Explicit Tone Transcription Improves ASR Performance in Extremely Low-Resource Languages: A Case Study in Bribri

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

Abstract

Linguistic tone is transcribed for input into ASR systems in numerous ways. This paper shows a systematic test of several transcription styles, using as an example the Chibchan language Bribri, an extremely low-resource language from Costa Rica. The most successful models separate the tone from the vowel, so that the ASR algorithms learn tone patterns independently. These models showed improvements ranging from 4% to 25% in character error rate (CER), and between 3% and 23% in word error rate (WER). This is true for both traditional GMM/HMM and end-to-end CTC algorithms. This paper also presents the first attempt to train ASR models for Bribri. The best performing models had a CER of 33% and a WER of 50%. Despite the disadvantage of using hand-engineered representations, these models were trained on only 68 minutes of data, and therefore show the potential of ASR to generate further training materials and aid in the documentation and revitalization of the language.

Resumen

Transcribir el tono de forma explícita mejora el rendimiento del reconocimiento de voz en idiomas extremadamente bajos en recursos: Un estudio de caso en bribri. Hay numerosas maneras de transcribir el tono lingüístico a la hora de proveer los datos de entrenamiento a los sistemas de reconocimiento de voz. Este artículo presenta un experimento sistemático de varias formas de transcripción usando como ejemplo la lengua chibcha bribri, una lengua de Costa Rica extremadamente baja en recursos. Los modelos más exitosos fueron aquellos en que el tono aparece separado de la vocal de tal forma que los algoritmos pudieran aprender los patrones tonales por separado. Estos modelos mostraron mejoras de entre 4% y 26% en el error de caracteres (CER), y de entre 3% y 25% en el error de palabras (WER). Esto se observó tanto en los algoritmos GMM/HMM como en los algoritmos CTC de secuencia-a-secuencia. Este artículo también presenta el primer intento de entrenar modelos de reconocimiento de voz en bribri. Los mejores modelos tuvieron un CER de 33% y un WER de 50%. A pesar de la desventaja de usar representaciones diseñadas a mano, estos modelos se entrenaron con solo 68 minutos de datos y muestran el potencial para generar más materiales de entrenamiento, así como de ayudar con la documentación y revitalización de la lengua.

1 Introduction

The documentation and revitalization of Indigenous languages relies on the transcription of speech recordings, which contain vital information about a community and its culture. However, the transcription of these recordings constitutes a major bottleneck in the process of making this information usable for researchers and practitioners. It typically takes up to 50 hours of an expert's time to transcribe each hour of audio in an Indigenous language (Shi et al., 2021). Moreover, there are usually few community members who have the expertise to transcribe this data and who have the time to do so. Because of this, extending automated speech recognition (ASR) to these languages and incorporating it into their documentation and revitalization workflows would alleviate the workload of linguists and community members and help accelerate their efforts.

Indigenous and other minority languages usually have few transcribed audio recordings, and so adapting data-hungry ASR algorithms to assist in their documentation is an active area of research (Besacier et al., 2014; Jimerson and Prud'hommeaux, 2018; Michaud et al., 2019; Adams et al., 2019; Foley et al., 2018; Gupta and Boulianne, 2020b,a; Zahrer et al., 2020; Thai et al., 2019; Li et al., 2020; Partanen et al., 2020; Zevallos et al., 2019; Matsuura et al., 2020; Levow et al., 2021). This paper will examine an element that

Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas, pages 173–184
June 11, 2021. ©2021 Association for Computational Linguistics

might appear obvious at first, but one where the literature is “inconclusive" (Adams, 2018), and which can have major consequences in performance: How should tones be transcribed when dealing with extremely low-resource languages? This will be examined by building ASR models for the language Bribri from Costa Rica. The results show that simple changes in the orthographic transcription, in the form of explicit tonal markings that are separate from the vowel information, can dramatically improve accuracy.

1.1 Tonal languages and ASR

A tonal language is a language where differences in pitch can change the meaning of a word, even if the consonants and vowels are the same (Yip, 2002). The best-known example of a tonal language is Mandarin Chinese. In Mandarin, the syllable [ma] means “mother" if it is produced with a high pitch. The same syllable means “horse" when pronounced with a dipping-rising pitch, but if it is pronounced with a falling pitch, it means “to scold”. Between 40% and 70% of the languages of the world are tonal (Yip, 2002; Maddieson, 2013), including numerous Indigenous languages of the Americas. Because tone is expressed as pitch variations, and those variations can only occur during the pronunciation of consonants and vowels, tonal cues overlap with those of the consonants and vowels in the word. Therefore, it is useful to distinguish between segments - consonants and vowels - and the information that is *suprasegmental*, such as tone, which occurs co-temporally with segments (Lehiste and Lass, 1976).

Precisely because of large tonal languages like Mandarin, there has been research into how tone can play a role in ASR. Many systems treat pitch (the main phonetic cue of tone) as a completely separate feature. In such systems, the traditional ASR algorithm learns the segments, and a separate machine learning module learns the pitch patterns and offers its inference of the tone (Kaur et al., 2020). This has been used for languages like Mandarin (Niu et al., 2013; Shan et al., 2010), Thai (Kertkeidkachorn et al., 2014) and Yoruba (Ọdélọbí, 2008; Yusof et al., 2013). On the other hand, there is research that suggests that, given that the tone and vowel information are co-temporal, these are best learned together. For example, an ASR system would be asked to learn a vowel and its tone as a single unit (e.g. a+highTone). Fusing the representation for vowel and tone, or *embedded tone modeling* (Lee et al., 2002), has been shown to be effective for larger languages like Mandarin (Chang et al., 2000), Vietnamese and Cantonese (Metze et al., 2013; Nguyen et al., 2018), as well as smaller languages like Yoloxóchitl Mixtec from Mexico (Shi et al., 2021) and Anyi from Côte d'Ivoire (Koffi, 2020). Finally, in some tonal languages like Hausa, in which the orthography does not mark any tone, the tone is not included at all in ASR models (Gauthier et al., 2016).

