transcription affected by the choice of transcription orthography? In this paper we present a range of experiments that indicate that, even when the tonal representations are kept the same, the way vowels and consonants are transcribed can affect tonal character outputs. Our results also indicate that using a Language Model (LM) for decoding can mitigate problems with tonal outputs, but tones remain the most difficult part of the transcription. In doing this we also present the first Automatic Speech Recognition (ASR) models for the Baima language, spoken in Sichuan and Gansu, China. We hope to use these models to contribute to ongoing documentation efforts.

1 Introduction

Researchers who start documenting endangered languages without writing systems often face the challenge of a race against the clock to collect and transcribe as much data as possible before the language disappears. With extremely limited access to native speakers who are not only essential when gathering, but also when transcribing and interpreting data, linguists and community members interested in preserving the language have to make crucial choices on how to spend limited time with informants. Is it worth the tremendous amount of time and effort to preserve every detail using the International Phonetic Alphabet (IPA) to facilitate further research in the sound system of the language? Or should they choose a local and/or romanised script to speed up transcription and to increase the possibilities of language revitalisation?

In this paper we present several ASR experiments to gain further insight into these important practical questions, focusing on the Baima language, spoken in Sichuan and Gansu, China. With three native tones, tone sandhi and tonal borrowings as well as complex consonantal onsets and epiglottalisation, this language forms the perfect case to test the trade-off of different transcription systems. In addition to tests with different base models, LMs and transcription systems, we will also do an in-depth error analysis of each of the tones to gain insight into which are more challenging for specific models. The results will therefore not only further work on ASR for tonal languages but also help researchers and speaker communities working on language documentation and revitalisation to choose how to best spend limited time and resources in order to get the best possible results.

1.1 Baima Language

Baima (/pêkê/, Chinese 白马语 *báimǎyǔ*, ISO-639 code *bqh*) is a Tibeto-Burman (Tibetic) language spoken at the border of Sichuan and Gansu provinces in China. The language has approximately 10,000 speakers, who reside in the counties of Pingwu, Songpan (Tib. *Zung chu*), and Jiuzhaigou (Tib. *Gzi rtsa sde dgu*) in Sichuan, and in the counties of Wenxian and Zhouqu (Tib. *'Brug chu*) in Gansu.

The area of distribution of the Baima language lies at the historical Sino-Tibetan border, in a multi-ethnic and multilingual region. In all counties of its present distribution, immediate linguistic neighbours of Baima include varieties of Mandarin (mostly Southwestern Mandarin) and Tibetic languages. To our knowledge, there are no longer any monolingual speakers of Baima, as all age groups are bilingual in the local varieties of Mandarin. Mandarin (both the local varieties and the closely related Standard Mandarin, the official language of the People's Republic of China) also dominates the education system and work in public domains. Baima is not used in writing or education and its use is mostly restricted to family and community events. For those reasons, it is severely endangered.[1]

[1] https://www.ethnologue.com/language/bqh/

Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 170–181
March 4-5, 2025 ©2025 Association for Computational Linguistics

Baima is little-studied. To date, most linguistic fieldwork on this language has concentrated on the Baima variety as spoken in Baima Township of Pingwu County, which is also the focus of the present study (Huang and Zhang, 1995; Chirkova, 2017; Sun et al., 2007). A small set of audio-visual non-annotated recordings of Pingwu Baima is available on the Pangloss archive of the Centre national de la recherche scientifique (CNRS).[2] Speakers of Baima are keen to preserve their language and cultural traditions and would greatly benefit from the development of tools that can facilitate this effort.

Baima is remarkable for its phonological complexity, and for a number of features that are typologically uncommon. These include non-modal phonation type contrasts in both consonants and vowels and a tonal system characterised by syllable-level contrasts, with redundant use of pitch, voice quality, and vowel length. The Baima consonant inventory consists of 57 phonemes, including 11 epiglottalised prenasalised, nasal, and approximant phonemes. The vowel inventory consists of 11 monophthongs, three native diphthongs, and one diphthong that only occurs in loanwords from Mandarin (/ua̱/). The three contrastive tonal categories are high falling (53), mid (44), and low (213). The high falling tone is correlated with a high falling pitch contour, tense vowel quality, and short vowel duration. The mid and low tone categories are correlated with long vowel duration. The mid tone has a mid level pitch contour and a modal voice quality. The low tonal category has a low falling-rising pitch contour and a breathy-like or lax voice quality. Detailed phonological analyses can be found in Chirkova et al. (2023) and Chirkova (2025), and examples of the tones are given in Table 1.

Baima Tones		
Category	**Example**	**Meaning**
1. Contrastive tones in the native lexicon		
High falling	no^{53}	inside
Mid level	no^{44}	sky, heaven
Low rising	no^{213}	exist, have
2. Tone sandhi		
Compound change	no^{31}mba^{53}	possessions
3. Tones in Chinese loans		
High level	t^{h}a^{55}	he/she/it
Mid rising	tʂ^{h}ɑ^{35}	examine

Table 1: Tones in Baima in IPA

1.2 ASR for No-Resource Tonal Languages

As there are no NLP efforts, corpora, dictionaries or other resources available for Baima, we have to resort to techniques to address the well-known transcription bottleneck for extremely low- (or no-)resource languages. Recent work by (Stoian et al., 2020; Prud'hommeaux et al., 2021; Coto-Solano et al., 2022) and others show getting good ASR results in those challenging situations is possible by relying on pre-trained models of acoustic data from other, high-resource languages. In addition, some techniques involve transfer learning or modification of the acoustic signal (Mitra et al., 2012; Meelen et al., 2024), data augmentation with written sources such as dictionaries and word lists (Hjortnaes et al., 2020; Arkhangelskiy, 2021).

