Evaluating phonemic transcription of low-resource tonal languages for language documentation

Oliver Adams, Trevor Cohn, Graham Neubig, Hilaria Cruz, Steven Bird, Alexis Michaud

Proceedings of LREC 2018: 11th edition of the Language Resources and Evaluation Conference, 7-12 May 2018, Miyazaki (Japan)

Evaluating Phonemic Transcription of Low-Resource Tonal Languages for Language Documentation

Oliver Adams,[1] Trevor Cohn,[1] Graham Neubig,[2] Hilaria Cruz,[3] Steven Bird,[4,5] Alexis Michaud[6]
[1]Computing and Information Systems, The University of Melbourne, Australia
[2]Language Technologies Institute, Carnegie Mellon University, USA
[3]Linguistics Program, Dartmouth College, USA
[4]Northern Institute, Charles Darwin University, Australia
[5]International Computer Science Institute, University of California Berkeley, USA
[6]CNRS-LACITO, National Center for Scientific Research, France
oliver.adams@gmail.com, trevor.cohn@unimelb.edu.au, gneubig@cs.cmu.edu,
steven.bird@cdu.edu.au, hilaria.cruz@dartmouth.edu, alexis.michaud@cnrs.fr

Abstract

Transcribing speech is an important part of language documentation, yet speech recognition technology has not been widely harnessed to aid linguists. We explore the use of a neural network architecture with the connectionist temporal classification loss function for phonemic and tonal transcription in a language documentation setting. In this framework, we explore jointly modelling phonemes and tones versus modelling them separately, and assess the importance of pitch information versus phonemic context for tonal prediction. Experiments on two tonal languages, Yongning Na and Eastern Chatino, show the changes in recognition performance as training data is scaled from 10 minutes up to 50 minutes for Chatino, and up to 224 minutes for Na. We discuss the findings from incorporating this technology into the linguistic workflow for documenting Yongning Na, which show the method's promise in improving efficiency, minimizing typographical errors, and maintaining the transcription's faithfulness to the acoustic signal, while highlighting phonetic and phonemic facts for linguistic consideration.

Keywords: low-resource languages; Asian languages; Mesoamerican languages; speech recognition; language documentation.

1. Introduction

Language documentation involves recording the speech of native speakers. Transcribing these recordings, which are rich cultural and linguistic resources, is an integral part of the language documentation process. However, transcription is slow: it often takes a linguist between 30 minutes to 2 hours to transcribe and translate a minute of speech, depending on the transcriber's familiarity with the language and the difficulty of the content. This is a bottleneck in the documentary linguistics workflow: linguists accumulate considerable amounts of speech, but do not transcribe and translate it all, and there is a risk that untranscribed recordings could end up as "data graveyards" (Himmelmann, 2006, 4,12-13). There is clearly a need for "devising better ways for linguists to do their work" (Thieberger, 2016, 92).

There has been work on low-resource speech recognition (Besacier et al., 2014), with approaches using cross-lingual information for better acoustic modelling (Burget et al., 2010; Vu et al., 2014; Xu et al., 2016; Müller et al., 2017) and language modelling (Xu and Fung, 2013). However, speech recognition technology has largely been ineffective for endangered languages since architectures based on hidden Markov models (HMMs), which generate orthographic transcriptions, require a large pronunciation lexicon and a language model trained on text. These speech recognition systems are usually trained on a variety of speakers and hundreds of hours of data (Hinton et al., 2012, 92), with the goal of generalisation to new speakers. Since large amounts of text are used for language model training, such systems often do not incorporate pitch information for speech recognition of tonal languages (Metze et al., 2013), as they can instead rely on contextual information for tonal disambiguation via the language model (Le and Besacier, 2009; Feng et al., 2012) even though there is no computational burden in additionally using pitch features.

In contrast, language documentation contexts often have just a few speakers for model training and little text for language model training. However, there may be benefit even in a system that overfits to these speakers. If a *phonemic* recognition tool can provide a canvas transcription for manual correction and linguistic analysis, it may be possible to improve the workflow and leverage of linguists. The transcriptions collected in this semi-automated workflow can then be used for refinement of the acoustic model, leading to a snowball effect of better and faster transcription.

In this paper we investigate the application of neural speech recognition models to the task of phonemic and tonal transcription in a language documentation setting where limited resources are available for model training. We use the connectionist temporal classification (CTC) formulation (Graves et al., 2006) for the purposes of direct prediction of phonemes and tones given an acoustic signal, thus bypassing the need for a pronunciation lexicon, language model, and time alignments of phonemes in the training data. By reducing the data requirements we make automatic transcription technology more feasible in a language documentation setting.