Representations where the tone is marked explicitly but is kept separate from the vowel (i.e. *explicit tone recognition* (Lee et al., 2002)) are not often used for larger languages, but they are very common in low-resource ASR. This is often done using phonetic representations, where the output of the algorithm is in the form of the International Phonetic Alphabet (IPA), which is then converted to the language's orthographic convention. For languages like Na from China and Chatino from Mexico (Ćavar et al., 2016; Adams et al., 2018), the characters representing the tone are separated from the vowel. Wisniewski et al. (2020) argue that it is the transparency of the representation (either orthographic or phonetic) that helps ASR to learn these tonal representations, and this transparency includes having characters that the algorithm can use to generalize the phonetic cues of the tones separate from those of the vowels.

Given the review above, there appears to be more than one way to represent tone effectively as input for ASR. In this paper several different methods will be tested using a language (and indeed, a language family) in which no ASR models have been trained before.

1.2 Chibchan Languages and Bribri

The Bribri language (Glottocode `brib1243`) is spoken by about 7000 people in Southern Costa Rica (INEC, 2011). It belongs to the Chibchan language family, which includes languages such as Cabécar and Malecu from Costa Rica, Kuna and Naso from Panama, and Kogi from Colombia. Bribri is a vulnerable language (Moseley, 2010; Sánchez Avendaño, 2013). This means that there are still children who speak it with their families but there are few circumstances when it is written, and indeed there are very few books published in the language. Bribri has four tones: high, falling, rising, and low tone. The first three are marked in the orthography using diacritics (respectively: *à*, *á*, *â*), while the low tone is left unmarked: *a*. Bribri tone can create differences in meaning: the word *alà* means 'child'; its first syllable is low and the second syllable is high. Contrast this with *alá* 'thunder', where the second syllable has a falling tone.

Bribri has an additional suprasegmental feature: Nasality. Like in French, vowels in Bribri can be oral or nasal. Therefore, *ù* with an oral vowel means 'house', but *ù̱* with a nasal vowel, marked with a line underneath the vowel,[1] means 'pot'.

Bribri orthographies are relatively transparent due to their recent invention, the oldest of which is from the 1970s (Constenla et al., 2004; Jara Murillo and García Segura, 2013; Margery, 2005). This works to our advantage, in that there is almost no difference between an orthographic and a phonetic representation for the input of Bribri ASR.

There has been some work on Bribri NLP, including the creation of digital dictionaries (Krohn, 2020) and morphological analyzers used for documentation (Flores Solórzano, 2019, 2017b). There have also been some experiments with untrained forced alignment (Coto-Solano and Flores Solórzano, 2016, 2017), and with neural machine translation (Feldman and Coto-Solano, 2020). However, there is a need to accelerate the documentation of Bribri and produce more written materials out of existing recordings, and here we face the bottleneck problem mentioned above. One of the main goals of this paper is to build a first ASR system for Bribri in order to alleviate the problems of transcription.

2 Transcription Methodology

The first step towards training an ASR model in Bribri was the selection of the training materials. The spontaneous speech corpus of Flores Solórzano (2017a) was used because of its public availability (it is available under a Creative Commons license) and because of its consistent transcription. This corpus contains 1571 utterances from 28 speakers (14 male and 14 female), for a total of 68 minutes of transcribed speech. These utterances contain a total of 13586 words, with 2221 unique words.

The main question in this paper is: How can we easily reformat Bribri text into the best possible input for ASR? Let's take the word *dikì̱* /di˩ˈki˥/ 'underneath' as an example. This word has two syllables, the first one with a low tone and the second one with a high tone, indicated by a grave accent. In addition to the tone, the second syllable is also nasal, and this is marked with a line underneath the vowel. One possible representation of this word would be to interpret it as four different characters, as is shown in condition 1 of table 1. Here, the character for the last vowel would carry in it the information that it is the vowel /i/, that the vowel is nasal, and that the vowel is produced with a high tone. This condition will be called AllFeats, or "all features together", because each character in the ASR alphabet carries with it all the suprasegmental features of the vowel. In this transcription, the Bribri ASR alphabet would have 48 separate vowel symbols: `A-HIGH`, `A-HIGH-NAS`, `A-LOW`, `A-LOW-NAS`, etc.

There are many other ways in which the word

Condition	Example transcription	Length	Symbols for vowels + feats
1. **AllFeats**: All features together	`D I-LOW K I-NAS-HIGH`	4	48
2. **NasSep**: Nasal as separate character	`D I-LOW K I-HIGH NAS`	5	28 + 1 = 29
3. **ToneNasSepWL**: Both tone and nasal separate; explicit indication of low tone	`D I LOW K I HIGH NAS`	7	7 + 5 = 12
4. **ToneNasSep**: Both tone and nasal separate; low tone as implicit default	`D I K I HIGH NAS`	6	7 + 4 = 11
5. **ToneSepWL**: Tone is separate; explicit indication of low tone	`D I LOW K I-NAS HIGH`	6	12 + 4 = 16
6. **ToneSep**: Tone is separate; low tone as implicit default	`D I K I-NAS HIGH`	5	12 + 3 = 15

Table 1: Different ways to transcribe the Bribri word *dikì̱* /di˩ˈki˥/ 'underneath'

[1] There are two main orthographic systems for Bribri. In the Constenla et al. (2004) system, the nasal is marked with a line under the vowel. In the Jara Murillo and García Segura (2013) system, the nasal is marked with a tilde over the vowel: *ũ̀* 'house'.