Languages like Baima with complex phoneme inventories and tones are generally more challenging for any ASR model, especially when data and resources are scarce or non-existent. For ASR systems that evaluate the Character Error Rate (CER), it is therefore important to think carefully about the transcription method, as CER has been shown to strongly correlate with orthographic complexity (Taguchi and Chiang, 2024). Representations where the tone is marked explicitly but kept separate from the vowel (i.e. explicit tone recognition, as discussed by Lee et al. (2002)) are not often used for larger languages, but they are common in low-resource ASR systems, such as those for Yongning Na from China and Eastern Chatino from Mexico (Adams et al., 2018). Coto-Solano (2021) shows that manipulating the transcription input can improve results in a language like Bribri, where not marking the most common tone and separating the tonal markings from the vowel can lead to major improvements in performance. Bribri has only four tones, however, transcribed with a limited number of additional segments, and only when necessary. Baima, on the other hand has three native tones, tone sandhi as well as additional tones on loanwords for Chinese. Following sinological tradition, those are all represented with Chao tone numbers (Chao, 1930), which means they consist of at least 2-3 additional characters on every tonal syllable. The current use of complex tone notation in Baima is in line with the research tradition that characterises Baima as a tonal language defined by pitch, favouring Chao tone numbers over IPA diacritics for tone representation (see Section 2.2). The fact

[2] https://pangloss.cnrs.fr/corpus/Baima?lang=en&mode=pro

that Baima tones are produced with both a particular f0 specification and a voice quality specification has only been recently discovered (Chirkova et al., 2023; Chirkova, 2025).

In this paper we therefore focus on transcription systems and how they might impact the automatic transcription of complex tones, testing different base models as well as the usefulness of adding an LM in an extremely low/no-resource context.

2 Methodology

2.1 Data collection

The data for Baima used in the present study was collected during two fieldtrips in November-December 2003 and October-November 2004 in several villages in the Baima Township of Pingwu County. We collected ca. 20 hours of traditional narratives, interviews, and descriptions of traditional activities. 191 minutes (4 hours 5 minutes) are fully transcribed. All of the transcribed materials are from recordings of traditional narratives from three native speakers (all male, between 50+ and 70+ years old at the time of recording).

To enhance efficiency of fine-tuning the base models and to avoid potential confounds in the results due to differences in segment length, we excluded segments longer than 15 seconds, reducing the dataset to 186 minutes. The total corpus contains 27,417 words (2715 unique words).

2.2 Transcription methods

The original transcriptions of recordings in the Baima language were done in IPA capturing all phonetic details of the language, including nasalisation, epiglottalisation and tones. While nasalisation is not phonemic, there is variation between different speakers. Epiglottalisation and tones are phonemic, however, and the latter are indicated with Chao tone numbers in our transcriptions.

The Pinyin-style transcription was created with the primary objective of being comprehensible to native speakers of Baima. It is rooted in the Hanyu Pinyin system, the official romanization system in China (Committee, 1956). The choice to establish a romanization system for the Baima language on the basis of the Hanyu Pinyin system was influenced by two crucial factors: (i) its widespread familiarity, which is a result of its extensive usage in elementary school education and public life, and (ii) its ease of adaptation for electronic applications, particularly mobile phones.

Over the past few decades, the Hanyu Pinyin system has served as the foundation for romanization systems of numerous minority languages in China, including large languages such as Nuosu (see (Ma et al., 2008)). It has also been instrumental in our own work on the Duoxu and Ersu languages (Chirkova and Han (2016); Chirkova and Wang (2017); Wang et al. (2019)). It is worth noting that the issue of tone notation in the Hanyu Pinyin system is intricate. The official system employs diacritics to represent the four tones of Standard Chinese. Nevertheless, these diacritics are often disregarded in various contexts, such as when spelling Chinese names. Alternatively, tones can be indicated by placing a tone number (1 to 4) at the end of each individual syllable.

In essence, transcribing tone remains a challenging aspect for speakers of tonal languages, such as Mandarin Chinese speakers and those whose languages we developed romanization systems for in the past. Therefore, it is crucial to engage in careful consultation with potential users of the system to address the issue of tone representation. We chose Chao tone numbers for several reasons. First, the complexity of the tonal system of Baima has only recently begun to be unravelled. While recent research has provided a better understanding of contrastive tones on monosyllabic words, tone sandhi in polysyllabic words remains largely understudied. Consequently, Chao tone numbers offer the most accurate and reliable method for noting tone variation before a comprehensive understanding of the tonal system is achieved. Secondly, the tradition of using Chao tone numbers in IPA transcription is deeply rooted in the field. The vast majority of publications on Baima, including the only reference grammar with the most comprehensive vocabulary list to date (Sun et al., 2007), rely on this system. Therefore, Chao tone numbers provide convenience for cross-reference and comparison between our work and previous descriptions of that language.

To facilitate testing of different transcription systems, we wrote one-way conversion scripts to create Pinyin and Simple romanisation equivalents of the detailed IPA transcriptions with tones.[3]. These conversions can only be done from IPA, as certain details are simplified in both alternative transcription systems. Table 2 shows examples for each

[3] Code and models can be found on https://github.com/rolandocoto/baima-asr

Transcription	With tones	No tones
IPA	ɲə^{53}	-
Pinyin	nyii^{53}	nyii
Simple	nyə^{53}	nyə

Table 2: Possible transcriptions for [ɲə^{53}] ‘man’

	Source	Hypothesis
1. Get hypothesis	[ɲə^{53} te^{53}]	[ɲə^{53} te^{44}]
2. Get only tones	53 53	53 44
3. One unit per tone	F F	F H
4. Calculate error	tCER=50, tWER=50	

Table 3: Example of the calculation of tonal CER and WER for the human-transcribed phrase [ɲə^{53} te^{53}] ‘that man’ and a potential automatic (and partially wrong) transcription of the phrase.

transcription type. All transcriptions have the same representation for tone: two numerical characters for contours (e.g. ^{53} for the falling tone) and three characters for the dipping tone (i.e. ^{213} dipping).