We evaluate this approach on two tonal languages, Yongning Na and Eastern Chatino. Na is a Sino-Tibetan language spoken in southwest China with three tonal levels, High (H), Mid (M) and Low (L), and a total of seven tone labels. East-

ern Chatino, spoken in Oaxaca, Mexico, has a richer tone set but both languages have extensive morphotonology (Cruz and Woodbury, 2006; Cruz, 2011; Michaud, 2017). Overall estimates of numbers of speakers for Chatino and Na are similar, standing at about 40,000 for both (Simons and Fennig, 2017), but there is a high degree of dialect differentiation within the languages. The data used in the present study are from the Alawa dialect of Yongning Na, and the San Juan Quiahije (SJQ) dialect of Eastern Chatino; as a rule-of-thumb estimate, it is likely that these materials would be intelligible to a population of less than 10,000.[1]

Though a significant amount of Chatino speech has been transcribed (Chatino Language Documentation Project, 2017), its rich tone system make it a useful point of comparison for our explorations of Na, the language for which automatic transcription is our primary practical concern. Though Na has previously had speech recognition applied in a pilot study (Do et al., 2014), phoneme error rates were not quantified and tone recognition was left as future work.

We perform experiments scaling the training data, comparing joint prediction of phonemes and tones with separate prediction, and assessing the influence of pitch information versus phonemic context on phonemic and tonal prediction in the CTC-based framework. Importantly, we qualitatively evaluate use of this automation in the transcription of Na. The effectiveness of the approach has resulted in its incorporation into the linguist's workflow. Our open-source implementation of this phonemic transcription tool, Persephone, is available online.[2]

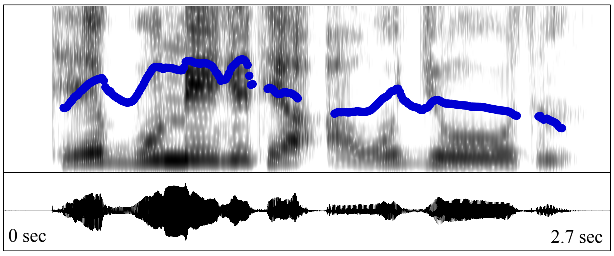
A preliminary version of this work was presented at the Australasian Language Technology Association Workshop (Adams et al., 2017), which we extend and improve upon in this paper by exploring: the effect of including elicited wordlists in training; how effective the model is at predicting tone group boundaries in Na and how this influences the model's capacity to learn tonal rules; discussion of the potential of this approach for reviewing transcriptions; analysis of the Chatino output; refined results for Na involving data preprocessing improvements and more data; results for both random and story-wise cross validation; and presentation of example utterances.

2. Model

The underlying model used is a long short-term memory (LSTM) recurrent neural network (Hochreiter and Schmidhuber, 1997) in a bidirectional configuration (Schuster and Paliwal, 1997). The network is trained with the connectionist temporal classification (CTC) loss function (Graves et al., 2006). Critically, this alleviates the need for alignments between speech frames and labels in the transcription, which we do not have. This is achieved through the use of a dynamic programming algorithm that efficiently sums over the probability of neural network output label that correspond to the gold transcription sequence when repeated labels are collapsed.

[1]For details on the situation for Eastern Chatino, see Cruz (2011, 18-23).

[2]`https://github.com/oadams/persephone`



/tʰi˩˥ | go˧mi˧-dʑo˥ | tʰi˩˥, | ɑ˩ʁo˧ dʑo˩ tsɯ˩ | mv̩˩. |/
Quant à la sœur, elle demeurait à la maison, dit-on.
As for the sister, she stayed at home.
而妹妹呢，留在家里。

Target label sequence:

1.	tʰ i g o m i dʑ o tʰ i ɑ ʁ o dʑ o t s ɯ m v̩
2.	˩˥ ˧ ˧ ˥ ˩˥ ˩ ˧ ˩ ˩ ˩
3.	tʰ i ˩˥ g o ˧ m i ˧ d ʑ o ˥ tʰ i ˩˥ ɑ ˩ ʁ o ˧ dʑ o ˩ t s ɯ ˩ m v̩ ˩
4.	tʰ i ˩˥ | g o ˧ m i ˧ d ʑ o ˥ | tʰ i ˩˥ | ɑ ˩ ʁ o ˧ dʑ o ˩ t s ɯ ˩ | m v̩ ˩

Figure 1: A sentence from the Na corpus. Top to bottom: spectrogram with F_0 in blue; waveform; phonemic transcription; French, English and Chinese translations; target label sequences: (1) phonemes only, (2) tones only, (3) phonemes and tones together, and (4) phonemes and tones with tone group boundary markers, "|".

The use of an underlying recurrent neural network allows the model to implicitly model context via the parameters of the LSTM, despite the independent frame-wise label predictions of the CTC network. It is this feature of the architecture that makes it a promising tool for tonal prediction, since tonal information is suprasegmental, spanning many frames (Mortensen et al., 2016). Context beyond the immediate local signal is indispensable for tonal prediction, and long-ranging context is especially important in the case of morphotonologically rich languages such as Na and Chatino.