could be transcribed. For example, as shown in the second condition, NasSep, the nasality could be written as a separate character and the tone and vowel could be represented together. In this transcription, the final vowel would be made up of two separate alphabetic symbols: I-HIGH and NAS. This idea of separating features could be taken further, and both the tone and the nasality could be represented as separate characters. This is represented in the third condition, TonesNasSepWL. Here, both the tones and the nasal feature follow the vowel as separate characters, and the final vowel of *dikì̠* 'underneath' would be expressed using three alphabetic symbols: I HIGH NAS. Notice that, in this condition, the low tone of the first syllable would be represented explicitly after the first vowel, I LOW, hence the condition includes the 'WL', "with low [tone]". However, this low tone is the most frequent tone in Bribri, and as a matter of fact it has no explicit diacritic in the Bribri writing system. Because of this, another option for the transcription could be to keep marking the tones and nasals separately from the vowels, but to only represent the three salient tones (high, falling, rising) and leave the low tone as a default, unwritten option in the transcription. This is shown in condition 4, ToneNasSep.

There are some combinations where the nasal marking stays with the vowel, but the tone is separate. In condition 5, ToneSepWL, the tones are indicated separately but the nasality is written jointly with the vowel. The final vowel of *dikì̠* 'underneath' would then be represented using two symbols: I-NAS HIGH. This means that there would be twelve vowel symbols[2] in the Bribri ASR alphabet (e.g. A, A-NAS, E, E-NAS, etc.), and separate indicators for the four tones: HIGH, FALL, RISE, LOW. But, given that the low tone is again the most frequent, we could assume it as a default tone and leave the LOW marking out. This is done in condition 6, ToneSep. In ToneSep, the second vowel has a high tone, and so it gets a separate HIGH tone marker. The first vowel, on the other hand, has a low tone, and therefore gets no marking.

In order to test the different performance of these conditions, two different ASR systems were used. First, the Bribri data was trained using a traditional Gaussian Mixture Models based Hidden Markov Model algorithm (GMM/HMM), implemented in the Kaldi ASR program (Povey et al., 2011). Given the paucity of data, this is likely the best option for training. However, end-to-end systems are also available, and while they are known not to perform well with small datasets (Goodfellow et al., 2016; Glasmachers, 2017), they were still tested to see if the differences in transcription caused any variation in performance. A Connectionist Temporal Classification (CTC) loss algorithm (Graves et al., 2006) with bidirectional recursive neural networks (RNNs) was used, implemented in the DeepSpeech program (Hannun et al., 2014).

3 Traditional ASR Results

Kaldi was used to train models for each of the transcription conditions described above. Two parameters were varied in the experiment: The number of phones in the acoustic model (monophone or triphone), and the number of words in a KenLM based language model (unigrams, bigrams and trigrams) (Heafield, 2011). All other hyperparameters were identical to those in the default Kaldi installation. Thirty models were trained for each of the six transcription conditions, using the six parameter combinations (phones x ngrams), for a total of 1080 models.[3] To train these models utterances were randomly shuffled for every model and then split so that 90% of the utterances were used for training (1571 utterances) and 10% were used for validation (174 utterances). Each of the models had two measures of error: the median character error rate (CER) and the median word error rate (WER), calculated over the input transcription for each condition. The results reported below correspond to the median of the 30 medians in each condition.

Figure 1 shows the summary of the training results. The condition with the best performance is ToneSep, where the tone symbol is kept separate (HIGH, FALL, RISE), the low tone is left out as a default, and the nasal feature remains connected to the vowel symbol (i.e.: A versus A-NAS).

Table 2 shows the summary of results for three conditions: ToneSep and AllFeats, which had the best performance, and ToneNasSepWL, which had the worst performance. The best performing of all conditions is ToneSep trained with triphones and with a trigram language model. This combination of factors produces models with a median of

[2] There are five vowels that can be both oral and nasal: /a, e, i, o, u/. There are two vowels, /ɪ, ʊ/, written 'ë' and 'ö', which can never be nasal.

[3] The models were trained using an Intel i7-10750H CPU, and each took approximately 5 minutes to train, for a total of 90 hours of processing. The electricity came from the ICE electric grid in Costa Rica, which uses 98% renewable energy.