2.3 ASR Training

Our next step was to create ASR models for the Baima language. We carried out monolingual fine-tuning using the Baima data, and we chose three base models for this[4]: Wav2Vec2 XLSR-53 Large (Baevski et al., 2020), henceforth Wav2Vec2; MMS 1b-all (Pratap et al., 2024), henceforth referred to as MMS; and Whisper Medium (Radford et al., 2023). For Wav2Vec2 and MMS we tried versions of the models with and without an LM for decoding. We used KenLM (Heafield, 2011) to produce the LMs, which were trained using the text in the training and validation partitions of the data.

In order to train the models, we took the total 186 minutes of data and created 20 randomly ordered versions of it. We split these 20 versions into train/dev/test sets, with ratios of 80%, 10% and 10%. We used these partitions to train the models, and the checkpoint before overfitting was saved. These were used to evaluate the test sets, and from there calculate the median CER and Word Error Rate (WER) for each test set. In section 3 we report the average values of the median error for each randomly assigned test set.[5] The total sample only has three speakers, so the speakers in the train/valid sets are also present in the test set.

2.4 Calculation of Tonal Errors

In addition to reporting the standard CER and WER, we also calculated the metrics of tonal character error rate (tonal CER) and tonal word error rate (WER). Table 3 shows an example of this process. Let’s assume we have the phrase [ɲə^{53} te^{53}] ‘that man’ as a human-transcribed phrase in the test set. It is transcribed [ɲə^{53} te^{53}] in IPA with Chao tone numbers. Let’s then assume that one of the ASR systems produces the wrong automatic transcription [ɲə^{53} te^{44}]. Here, the falling (53) tone of the first word is correct, but the tone of the second word is incorrectly tagged as a mid level (44) tone. We then strip both phrases of their consonants and vowels, leaving only the tones. This would result in `53 53` for the human transcription, and `53 44` for the incorrect automatic transcription. The next step is to convert the tones into single units, to avoid counting the start and end points of the falling contour tone (e.g. 5,3) as separate errors. When we do this, the transcriptions could take the form `F F` for the human transcription, and `F H` for the erroneous automated transcription. It is at this point that we can calculate the distance between the human and automated transcriptions, using the standard CER and WER algorithms. The *tonal CER* is the percentage of characters in this transcription that are wrong. The *tonal WER* is the percentage of words that have a tonal error in them. These tonal WER and CER will be reported for the transcriptions that do have tone (i.e. IPA, Pinyin, Simple).

3 Results

3.1 ASR Training

First, we performed a simple comparison of the base models to determine which had the best performance **without** the use of an LM. We restricted this test to the IPA transcription. When we compare the three base models (Wav2Vec2, MMS and Whisper), **Wav2Vec2** had the lowest character error (CER=18.3±1.1), compared to Whisper (CER=19.3±1.8) and MMS (CER=25.1±0.7), but Whisper has the lowest word error rate (WER=33.6±2.0), com-

[4] Detailed hyperparameters can be found in Appendix B.

[5] This paper reports the averages for 20 sets of Wav2Vec2 regular models (IPA, Pinyin, Simple), 20 sets of the Wav2Vec2 no-tone models (Pinyin, Simple), 20 MMS IPA models, and 5 Whisper IPA models. Each Wav2Vec2 model took approx. 93 mins to train and test; each MMS model took approx. 105 mins, and the Whisper models approx. 8.5 hrs. The results reported here needed a total of 155 hrs of an Nvidia A100 Tensor 80GB PCIe GPU and 4 CPU cores in an HPC environment (W2V2), as well as 78 hrs of an Nvidia L4 GPU with 8 CPUs in a cloud-based environment (MMS+Whisper).

pared to Wav2Vec2 (WER=47.5±2.7) and MMS (WER=69.5±2.8). Figures 1 and 2 show the summary of the results for the Wav2Vec2 models, figures 3 and 4 show a comparison between Wav2Vec2 and MMS, two base models which have the same architecture, but which differ on the number of languages used during the training phase. Based on these figures, we will answer questions about the interaction of the transcription style, the LM, and the presence of tones in the transcription.

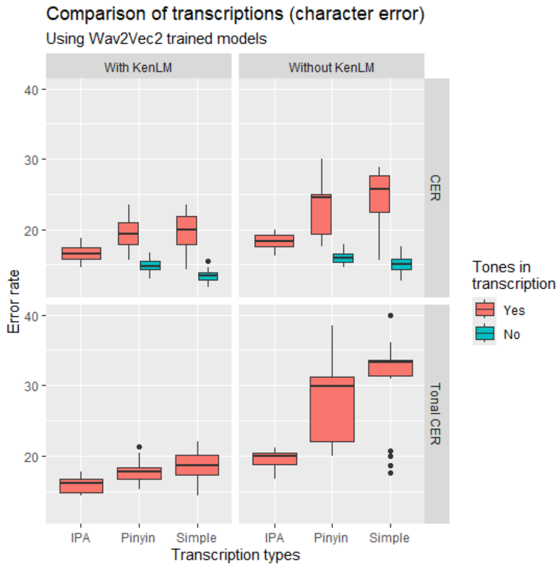

Figure 1: Character and tonal character error for models trained with Wav2Vec2.

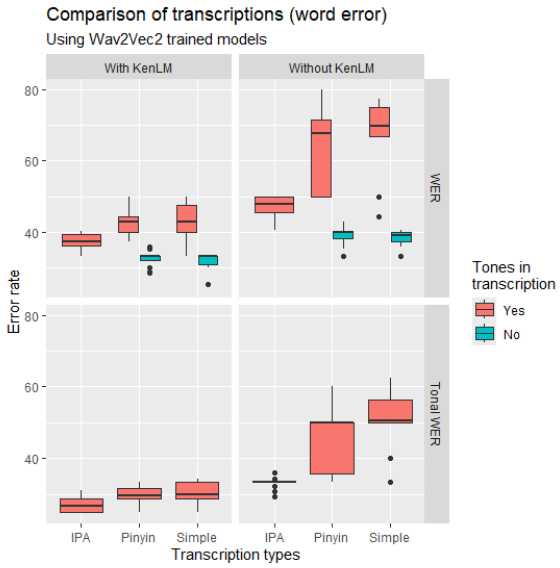

Figure 2: Word and tonal word error for models trained with Wav2Vec2.