Past work distinguishes between *embedded* tonal modelling, where phoneme and tone labels are jointly predicted, and *explicit* tonal modelling, where they are predicted separately (Lee et al., 2002). We compare several training objectives for the purposes of phoneme and tone prediction. This includes separate prediction of 1. phonemes and 2. tones, as well as 3. jointly predict phonemes and tones using one label set. We additionally explore 4. joint prediction of phonemes, tones and the tone group boundaries (TGBs) which delimit tone groups. Figure 1 presents an example sentence from the Na corpus described in §3.1., along with an example of these four objectives.

3. Experimental Setup

We designed the experiments to answer these primary questions:

1. How do the error rates scale with respect to training data?
2. How effective is tonal modelling in a CTC framework?
3. To what extent does phoneme context play a role in tone prediction?

4. Does joint prediction of phonemes and tones help minimize error rates?

We assess the performance of the systems as training data scales from 10 minutes to 224 minutes of spontaneous speech of a single Na speaker, and between 12 and 50 minutes for a single speaker of Chatino. Experimenting with this extremely limited training data gives us a sense of how much a linguist needs to transcribe before this technology can be profitably incorporated into their workflow.

We evaluate both the phoneme error rate (PER) and tone error rate (TER) of models based on the same neural architecture, but with varying input features and output objectives. Input features include log Filterbank features[3] (`fbank`), pitch features of Ghahremani et al. (2014) (`pitch`), and a combination of both (`fbank+pitch`). These input features vary in the amount of acoustic information relevant to tonal modelling that they include. The output objectives correspond to those discussed in §2.: tones only (`tone`), phonemes only (`phoneme`), or jointly modelling both (`joint`). We denote combinations of input features and target labellings as ⟨`input`⟩⇒⟨`output`⟩.

In case of tonal prediction we explore similar configurations to that of phoneme prediction, but with two additional points of comparison. The first is predicting tones given one-hot phoneme vectors (`phoneme`) of the gold phoneme transcription (`phoneme`⇒`tone`). The second predicts tones directly from pitch features (`pitch`⇒`tone`). These points of comparison serve to give us some understanding as to how much tonal information is being extracted directly from the acoustic signal versus the phoneme context.

In the `fbank+pitch`⇒`joint` configuration, we additionally explore the difference in performance between models that jointly predict tone group boundaries as well as phonemes and tones.

3.1. Data

We explore application of the model to the Na corpus (Michaud and Latami, 2017b) that is part of the Pangloss collection (Michailovsky et al., 2014). This corpus consists of around 100 narratives, constituting 11 hours of speech from one speaker in the form of traditional stories, and spontaneous narratives about life, family and customs (Michaud, 2017, 33). Several hours of recordings of spontaneous narratives have been phonemically transcribed, and we used up to 224 minutes of this for training, 24 minutes for validation and 23 minutes for testing. This represents an increase in training data from that used in preliminary reports on this work (Adams et al., 2017). Included in this additional data is 6 minutes and 30 seconds of semi-automatically transcribed speech of the narrative *Housebuilding 2* (Michaud and Latami, 2017a), where an automatic transcription of a model trained on less data was used as a canvas by the linguist during a field trip in 2017.

The total number of phoneme and tone labels used for automatic transcription was 90 and 7 respectively. This represents an increase in the number of phonemes from previously reported results, where 78 were used (Adams et al., 2017). This increase is the result of improved preprocessing of the linguist's transcriptions, taking place in parallel with improvements to the original transcriptions to arrive at full consistency. Concerning the preprocessing, (1) we represent fillers /əəə…/ and /mmm…/ with their own tokens; (2) onsetless syllables which are digraphs /wæ/, /wɑ/, /wɤ/, /jæ/, /jɤ/, /jo/ are also now represented as a single token because they constitute phonological units (syllable rhymes; Na syllables are composed of an onset, a rhyme, and a tone). Concerning improvements to the original transcriptions, we addressed cases where the same phoneme had inconsistent representation in the corpus, such as /wæ̃/ and /w̃æ/, as well as an instance where the unicode representation of a single phoneme was sometimes `v+nasality+syllabic diacritic` and sometimes `v+syllabic diacritic+nasality`. We computed the Na results of Tables 1-3 using the larger suite of 224 minutes and these preprocessing changes.