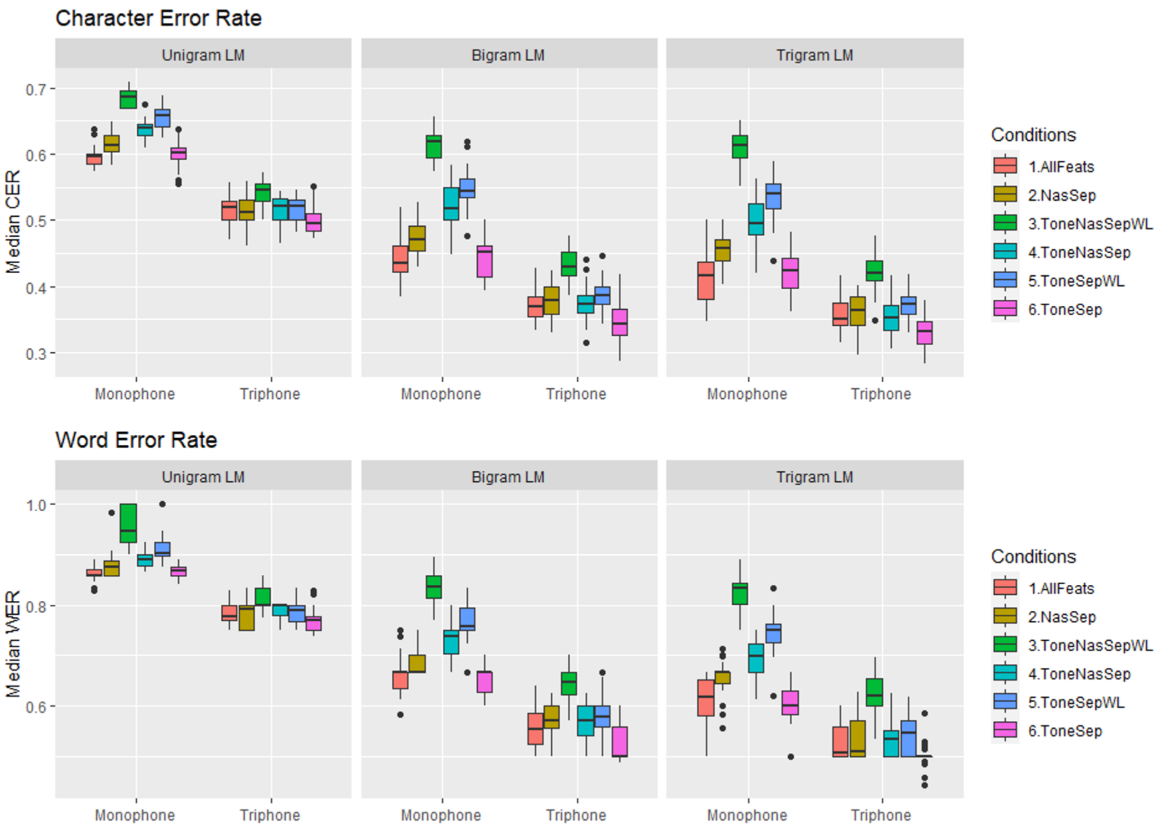

Figure 1: Medians for character error rate (CER) and word error rate (WER) for Kaldi training, using different phone (monophone, triphone) and language models (unigrams, bigrams, trigrams).

	ToneSep	AllFeats	ToneNasSepWL	MaxΔ
CER Mono	60 - 45 - 42	60 - 44 - 42	69 - 62 - 61	9 - 18 - 19
CER Tri	50 - 34 - **33**	52 - 37 - 35	54 - 43 - 42	4 - 9 - 9
WER Mono	87 - 67 - 60	86 - 67 - 62	95 - 84 - 83	9 - 17 - 23
WER Tri	77 - 50 - **50**	78 - 55 - 51	80 - 65 - 62	3 - 15 - 12

Table 2: Median character error rate (CER) and word error rate (WER) for the best conditions (ToneSep and AllFeats) and the worst condition (ToneNasSepWL). The three numbers indicate the error for unigram, bigram and trigram language models. MaxΔ indicates the difference between the worst and the best models.

33% CER and 50% WER. Very close is AllFeats with triphones and trigrams, with 35% CER and 51% WER. These two perform substantially better than ToneNasSepWL, with CER 42% and WER 62% using the same parameters. This means that the ToneSep transcription is associated with an improvement of 9% in CER and 12% in WER. The biggest improvements between conditions are seen with the monophone+trigram models, where ToneSep has a 19% lower CER and a 23% lower WER than ToneNasSepWL.

ToneSep is not the condition with the least vowel symbols, but it is the one with the best performance. This could be due to two reasons. First, what ToneSep appears to be doing is changing the behavior of the triphone window. Kaldi's acoustic model has states with three symbols in them. In a writing system that only has graphemes for segments, the triphone window would, indeed, look at the consonant or vowel in question and to its preceding and following segments. With ToneSep, the tone symbols are surrounded by the vowel the tone belongs to and the following consonant or vowel (or at the nasal symbol). This means that, in practice, when the triphone window looks at the tone, it is looking at two actual phones (the vowel, its tonal cues, and the following consonant/vowel), or even one actual phone (the vowel with its tonal and nasal cues). There are well known effects of tones in their preceding and following segments (Tang, 2008; DiCanio, 2012; Hanson, 2009), so this reduced window might be helping the computer generalize the relatively stable tone patterns of Bribri and their effect on the surrounding segments. The training chops the duration of the vowel into two segments; the first chunk is used to identify the vowel itself, and the second chunk is used to identify the tonal trajectory.[4]

A second reason for the advantage of ToneSep might be the phonetics of the low tone itself. It is not only the most frequent tone in Bribri, but it also the least stable phonetically. The low tone can actually appear as low or mid, depending on its surrounding tones (Coto-Solano, 2015). What Kaldi might be doing is simply learn the more stable patterns of the other tones and label all other pitch patterns as “low".

The reason why ToneNasSepWL is the worst performing transcription is unclear. It might be the case that the addition of the low tone creates an explosion in the number of HMM states, given that the low tone is the most frequent one. Another reason might be the separation of the nasal feature. It is possible that the nasal vowels of Bribri are different enough from their oral equivalents that trying to decouple the vowels from their nasality makes generalization more difficult. As can be seen in figure 1, the NasSep condition also performs poorly. This pattern matches results in languages like Portuguese (Meinedo et al., 2003) and Hindi (Jyothi and Hasegawa-Johnson, 2015), where the best results are obtained by keeping the nasal feature bound to the vowel representations.

Table 3 below shows examples of the transcriptions generated by Kaldi for the validation utterances. In this particular example, the transcription from ToneSep is only off by one space (it doesn’t separate the words *e’ ta̱* ‘so’). The transcription from AllFeats is also fairly good in terms of CER, but it is missing the pronoun *be’* ‘you’. Finally, the ToneNasSepWL transcription misses several words. For example, it transcribed the word *tsítsir* ‘young, small’ as the phonetically similar *chìchi* ‘dog’, and the adverb *wake’* ‘right, anyways’ as *wa* ‘with’.