Our next question was: **Does using an LM make a difference to the transcription?** The answer is **yes**; using a KenLM-style LM decreases the error rate. In this section we used paired Wilcoxon signed rank tests or paired t-tests to test significance, depending on whether the distributions met the assumption of normality or not, as determined by a Shapiro-Wilk test. The use of an LM reduces CER by 2.6±1.9 points (V=5050, p<0.0001) and WER by 13.9±8.8 points (V=5050, p<0.0001). The use of an LM also reduces tonal error: The tonal CER goes down by 8.8±5.2 points (V=1830, p<0.0001), and the tonal WER goes down by 14.1±8.1 points (V=1830, p<0.0001).

The third question is: **Does the amount of languages in the base model make a difference?** The answer is **yes**, but adding more languages does not seem to lead to an improvement in performance. Since Wav2Vec2 and MMS are based on the same architecture, but trained on a different number of languages (53 for Wav2Vec2 and 1162 for MMS), we decided to test this question. The answer was the opposite of what could be expected. The smaller model, Wav2Vec2, performed better. Its CER was lower by 5.5±1.6 points (V=820, p<0.0001), and its WER was lower by 15.3±7.4 points (V=820, p<0.0001). Wav2Vec2 also had a lower tonal error: the tonal CER was lower by 9.5±6.2 points (V=820, p<0.0001), and the tonal WER was lower by 14.8±8.0 points (V=820, p<0.0001). It seems that the additional languages in MMS did not aid in the transcription of Baima. Therefore, from this point on we will restrict the following tests to the Wav2Vec2 models.

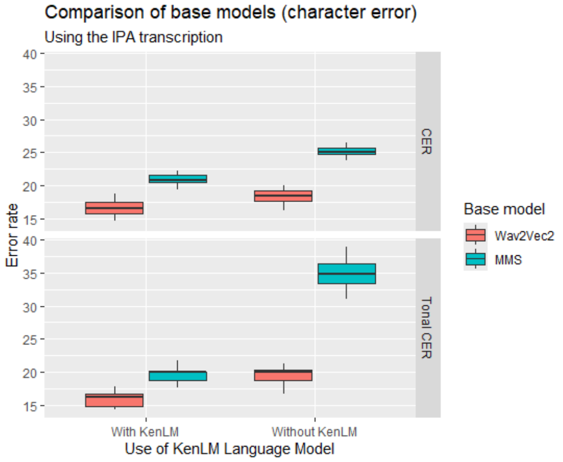

Figure 3: Comparison of Wav2Vec2 and MMS models for character and tonal character error.

The fourth question is: **Does adding tones**

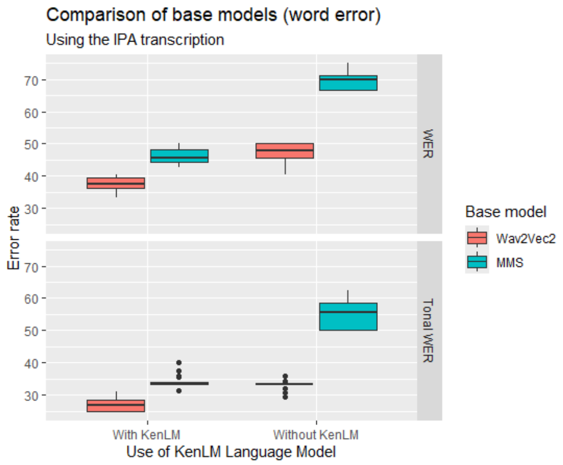

Figure 4: Comparison of Wav2Vec2 and MMS models for word and tonal word error.

to the transcription make transcribing Baima harder? Yes; the tones do take a toll on the transcription. When we studied the difference between the “tone” and “NoTone” versions of Pinyin and Simple romanisations, using tone increases the error rate. It increases CER by 6.9±3.6 points (t(79)=17, p<0.0001) and WER by 18.8±11.4 points (V=3240, p<0.0001).

Finally, **does the transcription style make a difference in the error rates? Yes**, the Pinyin transcription style leads to more errors overall, as well as more tonal errors, even if the tones themselves are the same in all transcription styles. Table 5 shows the (total) CER and WER, as well as the tonal CER and WER. (The consonants and vowels will be discussed in section 3.3). To test this we used the Kruskall-Wallis rank sum test to compare the error means between the three types of transcriptions (IPA, Pinyin, Simple). When the KenLM model is NOT used, there is a significant difference between transcriptions for the four metrics used. For example, IPA has CER=18, Pinyin CER=23 and Simple CER=24 (χ^2(2)=19, p<0.0001). The difference is more pronounced for the word error; IPA has WER=48, Pinyin WER=65 and Simple WER=67 (χ^2(2)=30, p<0.0001). This difference is attenuated by the use of the KenLM, but IPA still performs significantly better. As for the character error, IPA has CER=17, Pinyin CER=19 and Simple CER=20 (χ^2(2)=19, p<0.0001). As for the word error, IPA has WER=37 and both Pinyin and Simple have WER=43 (χ^2(2)=22, p<0.0001).

When we study the tonal errors without an LM, IPA is again the transcription with the least error. For tonal CER, IPA has tonal CER=20, compared to Pinyin tonal CER=28 and Simple tonal CER=31 (χ^2(2)=28, p<0.0001). For tonal WER, IPA has tonal WER=33, compared to Pinyin tonal WER=46 and Simple tonal WER=50 (χ^2(2)=32, p<0.0001). These differences are, again, attenuated by the use of an LM. In the case of character error, IPA has tonal CER=16, compared to Pinyin tonal CER=18 and Simple tonal CER=19 (χ^2(2)=23, p<0.0001). For tonal WER, IPA has tonal WER=27, compared to Pinyin and Simple tonal WER=30 (χ^2(2)=19, p<0.0001).

3.2 Tonal Errors

An additional question for this paper is: **Are there any tones that perform worse than others?** Table 4 shows the percentage of error for specific tones. It shows the average (across 20 models) of the percentage of all the occurrences of a certain tone (e.g. 53) that were predicted erroneously (e.g. 11% for 53, for Wav2Vec2+KenLM using IPA).