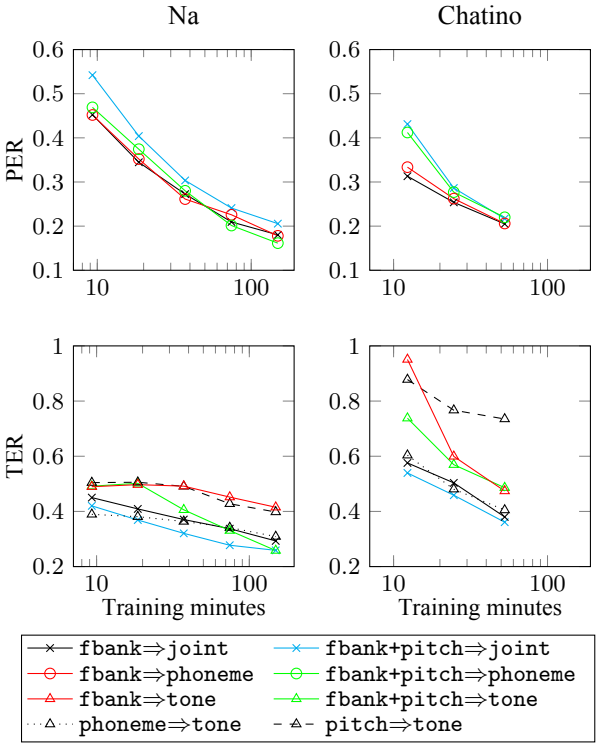


Figure 2: Phoneme error rate (PER) and tone error rate (TER) on test sets as training data is scaled from around 10 minutes up to 150 minutes for Na (left) and up to 50 minutes for Chatino (right). The legend entries are formatted as ⟨input⟩ ⇒ ⟨output⟩ to indicate model input and output.

For Chatino, we used data of Ćavar et al. (2016) from the GORILLA language archive for Eastern Chatino of San Juan Quiahije, Oaxaca, Mexico (Cavar et al., 2016) for the purposes of comparing phoneme and tone prediction with Na when data restriction is in place. We used up to 50 minutes of data for training, 6 minutes for validation and 6 minutes for testing. The phoneme inventory we used consists of 31 labels along with 14 tone labels. For both languages, preprocessing involved removing punctuation and any other symbols that are not phonemes, tones or the tone

[3] 41 log Filterbank features along with their first and second derivatives

		Chatino		Na		
Input	Output	PER ↓	TER ↓	PER ↓	TER ↓	TGB-F1 ↑
`fbank`	`joint`	0.205	0.380	0.119	0.219	0.845
`fbank+pitch`	`joint`	0.217	0.361	0.128	0.177	0.856
`fbank`	`phoneme`	0.206	-	0.112	-	-
`fbank+pitch`	`phoneme`	0.220	-	0.129	-	-
`fbank`	`tone`	-	0.474	-	0.394	0.830
`pitch`	`tone`	-	0.735	-	0.513	0.789
`fbank+pitch`	`tone`	-	0.486	-	0.240	0.847
`phoneme`	`tone`	-	0.405	-	0.267	0.850

Table 1: The phoneme error rate (PER) and tone error rate (TER) of a variety of models for transcription of Chatino and Na, along with tone group boundary F1 scores (TGB-F1) for Na. The Chatino models were trained on a total 50 minutes of training data, while the Na models were trained on 224 minutes.

group boundary (TGB) symbol, "|", such as hyphens connecting syllables within words.

3.2. Training and Tuning

We trained each configuration for a minimum of 30 epochs, stopping if no improvements on the validation set were found for 10 consecutive epochs thereafter. Occasionally ($<$ 10% of the time) model training failed to converge, in which case training was restarted. Batch size varied with the training set, between 4 and 64 utterances as the total number of training utterances scaled from 128 to 2,048.

We tuned the hyperparameters on the Na validation set, settling on 3 hidden layers with 250 hidden units as a consistently solid performer across `fbank`⇒`phoneme`, `fbank+pitch`⇒`joint` and `phoneme`⇒`tone` for varying amounts of training data up to 150 minutes. For the full Na data set of 224 minutes, 400 hidden layers performed better and was used for the results in Tables 1-3. We averaged results across 3 training runs.

4. Quantitative Results

Figure 2 shows the phoneme and tone error rates for Na and Chatino as training is scaled from around 10 minutes.

Note that the results for Na reported in Tables 1-3 are substantially better than the best results reported in Figure 2, on account of preprocessing changes, increased data and use of hyperparameters that are more effective for the larger amount of data. The settings used for Figure 2 were tailored to a smaller data set, but could be further improved. The Chatino data and its preprocessing remain unchanged.

Error rate scaling Error rates decrease logarithmically with training data. The best methods have a lower than 30% PER with 30 minutes of training data. We believe it is reasonable to expect similar trends in other languages with similar data quality and a single speaker. These results thus suggest how much audio linguists might need to transcribe before semi-automation can become part of their workflow.

Tonal modelling TER is always higher than PER for the same amount of training data, despite there being only 7 tone labels versus 90 phoneme labels in our Na experiment. This is true even when pitch features are present. However, it is unsurprising since the tones have overlapping pitch ranges, and can be realized with vastly different pitch over the course of a single sentence. This suggests that context is more important for predicting tones than phonemes, which are more context-independent.