4 End-to-End Results

End-to-end algorithms need massive amounts of data to train properly (Goodfellow et al., 2016; Glasmachers, 2017), so they are not the most appropriate way to train the small datasets characteristic of extremely low-resource languages. However, it would be useful to test whether the differences detected in the traditional ASR training are also visible in end-to-end training. A CTC loss algorithm with bidirectional RNNs was used, specifically that implemented in DeepSpeech. Two types of end-to-end learning were studied: First, models were trained using only the available Bribri data. This style of training will be called Just Bribri. Second, the Bribri data was incorporated into transfer learning models (Wang and Zheng, 2015; Kunze et al., 2017; Wang et al., 2020). DeepSpeech has existing English language models,[5] trained with 6-layer RNNs. The final two layers were removed and two new layers were grafted onto the RNN. The first four layers would, in theory, use their English model to encode the phonetic information, and the final two layers would receive that information and produce Bribri text as output. Removing two layers was found to be the optimal point of transfer learning, which matches previous results in literature

Utterance meaning:	‘So you were young then, right?’	
Target utterance:	e’ ta̱ be’ bák i̱a̱ tsítsir wake’	
ToneSep	e′ta̱ be′ bák i̱a̱ tsítsir wake′	CER: 3%
AllFeats	e′ta bák i̱a̱ tsítsir wake′	CER: 16%
ToneNasSepWL	e′ ta̱ wake′ chìchi wa	CER: 61%

Table 3: Example of Kaldi transcriptions for three of the experimental conditions, trained with triphone-trigram models. More examples are shown in Appendix A.

[4] No experiment was conducted to test the effect of placing the tone indicator before the vowel (e.g. `d LOW i k HIGH i NAS` for *dikì̱* ‘underneath’). In theory, the performance would be worse given that, in the early milliseconds of a vowel, tones can be phonetically co-articulated with their preceding tone and these two cues would blend together (Xu, 1997; Nguyễn and Trần, 2012; DiCanio, 2014). This effect, called *carryover*, causes greater deformations in pitch than the effect of anticipating the following tone, or *anticipatory assimilation* (Gandour et al., 1993; Coto-Solano, 2017, 93-99). Therefore, the second part of the vowel would provide a clearer tonal cue.

[5] A short experiment was run with the Mandarin DeepSpeech models as the base for transfer training, given that both languages are tonal. However, these models had worse performance than with transfer from the English model.

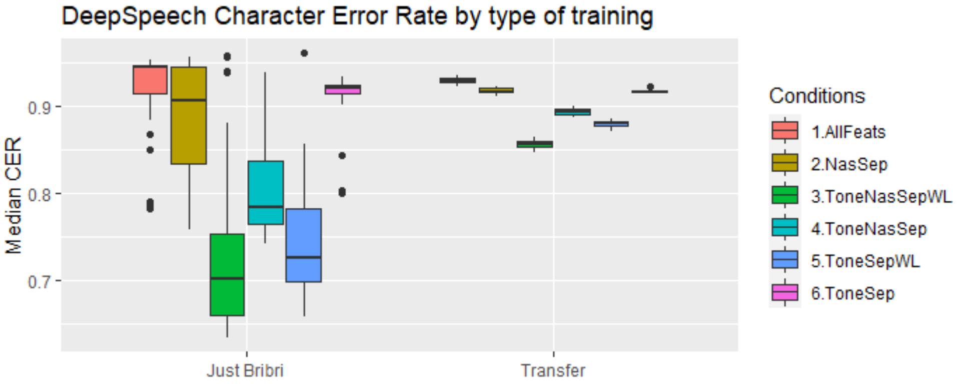

Figure 2: Medians for character error rate (CER) for DeepSpeech models.

(Meyer, 2019; Hjortnaes et al., 2020). This training style will be called Transfer. Both the Just Bribri and Transfer models were trained for 20 epochs, and all other hyperparameters were the same as in the default installation of DeepSpeech.

	Just Bribri	Transfer
AllFeats	95	93
NasSep	91	92
ToneNasSepWL	**70**	**86**
ToneNasSep	78	89
ToneSepWL	73	88
ToneSep	92	91
MaxΔ	25	7

Table 4: Median character error rate (CER) for models trained with CTC (DeepSpeech). MaxΔ indicates the difference between the worst and the best models.

The six transcription conditions were used to train models in both training styles. Same as before, thirty models were trained for each condition. The utterances were randomly shuffled before preparing each model, and then 80% of the utterances were used in the training set (1397 utterances), 10% of the utterances were used for validation (174 utterances), and the final 10% were used for testing. After the training was complete, the median CER and WER were extracted for each model. The median CER for the thirty models in each condition are shown in figure 2.[6]

In the CTC training, the tables have completely turned: ToneSep and AllFeats are the worst performing conditions, and ToneNasSepWL has the best performance. Table 4 shows the median of the 30 medians for each transcription condition. The ToneNasSepWL models trained with Just Bribri have a median of 70% CER, whereas the AllFeats models have a median of 95%, a full 25% worse. As a matter of fact, both WL conditions now have the best performance. This pattern is also visible in the Transfer models: The ToneNasSepWL transcription has a CER of 86%, 7% better than the AllFeats transcription. The median WER is not shown because, for all conditions, the median of the thirty medians was WER=1.

There might be several reasons why the situation has reversed in the CTC models. First, providing an explicit symbol for the low tone might force DeepSpeech to look for more words in the transcription. As can be seen in table 5, the ToneNasSepWL transcription uses the character *4* for the explicit indication of the low tone, which is then eliminated in post-processing to produce a human readable form. The explicit symbol for the low tone appears to force the CTC algorithm to keep looking for tones, and therefore words, whereas, in the other conditions, the CTC algorithm gives up on the search sooner. A second reason why WL performs better is that it provides a clear indication of where a syllable ends, and therefore makes the traverse through the CTC trellis simpler to navigate. Without an explicit low tone, any vowel could be followed by tones, vowels or consonants. On the other hand, when all tones have explicit marking, vowels can only be followed by a tone, which potentially simplifies the path to finding the word.

A third reason for this improvement might have to do with the size of the alphabet: The WL conditions have relatively few symbols for the vowels (12 symbols for ToneNasSepWL versus 48 for

[6] The models were trained using the HPC infrastructure at Dartmouth College in New Hampshire. Each model used 16 CPUs and took approximately 65 minutes to train, for an approximate total of 78 hours of processing.