In order to understand the patterns in the table, we used an ANOVA test with the percentage of error as the dependent variable, and four independent variables and their interactions: (i) tone {53, 44, 213, 31, 35 and 55}, (ii) transcription style {IPA, Pinyin, Simple}, (iii) use or not of a KenLM LM, and (iv) base model (Wav2Vec2 vs MMS). There is a significant three-way interaction between tone, transcription and base model (F(5,912)=11.9, p<0.0001). In general the Baima and Sandhi tones have less error than the borrowed tones. The use of an LM decreases the error. On average, tones have an error of 47±32% when using KenLM, compared to 55±30% without it. As for the type of model, MMS transcriptions have more errors in general, but this depends on the tone: Wav2Vec2 and MMS have almost identical error rates for tone 53 (both of them 16%), but they have very different error rates for tone 44 (57% versus 40%).

Perhaps the most relevant for the present is the interaction between tone and transcription. There are tones that have much lower error rates than others. As can be seen in table 4, the dipping tone 213 has a lower error rate in the IPA transcription (24% for Wav2Vec2 with KenLM) than in the Pinyin and Simple transcriptions (29% and 30% respectively). This pattern is different from the falling tone 53, which has almost identical error rates across all transcriptions (approx. 12% when using KenLM). Tone 44 also shows large differences between transcriptions, whereas the tones in borrowed words,

			Baima tones			**Sandhi tone**	**Borrowed tones**	
	Model	**Transcription**	**213**	**53**	**44**	**31**	**35**	**55**
With LM	MMS	IPA	34	11	46	29	97	99
	Wav2Vec2	IPA	24	11	30	21	75	78
		Pinyin	29	12	36	26	89	96
		Simple	30	13	38	28	91	97
Without LM	MMS	IPA	59	21	68	36	100	100
	Wav2Vec2	IPA	28	14	34	23	80	83
		Pinyin	34	21	50	35	94	98
		Simple	40	26	55	36	96	100

Table 4: Percentage of errors for each tone by transcription, base model, and use of a KenLM LM. The underlined number is the largest error amongst the different transcription models per tone.

35 and 55, are almost identical (i.e. equally poor) for all transcription styles. These patterns will be further discussed in section 4.

3.3 Tonal versus consonant and vowel errors

We ask one final question which is important for anyone working in the documentation of a tonal language: **Are tones more difficult to transcribe than other parts of the phonology, like the consonants and the vowels? They are**, but mainly when an LM is NOT used. Table 5 shows the CER and WER when only the tones, consonants and vowels were considered. Using a technique similar to that described in section 2.4, we made versions of the transcriptions that had only the consonants and the vowels. For example, [ɲə^{53} te^{53}] 'that man' would be *ɲ t* in the IPA consonant transcription, and *ə e* in the IPA vowel transcription.

We used two ANOVA models, one for the character errors, and one for the word errors. Each of these had the percentage of error as the dependent variable, and three independent variables: transcription (IPA, Pinyin, Simple; all of them with tones), type of phone (Tone, Consonant, Vowel) and use or not of a KenLM LM. The CER model had a significant three-way interaction (F(4,342)=3.6, $p<0.01$), and the WER model had significant two-way interactions.

In the case of the LM, the CER shows a pattern where the use of a KenLM LM reduces the error, but it reduces it more for tones than for the other segments. This is also true for the WER (F(2,342)=5.2, $p<0.01$), where tones improved an average of 9 points, but consonants and vowels only improve by an average of 3 points.

In the case of the transcriptions, the use of KenLM led to a bigger improvement in the Pinyin and Simple transcriptions. This pattern is also true for the WER; where the Pinyin KenLM transcriptions improve by an average of 14 points and the Simple improve by an average of 15.7 points, compared to 5.3 for (F(2,342)=32, $p<0.0001$).

The main difference between CER and WER is in the way they interact with the transcriptions. The tones always have a larger CER when an LM is not used, and they always have amongst the highest CERs even if an LM is used. However, in the case of the WER, the tones are always the worst performers when an LM is absent, but the consonants and vowels behave slightly worse than the tones when an LM is present (F(4,342)=3.0, $p<0.05$).

4 Qualitative Error Analysis

In this section we shift our focus to the specific errors that the models make when transcribing, and how those might affect linguistic work.

4.1 Specific errors

Table 6 provides specific examples of transcription output. Further examples, including for the contrast between transcription systems are available in the Appendix. Examples (1) and (2) show the difference between the base models (without LM) and Wav2Vec2 with and without LM. It is clear that without an LM both base models, but especially MMS, struggle to get the right word boundaries for words that are acoustically merged together, like the copula [re^{213}] and the following question marker [a]. The target transcriptions actually give the original (lexical) tones of the two morphemes, whereas the models provide the actual pronunciation: a fused syllable with the overlaid interrogative intonation, which is closer to actual acoustic signal. Both models also appear to make errors at the end of the segment in (1). The acous-

		CER				WER			
	Transcription	**Total**	**Tone**	**Cons**	**Vowel**	**Total**	**Tone**	**Cons**	**Vowel**
With LM	IPA	17	16	13	16	37	27	25	28
	Pinyin	19	18	18	17	43	30	32	32
	Simple	20	19	17	19	43	30	34	31
Without LM	IPA	18	20	15	18	48	33	31	32
	Pinyin	23	28	22	21	65	46	45	45
	Simple	24	31	21	23	67	50	48	44

Table 5: Error for each type of character (tones, Cons=consonants, vowels) for Wav2Vec2 models. The underlined number is the largest error amongst the three types of characters.