`fbank`⇒`tone` and `pitch`⇒`tone` are vastly inferior to other methods, all of which are privy to phonemic information via training labels or input. However, combining the fbank and pitch input features (`fbank+pitch`⇒`tone`) makes for a competitive approach for tonal prediction in Na at maximum training data. This indicates both that these features are complementary and that the model has learnt a representation useful for tonal prediction that is on par with explicit phonemic information.

Though tonal prediction is more challenging than phoneme prediction, these results suggest automatic tone transcription is feasible using this architecture, even without inclusion of explicit linguistic information such as constraints on valid tone sequences, which is a promising line of work.

In the case of phoneme-only prediction, the use of pitch information doesn't help reduce the PER, which differs from previous work (Metze et al., 2013), including our own preliminary results (Adams et al., 2017).

Phoneme context To assess the importance of context in tone prediction, `phoneme`⇒`tone` gives us a point of comparison where no acoustic information is available at all. It performs reasonably well for Na, and competitively for Chatino. One likely reason for its solid performance is that long-range context is modelled more effectively with phoneme input features, as there are vastly fewer phonemes per sentence than speech frames. The rich morphotonology of Na and Chatino means context is important in the realisation of tones, explaining why `phoneme`⇒`tone` can perform almost as well as methods using acoustic features.

Joint prediction Interestingly, joint prediction of phonemes and tones does not consistently outperform the best methods for phoneme-only prediction. In light of the celebrated successes of multitask learning in various domains (Collobert et al., 2011; Deng et al., 2013; Girshick, 2015; Ramsundar et al., 2015; Ruder, 2017), one might expect training with joint prediction of phonemes and tones to help, since it gives more relevant contextual information

to the model. The TER, on the other hand, is always at its lowest when tones are jointly predicted with phonemes.

Na versus Chatino The trends observed in the experimentation on Chatino were largely consistent with those of Na, but with higher error rates owing to less training data and a larger tone label set. There are two differences with the Na results worth noting. One is that `phoneme⇒tone` is somewhat more competitive in the case of Chatino, suggesting that phoneme context plays a more important role in tonal prediction in Chatino. The second is that `fbank⇒tone` outperforms `pitch⇒tone`, and that adding pitch features to Filterbank features offers less benefit than in Na. This may be because the phonological interpretation of the pitch features is less straightforward in Chatino than it is in Na.

Tone group boundary markers An important concept in the morphotonology of Na is the notion of a *tone group*. Tone groups (phonological phrases) describe segments of speech in which the realization of tones may have interdependent relationships. Since the rules by which tones influence neighbouring tones in a tone group are well described (Michaud, 2017), there is potential to harness these rules to improve tone transcription in a rule-based fashion by enforcing transcriptions to comport with these rules. While the specifics of the rules are language dependent, there is potential to enable linguists to describe these rules to the model in order to improve transcription in the language documentation setting. However, this relies on identifying tone groups since the tonal rules do not hold across the dividing lines between tone groups.

Tone group boundaries (TGBs) are the points that demarcate tone groups. While TGBs are a somewhat more abstract concept than that of phones or tones, there are acoustic features that may be harnessed to determine these tone group boundaries, including rhythm, duration, F_0, and details in the articulation of vowels and consonants. In light of the performance achieved when predicting tones from phonemes without acoustic information (`phoneme⇒tone`), and the potential value in harnessing tone group information for improved tonal prediction or rule-based methods, we additionally evaluated the performance of TGB prediction.

The first two rows of Table 2 show the performance of models trained without and with TGB prediction, respectively. TGB prediction is surprisingly accurate as per the F1 score. Prediction of TGBs decreases phoneme error rates somewhat, and decreases tone error rate moreso. TGBs have influence on the tones that precede and follow them, so this is unsurprising. While there is in principle no inherent link between tone group boundaries and phonemes (any syllable can occur before a tone-group boundary), some morphemes, such as the reported speech particle /tsɯ/, are frequent at the end of sentences, and the end of sentences also means the end of a tone group. Topic markers (the most frequent morphemes in the language) are also often found at TGBs and that can create a bias. This can potentially account for the decrease in PER when the model is trained to predict TGBs.

Elicited speech It is common practice in field linguistics to elicit clear non-spontaneous speech of interesting grammatical constructs to complement the collection of natural spontaneous speech. Such elicitation is useful for linguistic analysis, since some forms are unlikely to be found in a small corpus of spontaneous speech, and thus one cannot arrive at complete morphological paradigms. On the other hand, elicited speech tends to have bias and lacks many properties of spontaneous speech and so a balance of spontaneous and elicited speech is considered important (Cruz and Woodbury, 2014; Niebuhr and Michaud, 2015).