Utterance meaning:	‘So you were young then, right?’		
Target utterance:	*e’ ta̲ be’ bák i̲a̲ tsítsir wake’*		
Condition	DeepSpeech output	Human-readable output	CER
ToneNasSepWL	`e4′ tax4 i4e4′ i4`	`e′ ta̲ ie′ i`	65%
ToneSep	`e′`	`e′`	91%
AllFeats	`i`	`i`	93%

Table 5: Example of DeepSpeech transcriptions for three of the experimental conditions

AllFeats), which would result in a smaller output layer for the RNNs. Notice that, as with the triphones in Kaldi, the RNNs might be splitting the vowel into separate chunks. It would then proceed to identify the type of vowel from the first chunk, the tone in the second and the nasality in the final part. It would also benefit from the bidirectionality of the neural networks, finding tonal cues in the surrounding segments without the disadvantages of GMM/HMM systems.

Finally, it should be noted that the Transfer models did not provide an improvement in performance. This is somewhat surprising; this might indicate that the Bribri dataset is too small to benefit from the transfer, or that the knowledge of English phones does not overlap sufficiently with the Bribri sound system to produce a boost. Even then, the Transfer models also show effects due to the different transcription conditions, and they also benefited from separating the tone and nasal features from the vowel. This effects will have to be confirmed in the future with other end-to-end techniques, such as *Listen, Attend and Spell* algorithms (Chan et al., 2016) and wav2vec pretraining (Baevski et al., 2020).

5 Conclusions

While hand-engineered representations are suboptimal for high-resource languages, these can still be helpful in low-resource environments, where they can help set up a virtuous cycle of creating imperfect but rapid transcriptions, which can then be improved to create more training materials, improve ASR algorithms, and start helping documentation and revitalization projects right away.

The results above show that performing relatively easy transformations in the input (e.g. not marking the most common tone, separating the tonal markings from the vowel) can lead to major improvements in performance. It also shows that NLP practitioners and linguists can fruitfully combine their knowledge to understand the different features involved in the writing system of a language. Additionally, it provides evidence that the benefits of phonetic transcription can also be gained using semi-orthographic representations. The following recommendations provide a short summary of the results: (i) Separate the tones from the vowels. This will help ASR systems learn their regularities. (ii) Experiment with other features, such as nasality; if they modify the formants of the vowel, they should probably be grouped with the vowel.

Finally, this work is the first attempt at training speech recognition for a Chibchan language. As shown in table 3 and Appendix A, it is feasible to transcribe these languages automatically, and these methods will be refined in the future to incorporate ASR into the documentation pipelines for this language family.

Acknowledgements

The author would like to thank Dr. Sofía Flores for her work on the Bribri corpus, the personnel of Research Computing Department at Dartmouth College for their support with processing resources, and Dr. Samantha Wray and three anonymous reviewers for their work in reviewing and offering helpful suggestions to improve this paper.

References

Oliver Adams. 2018. Persephone Quickstart. `https://persephone.readthedocs.io/en/stable/quickstart.html#using-your-own-data`.

Oliver Adams, Trevor Cohn, Graham Neubig, Hilaria Cruz, Steven Bird, and Alexis Michaud. 2018. Evaluating Phonemic Transcription of Low-resource Tonal languages for Language Documentation. In *LREC 2018 (Language Resources and Evaluation Conference)*, pages 3356–3365.

Oliver Adams, Matthew Wiesner, Shinji Watanabe, and David Yarowsky. 2019. Massively Multilingual Adversarial Speech Recognition. *arXiv preprint arXiv:1904.02210*.

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*.

Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. 2014. Automatic Speech Recognition for Under-resourced Languages: A Survey. *Speech communication*, 56:85–100.

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

William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4960–4964. IEEE.

Eric Chang, Jianlai Zhou, Shuo Di, Chao Huang, and Kai-Fu Lee. 2000. Large Vocabulary Mandarin Speech Recognition with Different Approaches in Modeling Tones. In *Sixth International Conference on Spoken Language Processing*.

Adolfo Constenla, Feliciano Elizondo, and Francisco Pereira. 2004. *Curso Básico de Bribri*. Editorial de la Universidad de Costa Rica.

Rolando Coto-Solano. 2015. The Phonetics, Phonology and Phonotactics of the Bribri Language. In *2nd International Conference on Mesoamerican Linguistics*, volume 25. Los Angeles: California State University.

Rolando Coto-Solano and Sofía Flores Solórzano. 2016. Alineación forzada sin entrenamiento para la anotación automática de corpus orales de las lenguas indígenas de Costa Rica. *Kánina*, 40(4):175–199.

Rolando Coto-Solano and Sofía Flores Solórzano. 2017. Comparison of Two Forced Alignment Systems for Aligning Bribri Speech. *CLEI Electron. J.*, 20(1):2–1.

Rolando Alberto Coto-Solano. 2017. *Tonal Reduction and Literacy in Me̱'pha̱a̱ Vátháá*. Ph.D. thesis, University of Arizona.

Christian DiCanio. 2014. Triqui Tonal Coarticulation and Contrast Preservation in Tonal Phonology. In *Proceedings of the Workshop on the Sound Systems of Mexico and Central America*, New Haven, CT: Department of Linguistics, Yale University.

Christian T DiCanio. 2012. Coarticulation between Tone and Glottal Consonants in Itunyoso Trique. *Journal of Phonetics*, 40(1):162–176.

Ọdétúnjí Àjàdí Ọdélọbí. 2008. Recognition of Tones in Yorùbá Speech: Experiments with Artificial Neural Networks. In *Speech, Audio, Image and Biomedical Signal Processing using Neural Networks*, pages 23–47. Springer.