	Different base models with IPA transcription (Without LM)		
1. SPX-bqh-018-193	“[He] asked (literally: said): "Is it the herdsman’s horse or this young wanderer’s horse."”		
Target transcription	ta^{53} ndzʊ^{213} s^{h}e^{31}pu^{53} ta^{53} re^{213} a tɕ^{h}o^{31}mba^{53} ɡo^{31}dʑy^{53} ta^{53} re^{213} te^{53} dzɛ^{213} ʃə	CER	WER
MMS prediction	ta^{53} ndzo^{53} se^{31}pu^{53} ta^{53} ra^{3} tɕ^{h}o^{31}mba^{53} ŋgo^{31}ʑy^{53} ta^{53} re^{2} ə	27	67
Wav2Vec2 prediction	ta^{53} ndzʊ^{213} s^{h}e^{31}pu^{53} ta^{53} re^{213} tɕ^{h}o^{31}mba^{53} ŋgo^{31}dʑy^{53} ta^{53} re^{213}z^{2}	14	42
	Wav2Vec2 IPA transcription Without vs With LM		
2. SPX-bqh-020-053	“When the two of them were hunting, [they accidentally] fired an arrow into a tree, and that tree turned into a young man [= a tree brother appeared], then they... ”		
Target transcription	ɲʕi^{53} ɲge^{53} nde^{53} sə õ^{213} ʃ^{h}e^{213} ke^{53} nda^{53} dʑʉ^{53} ɕɛ^{44} ʃ^{h}e^{213} ɲa^{31}ɲu^{53} ly^{213} ue^{44} ɲi to^{44} ʧo^{31}rə^{53}	CER	WER
Without LM prediction	ɲʕi^{53} ɲge^{53} nde^{53} ʃə õ^{213} ʃ^{h}e^{213} ke^{53} nda^{53} dʑʉ^{53} sə ʃ^{h}e^{213} ɲa^{31}ɲu^{53} ly^{213} ue^{44} ɲi to^{44} ʧo^{31}rʊ^{53}	6	18
With LM prediction	ɲʕi^{53} ɲge^{53} nde^{53} ʃə õ^{213} ʃ^{h}e^{213} ke^{53} nda^{53} dʑʉ^{53} ɕɛ ʃ^{h}e^{213} ɲa^{31}ɲu^{53} ly^{213} ue^{44} ɲi to^{44} ʧo^{31}rə^{53}	4	12
	Perfect CER and WER (even in detailed IPA with tone)		
3. SPX-bqh-011-121	“[You] need to go to my place, so [the emperor] said.”	CER	WER
Target & Prediction	k^{h}ʉ^{53} tsa^{44} ndʑi^{53} ɡo^{53} re^{213} ndʐu^{53} dzɛ^{213} ʃə	0	0
	Bad CER/WER most challenging IPA and ‘easiest’ Simple NoTone transcriptions		
4. SPX-bqh-002-277	“The big sister looked around, looked up, looked sideways, [then she] returned home, shook her head and said, there’s nothing there.”		
IPA			
Target transcription	pu^{44} tʃ^{h}e^{213} ŋgo^{31}kɛ^{31} tʂa^{53} tyʉ^{44} mbo tɕe^{53} tyʉ^{44} ndʐe^{44} tyʉ^{44} ɕi^{53} tse^{53} a^{31} ã^{53} tʃo^{53} mu^{31}=no^{213}	CER	WER
With LM prediction	pu^{44} tʃ^{h}e^{213} te^{53} ŋgo^{213} ke^{31}ʧɑ^{53} te^{53} kumbo tɕɛ^{213} te^{53} ndʐe^{53} ɕi^{53} ɑ^{213}ɑ^{213} tʃo^{53} mu^{31} no^{213} ndʐu^{53} dzɛ^{213} ʃə	56	87
Simple NoTone			
Target transcription	pu tsyhe nggookëtra tyue mboo tsyë tyue ndrqe tyue syi tse aã tsyoo mu noo	CER	WER
With LM prediction	pu tsyhë nyi ngoo ketsya te khu mboo tsyë te ndu aa tsyoo mu noo ndrqu dzë syə	58	93
Without LM prediction	pu tsyhë nyə ngoo ketsya te khumboo tsyë te nduë i aa tsyoo mu noo ndrqu dzë syə	58	93

Table 6: ASR results from various experiments for Baima - Part I: Base and Language Models

tic signal is actually deprecated here, showing the real benefit of adding an LM that can add words in often-seen contexts even if they are barely audible in the recordings. Finally, the MMS base model in particular seems to struggle with clusters at the start of syllable like [ʃ], [dz] and [dʑ].

Examples (3) and (4) illustrate that the models have outliers too, yielding both very good examples (in 3) or seemingly very bad examples (in 4), judging by the error rates. While the recording for (3) is rather short and clear, the articulation of the speaker uttering (4) is much less clear. The suggestion of the model to transcribe a particularly unclear part of the segment as [te^{53} kumbo] is actually probably more plausible than what the original transcriber first proposed. Furthermore, the final part of the utterance is completely ‘swallowed’ in the recording, but the model still proposes a very good transcription for those final words. Overall, zooming in on specific errors shows that even when results look very bad when simply calculating the CER and WER, in reality the models may actually be more useful than originally thought.

4.2 Tonal error analysis

Generally, the models for transcription types without the tones perform better. This could be due to the fact that it is genuinely ‘easier’ to ignore suprasegmental features like tones, and because the Chao tone numbers simply add further characters to the target inventory, especially when they are counted as separate characters. When it comes to tonal errors we can make one clear observation from these qualitative data: some errors are due to the fact that transcriptions only note etymological tones and disregard sentence-level stress, that is, distinctive pitch contours that serve to mark words 'in focus' position and overlay the etymological tone of the word in focus.

As for specific tones, out of the six different options the 53 tone is the easiest to recognise, probably due to its high frequency, whereas the high tone 55, which only occurs on a handful of Chinese borrowings proves the most challenging.[6]

4.3 General errors

In general, based on the examples above, the main reasons for discrepancies between transcriptions and predictions are easily explained. For example, weakening in unstressed position can lead to the models predicting schwas, which is no doubt a frequent occurrence in any base model. Mainly, however, we note that all models suffer significantly from bad quality of the recording: background noise, unclear articulation, etc. lead to an increase in both CER and WER. However, when these increases are there because of incomplete or inexact original transcriptions, we also see that the models (especially those enriched with a Baima-specific KenLM LM) can actually yield transcriptions that are even better than the original.