TGB	Wordlist	PER ↓	TER ↓	TGB-F1 ↑
No	No	0.131	0.184	-
Yes	No	0.128	0.177	0.856
No	Yes	0.129	0.178	-
Yes	Yes	0.135	0.179	0.858

Table 2: Results for Na with 224 minutes of spontaneous narrative speech and, where applicable, and 105 minutes of elicited wordlist speech. TGB-F1 is the F1 score of tone group boundary prediction.

Cross-validation	PER ↓	TER ↓	TGB-F1 ↑
Story-wise	0.163	0.205	0.842
Random	0.150	0.189	0.855
Default test set	0.128	0.177	0.856

Table 3: *Story-wise* cross-validation results for Na. Since a linguist will apply their model to utterances in narratives unseen in the training data, the first row is perhaps most representative of the results that can be expected in practice.

Supplementing the 224 minutes of Na training data is 104 minutes of elicited speech in the form of wordlists. For linguists interested in incorporating automation into their transcription workflow, one question is: what is the relative value of elicited speech versus spontaneous speech for improving the system? For insight into this we additionally include the elicited speech in the Na training set, constituting a 46% increase in the total duration of training data.

Table 2 shows the performance change when wordlists are additionally included in the training data. Using wordlists and not predicting TGBs yields a comparable improvement to adding prediction of TGBs without wordlists. However, when both TGBs are predicted and wordlists are used, the PER goes up, even though the TGB-F1 does not suffer. TGBs are easy to predict in the wordlists because there tend to be fewer of them and they tend to delimit repeats of the same word. As a result the phonemes TGBs co-occur with in that context is biased differently to those that they co-occur with in the spontaneous narratives, even though the TGB serves the same function. The nature of the interdependence between TGBs and phonemes is thus different in the narratives versus the wordlists. This observation illustrates how speech processing tools can help characterize how the function of a given sign differs across linguistic data sets, opening up new possibilities for linguistic-semiotic studies of speech corpora.

Our results are consistent with conventional machine learning understanding that training data should be similar to test

data. Thus, slower hyper-articulated training data may not necessarily help a model transcribe faster hypo-articulated speech that the linguist ultimately wants to transcribe.

***Story-wise* cross-validation** The quantitative results reported up to this point are based on training, validation and test sets randomly selected at the utterance level. However, this means the training set gets a fair representation of utterances from all the narratives present in the test set. Since the vocabulary of different narratives is different, it may be more insightful to evaluate the performance on a held-out narrative. We performed cross-validation where each fold involves testing on one of 27 held-out narratives (stories). Results are presented in Table 3, using the `fbank+pitch⇒joint` configuration with TGB prediction and without wordlists. Performance is substantially worse than the performance on the test set, with large variation between the stories for a minimum and maximum PER of 0.125 and 0.249, respectively and a minimum and maximum TER of 0.157 and 0.241. To ensure that this difference between story-wise cross-validation and the default test set reflects the challenge of encountering new story-specific content and isn't simply an artefact of the test set, we performed random cross-validation with held-out sets of the size of the average narrative. Performance for this task was worse than on the test set, but substantially better than story-wise cross-validation. The story-wise cross validation is thus most representative of the error rates to be expected in the transcription of subsequent Na data.

Implicit learning of tone rules In Na, a set of phonological tone rules govern how the underlying tones of words are converted to surface tones when realised in a tone group. To gain some insight into how prediction of TGBs can inform how tones should be predicted, we consider instances in the transcription where tone group boundaries influence the realisation of tones, and compare performance of the model that predicts TGBs (denoted as `TGB`) versus the one that doesn't (denoted as `¬TGB`).

Tone rule 6 (Michaud, 2017, 323) states *"In tone-group-final position, H and M are neutralized to H if they follow a L tone."* The Mid and High tones are acoustically identical in the tone group final position, and so the transcription is normalized to H. For example, /|dʑɤ˩ts^{h}o˦|/ becomes /|dʑɤ˩ts^{h}o˥|/.

In the test set, in 26.4% of the instances where this rule applies, it had not been applied by the human transcriber. Michaud hesitated on whether to transcribe according to the surface phonology or the tonal string prior to the application of this phonological rule. A decision was only made in 2015 and not all the narratives have yet been normalized. This means the training set used a mix of both tones and generally biased towards the high tone. "The transcription of spontaneous speech in little-known languages (...) is built on a more or less shaky foundation of linguistic hypotheses (...). It is certainly not raw or unchanging data." (Jacobson et al., 2001, 81).

The percentage of tones that were M in such positions was 27.7% and 23.0% for `TGB` and `¬TGB` respectively, which is close to the ratio in the reference transcriptions.