Isaac Feldman and Rolando Coto-Solano. 2020. Neural Machine Translation Models with Back-Translation for the Extremely Low-Resource Indigenous Language Bribri. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3965–3976.

Sofía Flores Solórzano. 2017a. Corpus oral pandialectal de la lengua bribri. `http://bribri.net`.

Sofía Flores Solórzano. 2019. La modelización de la morfología verbal bribri - Modeling the Verbal Morphology of Bribri. *Revista de Procesamiento del Lenguaje Natural*, 62:85–92.

Sofía Margarita Flores Solórzano. 2017b. *Un primer corpus pandialectal oral de la lengua bribri y su anotación morfológica con base en el modelo de estados finitos*. Ph.D. thesis, Universidad Autónoma de Madrid.

Ben Foley, Josh Arnold, Rolando Coto-Solano, Gautier Durantin, T. Mark Ellison, Daan van Esch, Scott Heath, Frantisek Kratochvil, Zara Maxwell-Smith, David Nash, Ola Olsson, Mark Richards, Nay San, Hywel Stoakes, Nick Thieberger, and Janet Wiles. 2018. Building Speech Recognition Systems for Language Documentation: The CoEDL Endangered Language Pipeline and Inference System (ELPIS). In *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*, pages 205–209.

Jack Gandour, Suvit Ponglorpisit, Sumalee Dechongkit, Fuangfa Khunadorn, Prasert Boongird, and Siripong Potisuk. 1993. Anticipatory tonal coarticulation in thai noun compounds after unilateral brain damage. *Brain and language*, 45(1):1–20.

Elodie Gauthier, Laurent Besacier, and Sylvie Voisin. 2016. Automatic Speech Recognition for African Languages with Vowel Length Contrast. *Procedia Computer Science*, 81:136–143.

Tobias Glasmachers. 2017. Limits of End-to-end Learning. In *Asian Conference on Machine Learning*, pages 17–32. PMLR.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Vishwa Gupta and Gilles Boulianne. 2020a. Automatic transcription challenges for Inuktitut, a low-resource polysynthetic language. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2521–2527.

Vishwa Gupta and Gilles Boulianne. 2020b. 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.

Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. 2014. Deep Speech: Scaling up End-to-End Speech Recognition. *arXiv preprint arXiv:1412.5567*.

Helen M Hanson. 2009. Effects of Obstruent Consonants on Fundamental Frequency at Vowel Onset in English. *The Journal of the Acoustical Society of America*, 125(1):425–441.

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, 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.

INEC. 2011. Población total en territorios indígenas por autoidentificación a la etnia indígena y habla de alguna lengua indígena, según pueblo y territorio indígena. In Instituto Nacional de Estadística y Censos, editor, *Censo 2011*.

Carla Victoria Jara Murillo and Alí García Segura. 2013. *Se' ttö́ bribri ie Hablemos en bribri*. EDigital.

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)*.

Preethi Jyothi and Mark Hasegawa-Johnson. 2015. Improved Hindi broadcast ASR by adapting the language model and pronunciation model using a priori syntactic and morphophonemic knowledge. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Jaspreet Kaur, Amitoj Singh, and Virender Kadyan. 2020. Automatic Speech Recognition System for Tonal Languages: State-of-the-Art Survey. *Archives of Computational Methods in Engineering*, pages 1–30.

Natthawut Kertkeidkachorn, Proadpran Punyabukkana, and Atiwong Suchato. 2014. Using tone information in Thai spelling speech recognition. In *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing*, pages 178–184.

Ettien Koffi. 2020. A Tutorial on Acoustic Phonetic Feature Extraction for Automatic Speech Recognition (ASR) and Text-to-Speech (TTS) Applications in African Languages. *Linguistic Portfolios*, 9(1):11.

Haakon S. Krohn. 2020. Diccionario digital bilingüe bribri. `http://www.haakonkrohn.com/bribri`.

Julius Kunze, Louis Kirsch, Ilia Kurenkov, Andreas Krug, Jens Johannsmeier, and Sebastian Stober. 2017. Transfer Learning for Speech Recognition on a Budget. *arXiv preprint arXiv:1706.00290*.

Tan Lee, Wai Lau, Yiu Wing Wong, and PC Ching. 2002. Using tone information in Cantonese continuous speech recognition. *ACM Transactions on Asian Language Information Processing (TALIP)*, 1(1):83–102.

Ilse Lehiste and Norman J Lass. 1976. Suprasegmental Features of Speech. *Contemporary issues in experimental phonetics*, 225:239.

Gina-Anne Levow, Emily P Ahn, and Emily M Bender. 2021. Developing a Shared Task for Speech Processing on Endangered Languages. In *Proceedings of the Workshop on Computational Methods for Endangered Languages*, volume 1, pages 96–106.

Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, David R Mortensen, Graham Neubig, Alan W Black, et al. 2020. Universal phone recognition with a multilingual allophone system. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8249–8253. IEEE.

Ian Maddieson. 2013. Tone. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Enrique Margery. 2005. *Diccionario Fraseológico Bribri-Español Español-Bribri*, second edition. Editorial de la Universidad de Costa Rica.

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. *arXiv preprint arXiv:2002.06675*.

Hugo Meinedo, Diamantino Caseiro, Joao Neto, and Isabel Trancoso. 2003. AUDIMUS. media: a Broadcast News speech recognition system for the European Portuguese language. In *International Workshop on Computational Processing of the Portuguese Language*, pages 9–17. Springer.

Florian Metze, Zaid AW Sheikh, Alex Waibel, Jonas Gehring, Kevin Kilgour, Quoc Bao Nguyen, et al. 2013. Models of tone for tonal and non-tonal languages. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 261–266. IEEE.

Josh Meyer. 2019. *Multi-task and transfer learning in low-resource speech recognition*. Ph.D. thesis, The University of Arizona.