5 Conclusion

In this paper we tested tonal accuracy and the effect of transcription type, base model as well as the option of adding a KenLM LM to the ASR pipeline of the Baima language, which has phonological features, including six tones, and is extremely limited in resources.

First, we found that more languages in a similar architecture for the base model (i.e. MMS vs Wav2Vec2) does not lead to better outcomes when transcribing smaller languages, perhaps because the extra languages are not phonologically similar to Baima. Wav2Vec2 has 5 tonal languages (Mandarin Chinese, Hakka, Cantonese, Lao and Zulu). MMS has these, plus many others, including small Indigenous languages with a wealth of tones. However, maybe the specific typology of the tonal system in Baima (where tones are consistently produced with both a particular f0 specification and a voice quality specification) poses a problem for the model. We furthermore showed that complex tones remain the most difficult part of the phonology to transcribe, despite the complexity of Baima vowel phonology. However, adding an LM to the decoding process can help to mitigate this problem.

Overall, non-tonal romanised transcriptions trained with a Wav2Vec2 base model and enhanced with a KenLM LM show the best results, but even detailed IPA models with Chao-numbered tones perform reasonably well, considering the very small amount of input data (186 mins). While it remains essential to reliably document and describe the sound system of the language using detailed IPA, it may at times be preferable to use a simplified romanisation system to speed up transcription of larger speech samples. Pinyin results are generally worse than both detailed IPA and Simple romanisation, but it would be naturally easier to learn for speakers as they are familiar with this type of transcription system thus facilitating language preservation. While conversion from Simple romanisation or Baima Pinyin to IPA is impossible as too many details are lost, it is possible to convert into Pinyin and Simple romanised script from the better-performing IPA model, making the latter potentially the most useful, not just for phoneticians, but also the local community.

To conclude, the way the language is transcribed can affect tonal outputs, even when the tonal markings themselves remain the same throughout different transcriptions. This underlines the difficulties in using deep-learning based technology, where the various orthographies produce opaque but significant differences in how the system outputs tone.

Ethics Statement

Ethics approval was obtained prior to data collection from the Research Ethics Office of CNRS.

Acknowledgements

We would like to thank the Baima speakers. We also gratefully acknowledge funding from the ANR (PhoTon, ANR-23-CE54-0003) and the ERC (PaganTibet Advanced Grant 101097364).

[6] Overall frequencies can vary slightly due to the different splits in training/validation/test data, but to give an impression, in the test set #3 the frequencies are (in descending order): Tone 53 - n=1626 (53%), Tone 31 - n=551 (18%), Tone 213 - n=495 (16%), Tone 44 - n=363 (12%), Tone 35 - n=22 (0.7%) and Tone 55 - n=15 (0.5%).

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)*.

Timofey Arkhangelskiy. 2021. Low-resource asr with an augmented language model. In *Proceedings of the Seventh International Workshop on Computational Linguistics of Uralic Languages*, pages 40–46.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*.

Yuen Ren Chao. 1930. ə sɪstəm əv “toun-lɛtəz”. *Le Maître Phonetique*, 30(1):24–27.

Katia Chirkova. 2017. 14 evidentials in pingwu baima. *Evidential systems of Tibetan languages*, 302:445.

Katia Chirkova. 2025. Pitch, vowel duration, and phonation in baima and neighboring languages. *Language and Linguistics*, 26.2.

Katia Chirkova and Zhengkang Han. 2016. *Shiyong Duoxuyu Yufa 实用多续语语法 [Practical Grammar of Duoxu]*. Beijing: Minzu Chubanshe.

Katia Chirkova, Tanja Kocjančič Antolík, and Angélique Amelot. 2023. Baima. *Journal of the International Phonetic Association*, 53(2):547–576.

Katia Chirkova and Dehe Wang. 2017. Binwei yuyan diancang yu ersuyu pinyin fang’an 濒危语言典藏与尔苏语拼音方案 [endangered languages documentation and ersu romanization system]. *Xinan Renmin Daxue Xuebao 西南民族大学学报 [Journal of Southwest University for Nationalities]*, pages 69–75.

Chinese Script Reform Committee. 1956. Hanyu pinyin fang’an 汉语拼音方案 [scheme for the chinese phonetic alphabet].

Rolando Coto-Solano. 2021. Explicit tone transcription improves ASR performance in extremely low-resource languages: A Case Study in Bribri. Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas (pp. 173–184). Association for Computational Linguistics.

Rolando Coto-Solano, Sally Akevai Nicholas, Samiha Datta, Victoria Quint, Piripi Wills, Emma Ngakuravaru Powell, and Isaac Feldman. 2022. Development of Automatic Speech Recognition for the Documentation of Cook Islands Māori. Proceedings of the Thirteenth Language Resources and Evaluation Conference (pp. 3872-3882). https://aclanthology.org/2022.lrec-1.412.

Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the sixth workshop on statistical machine translation*, pages 187–197.

Nils Hjortnaes, Timofey Arkhangelskiy, Niko Partanen, Michael Rießler, and Francis M Tyers. 2020. Improving the language model for low-resource asr with online text corpora. In *Proceedings of the 1st joint SLTU and CCURL workshop (SLTU-CCURL 2020)*. European Language Resources Association (ELRA).

Bufan Huang and Minghui Zhang. 1995. Baimahua zhishu wenti yanjiu [a study of the genetic affiliation of baima]. *Tibetology in China*, 1995:79–118.

Jesin James, Deepa P Gopinath, et al. 2024. Advocating character error rate for multilingual asr evaluation. *arXiv preprint arXiv:2410.07400*.

Tan Lee, Wai Lau, Y. W. Wong, and P. C. Ching. 2002. Using tone information in cantonese continuous speech recognition. *ACM Transactions on Asian Language Information Processing (TALIP)*, 1(1):83–102.