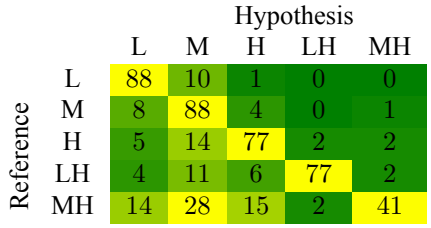


Figure 3: Confusion matrix showing the rates of substitution between Na tones (as a percentage, normalized per reference tone).

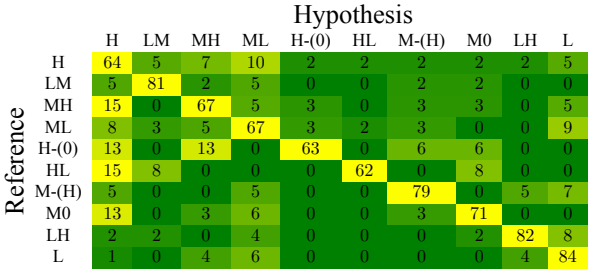


Figure 4: Confusion matrix showing the rates of substitution between Chatino tones (as a percentage, normalized per reference tone).

`¬TGB` transcribed with either an M or H tone 63.7% of the time in instances where the rule applies. However, `TGB` predicted M or H 80.4% of the time. Considering only instances when the latter model predicted a following TGB (TGB recall was 87.5%), this probability increased to 83.1%. This does suggest that predicting TGBs helps the model to learn that these tones should be transcribed as M or H in this situation.[4]

For tone rule 3 ("*In tone-group-initial position, H and M are neutralized to M*") the human annotator was 100% accurate. For instances where rule 3 applies, an M tone was predicted 88.6% by `¬TGB`, and 89.4% by `TGB`. Considering instances where the latter model accurately predicts the TGB (TGB recall was 83.2%), the accuracy increases to 93.1%.

There is thus some evidence to suggest that these tonal rules are learnt implicitly and benefit from TGB prediction.

Na tone errors Figures 3 and 4 show the most common tone substitution mistakes for Na and Chatino respectively using the `fbank+pitch⇒joint` configuration. The relative rates of substitution were similar for other methods. For Na, the most mis-recognized tone was the MH contour, which was frequently misclassified as M, H and L. These three tones are far more common, giving a bias to the training data. Moreover, in running speech the M and H tones have pitch ranges and phonetic contours that overlap substantially with the MH tone (due to tonal coarticulation as well as intonation: the conveyance of prominence and phrasing).

Chatino tone errors For Chatino, the most common errors were mislabelling tones as tone 1 (H) instead of 32 (MH), 0 ("super high"), 14 (HL-(0)) and 20 (M0). These tones generally have a similar pitch to tone 1. The speaker's

[4]These percentages were based on the test hypotheses across 4 different trained models.

Reference	tʰ i ꜔ k i ꜔ s e ꜓ | tʰ i ꜔ dz ɯ ꜓ s e ꜓ |
TGB	tʰ i ꜔ k i ꜓ s e ꜓ | tʰ i ꜔ dz i ꜓ s e ꜓|
¬TGB	tʰ i ꜔ k i ꜓ s e ꜓ tʰ i ꜖ dz i ꜓ s e ꜓

Table 4: An example Na transcription (from the narrative *Sister V1*, utterance #30), and the automatic transcriptions of two models from the joint phoneme and tone prediction task. TGB is a model that additionally predicts tone group boundary markers, while ¬TGB does not. The reference transcription has punctuation and syllable boundaries removed.

Reference	n d e 2 j y an 1 w a 42 n e 2
Hypothesis	n d e 2 j y o 14 w a 42 r e 2 n e 2
Revised	n d e 2 j y an 1 w a 42 r e 2 ne 2

Table 5: An example transcription of a Chatino sentence. There were common errors made in the model's hypothesis, such as confusing low and high back nasal vowels. The automatic transcription also highlighted errors in the reference transcription, leading to its revision.

tone 1 goes from 310-315HZ. Tone 32 runs roughly 270-290Hz, so it is possible the model is catching the top range of the contour. Similarly, tone 14 and 1 begin at the same level, with 14 going from 300-170Hz. This suggests that the model isn't considering the whole contour (see §5.2.). In isolation, it is difficult to distinguish the 0 and 1 tone, and tone 20 also starts roughly at 310Hz before going up to 370Hz and back down to 340Hz. In contrast, there was limited confusion between tones with substantially different pitch, even frequently occurring ones such as tone 4 (L).

5. Qualitative Discussion

5.1. Na

The error rates in the above quantitative analysis are promising, but is this system of practical use in a linguistic workflow? We discuss here the experience of a linguist (Michaud) in applying this model to Na data to aid in transcription of 9 minutes and 30 seconds of speech.

The phonemic errors typically make linguistic sense: they are not random added noise and often bring the linguist's attention to phonetic facts that are easily overlooked because they are not phonemically contrastive.