Alexis Michaud, Oliver Adams, Christopher Cox, and Séverine Guillaume. 2019. Phonetic lessons from automatic phonemic transcription: preliminary reflections on Na (Sino-Tibetan) and Tsuut'ina (Dene) data. In *ICPhS XIX (19th International Congress of Phonetic Sciences)*.

Christopher Moseley. 2010. *Atlas of the World's Languages in Danger*. Unesco.

Quoc Bao Nguyen, Van Tuan Mai, Quang Trung Le, Ba Quyen Dam, and Van Hai Do. 2018. Development of a Vietnamese Large Vocabulary Continuous Speech Recognition System under Noisy Conditions. In *Proceedings of the Ninth International Symposium on Information and Communication Technology*, pages 222–226.

Thị Lan Nguyễn and Đỗ Đạt Trần. 2012. Tonal Coarticulation on Particles in Vietnamese Language. In *International Conference on Asian Language Processing*, pages 221–224.

Jianwei Niu, Lei Xie, Lei Jia, and Na Hu. 2013. Context-dependent deep neural networks for commercial Mandarin speech recognition applications. In *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 1–5. IEEE.

Niko Partanen, Mika Hämäläinen, and Tiina Klooster. 2020. Speech Recognition for Endangered and Extinct Samoyedic languages. *arXiv preprint arXiv:2012.05331*.

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

Carlos Sánchez Avendaño. 2013. Lenguas en peligro en Costa Rica: vitalidad, documentación y descripción. *Revista Káñina*, 37(1):219–250.

Jiulong Shan, Genqing Wu, Zhihong Hu, Xiliu Tang, Martin Jansche, and Pedro J Moreno. 2010. Search by Voice in Mandarin Chinese. In *Eleventh Annual Conference of the International Speech Communication Association*.

Jiatong Shi, Jonathan Amith, Rey Castillo García, Esteban Guadalupe Sierra, Kevin Duh, and Shinji Watanabe. 2021. Leveraging End-to-End ASR for Endangered Language Documentation: An Empirical Study on Yoloxóchitl Mixtec. *arXiv preprint arXiv:2101.10877*.

Katrina Elizabeth Tang. 2008. *The Phonology and Phonetics of Consonant-Tone Interaction*. Ph.D. thesis.

Bao Thai, Robert Jimerson, Dominic Arcoraci, Emily Prud'hommeaux, and Raymond Ptucha. 2019. Synthetic data augmentation for improving low-resource ASR. In *2019 IEEE Western New York Image and Signal Processing Workshop (WNYISPW)*, pages 1–9. IEEE.

Changhan Wang, Juan Pino, and Jiatao Gu. 2020. Improving Cross-Lingual Transfer Learning for End-to-End Speech Recognition with Speech Translation. *arXiv preprint arXiv:2006.05474*.

Dong Wang and Thomas Fang Zheng. 2015. Transfer learning for speech and language processing. In *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 1225–1237. IEEE.

Guillaume Wisniewski, Alexis Michaud, and Séverine Guillaume. 2020. Phonemic transcription of low-resource languages: To what extent can preprocessing be automated? In *1st Joint SLTU (Spoken Language Technologies for Under-resourced languages) and CCURL (Collaboration and Computing for Under-Resourced Languages) Workshop*, pages 306–315. European Language Resources Association (ELRA).

Yi Xu. 1997. Contextual Tonal Variations in Mandarin. *Journal of phonetics*, 25(1):61–83.

Moira Yip. 2002. *Tone. Cambridge Textbooks in Linguistics*. Cambridge University Press.

Shahrul Azmi Mohd Yusof, Abdulwahab Funsho Atanda, and M Hariharan. 2013. A Review of Yorùbá Automatic Speech Recognition. In *2013 IEEE 3rd International Conference on System Engineering and Technology*, pages 242–247. IEEE.

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.

Rodolfo Zevallos, Johanna Cordova, and Luis Camacho. 2019. Automatic Speech Recognition of Quechua Language Using HMM Toolkit. In *Annual International Symposium on Information Management and Big Data*, pages 61–68. Springer.

Appendix A: Additional Transcription Examples

Target	ToneSep	AllFeats	ToneNasSepWL	Meaning
dawáska e’ ta̱ be’ mi̱’ke̱ sulë̀ wa i wéblök	dawáska e′ ta̱ wa̱ e′ mi̱′ke̱ sulë̀ wa wéblö 14%	dawáska e′ ta̱ ma̱ mi̱′ke̱ sulë̀ wa wéblö 14%	dawáska ta̱ mi̱′ke̱ sulë̀ wa wé̱rö 28%	‘during the summer then, you go with your arrow to see them’
dùala tso’i̱a̱ kàl a̱	dùla tso′i̱a̱ kàl a̱ 5%	dùala tso′i̱a̱ kàl ta̱ 5%	dúla tso′ akàla 42%	‘There are birds on the trees.’
iku̱á̱ki̱ iku̱á̱ki̱ sa’ é̱n a̱ i̱à̱ne̱ bua’ë	iku̱á̱ki̱ iku̱á̱ki̱ sa′ i̱à̱ne̱ bua′ë 14%	iku̱á̱ki̱ iku̱á̱ki̱ se′ mí̱a̱ irir bua′ë 26%	wè̱k iku̱á̱ki̱ sa′ i̱à̱ne̱ bua′ë 30%	‘the others, the others, we understand them well’
sìkua i kiè setenta años	sìkua i kiè setenta años 0%	sìkua i kiè setenta a̱ñì̱ 13%	síkwa kë̀ se′ kë̀ ta̱′ 63%	‘[in the] Spanish [language] they say *seventy years* [old]’

Table 6: Additional examples of Kaldi transcriptions for three of the experimental conditions, trained with triphone-trigram models. The numbers represent the character error rate (CER) between the transcription and the target sentence. The fourth example includes code-switching into Spanish.