Linying Ma, Dennis Elton Walters, and Susan Gary Walters. 2008. *Nuosu Yi–Chinese–English glossary*. Beijing: Minzu Chubanshe.

Marieke Meelen, Alexander O’Neill, and Rolando Coto-Solano. 2024. End-to-end speech recognition for endangered languages of nepal. In *Proceedings of the Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 83–93.

Vikramjit Mitra, Horacio Franco, Martin Graciarena, and Arindam Mandal. 2012. Normalized amplitude modulation features for large vocabulary noise-robust speech recognition. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4117–4120. IEEE.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.

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.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Mihaela C Stoian, Sameer Bansal, and Sharon Goldwater. 2020. Analyzing asr pretraining for low-resource speech-to-text translation. In *ICASSP 2020-2020*

IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 7909–7913. IEEE.

Hongkai Sun, Katia Chirkova, and Guangkun Liu. 2007. *Baimayu Yanjiu* 《白马语研究》. Beijing: Nationalities Press 民族出版社.

Chihiro Taguchi and David Chiang. 2024. Language complexity and speech recognition accuracy: Orthographic complexity hurts, phonological complexity doesn't. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15493–15503, Bangkok, Thailand. Association for Computational Linguistics.

Dehe Wang, Ke Wang, Xuan Wang, Katia Chirkova, and Tao Gu. 2019. *Ersu-Chinese Dictionary* 尔苏语词汇通释. Hefei 合肥: Anhui Publishing House 安徽出版社.

A Appendix: More Sample Outputs

This appendix provides additional transcription examples to enable full comparison between different transcription systems. It is clear that all five options struggle with the same Baima words and the same phonemes, namely the first vowel in [ʧ^{h}u^{31}jo^{213}] and the onset of [wo^{44}]. The vowel [u] is in an unstressed position here, which may explain why all models predict a schwa (or similar). Similarly, all of the converted models (i.e. all apart from the original IPA transcription) appear to struggle with onset glides [j-] vs [w-] or zero. The WER in all models apart from the Simple NoTone version is mainly higher because of the failure to recognise [ʧ^{h}u^{31}jo^{213}] as one word. Overall, WER is very similar for all transcription forms, which provides additional support for the importance of reporting both CER and WER, especially when it comes to ASR for extremely low-resource and highly-endangered languages (James et al., 2024).

B Appendix: Hyperparameters

The following are the hyperparameters for the Wav2Vec2 training, using the `wav2vec2-large-xlsr-53` base model:

1. `attention_dropout = 0.1`
2. `hidden_dropout = 0.1`
3. `feat_proj_dropout = 0.0`
4. `mask_time_prob = 0.05`
5. `layerdrop = 0.1`
6. `gradient_checkpointing = true`
7. `ctc_loss_reduction = mean`
8. `per_device_train_batch = 8`
9. `gradient_accumulation_steps = 2`
10. `evaluation_strategy = steps`
11. `num_train_epochs = 29` (4000 steps)
12. `fp16 = true`
13. `save_steps = 400`
14. `eval_steps = 100`
15. `learning_rate = 3e-4`
16. `warmup_steps = 500`
17. `kenlm_ngrams = 4`

The following are the hyperparameters for the MMS training, using the `mms-1b-all` model:

1. `attention_dropout = 0.0`
2. `hidden_dropout = 0.0`
3. `feat_proj_dropout = 0.0`
4. `ctc_loss_reduction = mean`
5. `per_device_train_batch = 2`
6. `evaluation_strategy = steps`
7. `num_train_epochs = 4` (4872 steps)
8. `gradient_checkpointing = true`
9. `fp16 = true`
10. `save_steps = 400`
11. `eval_steps = 100`
12. `learning_rate = 1e-3`
13. `warmup_steps = 100`
14. `kenlm_ngrams = 4`

The following are the hyperparameters for the Whisper training, using the `whisper-medium Multilingual` model:

1. `per_device_train_batch_size = 2`
2. `per_device_eval_batch_size = 1`
3. `gradient_accumulation_steps = 1`
4. `learning_rate = 1e-5`
5. `warmup_steps = 500`
6. `max_steps = 4001`
7. `gradient_checkpointing = true`
8. `evaluation_strategy = steps`
9. `predict_with_generate = true`
10. `generation_max_length = 225`
11. `fp16 = true`
12. `metric_for_best_model = wer`
13. `greater_is_better = false`

	Five different transcription systems (withoutLM, W2v2)		
1. SPX-bqh-018-453	“I have a buffalo hide soaked in water [if you can tan the hide...].”		
IPA			
Target transcription	k^{h}ʉ^{53} la^{53} ʧ^{h}u^{31}jo^{213} ʃu^{31}mba^{53} wo^{44} ʒɑ^{53}ʐu^{53}	CER	WER
Prediction	k^{h}ʉ^{53} la^{53} ʧ^{h}ə^{53} jo^{213} ʃu^{31}mba^{53} wo^{213} rʒɑ^{53}ʐu^{53}	15	57
Pinyin			
Target transcription	gue^{53} la^{53} chu^{31}yoo^{213} syu^{31}nbba^{53} woo^{44} ssha^{53} xxu^{53}		
Prediction	gue^{53} lu^{31}ei^{53} chii^{53}oo^{213} syu^{31}nbba^{53} oo^{213} zzei^{213} xxu^{53}	36	57
Pinyin NoTone			
Target transcription	gue la chuyoo syunbba woo ssha xxu		
Prediction	gue la chii yoo syunbba oo ra xxu	21	57
Simple			
Target transcription	khue^{53} la^{53} tsyhu^{31}yoo^{213} syu^{31}mba^{53} woo^{44} zya^{53} zyu^{53}		
Prediction	khue^{53} la^{53} tsyhə^{31}yoo^{413} shu^{31}mba^{53} oo^{213} zyu^{53}	24	57
Simple NoTone			
Target transcription	khue la tsyhuyoo syumba woo zya zyu		
Prediction	khue la tsyhəyoo syumba oo dzya zyu	9	43

Table 7: ASR results from various experiments for Baima - Part II: Transcription systems