One set of such errors is due to differences in articulation between different morphosyntactic classes. For example, the noun 'person' /hĩ˥/ and the relativizer suffix /-hĩ˥/ are segmentally identical, but the latter is articulated much more weakly than the former and it is often recognized as /ĩ/ in automatic transcription, without an initial /h/. Likewise, in the demonstrative /ʈʂʰɯ˥/ the initial consonant /ʈʂʰ/ is often strongly hypo-articulated, resulting in its recognition as a fricative /ʂ/, /ʐ̩/, or /ʑ/ instead of an aspirated affricate. The extent to which a word's morphosyntactic category influences its pronunciation is known to be language-specific (Brunelle et al., 2015); the phonemic transcription tool indirectly reveals that this influence is considerable in Na.

A second set is due to loanwords containing combinations of phonemes that are unattested in the training set. For example /ʑɯ˩pe˧/, from Mandarin *rìběn* (日本 , 'Japan'). /pe/ is otherwise unattested in Na, which only has /pi/; accordingly, the syllable was identified as /pi/. In documenting Na, Mandarin loanwords were initially transcribed with Chinese characters, and thus cast aside from analyses, instead of confronting the issue of how different phonological systems coexist and interact in language use.

A third set of errors made by the system result in an output that is not phonologically well formed, such as syllables without tones and sequences with consonant clusters such as /kgɣ̩/. These cases are easy for the linguist to identify and amend.

The recognition system currently makes tonal mistakes that are easy to correct on the basis of elementary phonological knowledge: it produces some impossible tone sequences such as M+L+M inside the same tone group. Very long-ranging tonal dependencies are not harnessed so well by the current tone identification tool. This is consistent with quantitative indications in §4. and is a case for including a tonal language model or refining the neural architecture to better harness long-range contextual information.

Table 4 exemplifies common errors and successes of the models. Erroneous replacement of the mid tone (/˧/) with the low tone (/˩/) was one of the most common mistakes for all models. In the second tone group, the absence of a tone group boundary following the /tʰ i ˧/ precludes the use of the mid-high tone (/˦/), even though phonetically there is a rise there. The model with tone group boundary prediction (TGB) halved the number of mis-transcriptions of /˧/ as /˦/, suggesting it used information about tone groups to learn a more phonological representation. In all models, misclassification of /ɯ / as /i/ was one of the most common errors.

5.2. Chatino

For tonal prediction in Chatino, the model has issues distinguishing between ascending and descending tones that have overlapping pitch ranges. There is additional trouble with distinguishing contours and floating tones. It was noted by the linguist (Hilaria Cruz) that in many of these cases it appears as though the model likes to pick up just one point in the tonal range. This is not inconsistent with typical behaviour of CTC-based networks, where label probabilities tend to spike in narrow windows of time (Graves et al., 2006). The model may be getting overconfident in the prediction of tone in a narrow part of the contour, but a more thorough investigation into the timing and cause of label probability peaks is required to be conclusive.

As for phonemes, the system had issues recognizing laminal sounds (eg. *ndyke32wan4* recognized as *ne32wan4*), prenasalized stops and glottal stops (eg. *ntyqya24qa1* recognized as *ya140qa1*). All of these are phonemically contrastive and are key sounds in SJQ Chatino.

There are also frequent issues with back nasal vowels. In the example in Table 5, there is a confusion between a low back and high back nasal vowel. There also tend to be is-

sues with double mora (eg. *ja4jlyo20ren2enq1* recognized as *ja4jlyo20ren1*).

6. Benefits for the Linguist

Using this automatic transcription as a starting point for manual correction was found to confer several benefits to the linguists.

Faithfulness to acoustic signal The model produces output that is faithful to the acoustic signal. In casual oral speech there are repetitions and hesitations that are sometimes overlooked by the transcribing linguist. When using an automatically generated transcription as a canvas, there can be full confidence in the *linearity* of transcription, and more attention can be placed on linguistically meaningful dialogue with the language consultant. There are also perceived benefits to this faithfulness even in the case of SJQ Chatino, where the linguist is a native speaker (Cruz).

Typographical errors and the transcriber's mindset Typographic errors are common, with a large number of phoneme labels and significant use of combinations of keys (Shift, Alternative Graph, etc). By providing a high-accuracy first-pass automatic transcription, much of this manual data entry is avoided. Enlisting the linguist solely for correction of errors also allows them to embrace a critical mindset, putting them in "proofreading mode," where focus can be entirely centred on assessing the correctness of the system output without the additional distracting burden of data entry.

In the Na documentation workflow, the importance of this effect is amplified, since the linguist is not a native speaker: transcriptions are made during fieldwork with a language consultant and are difficult to correct later on based only on auditory impression when the consultant is not available.

Although native speaker linguists have the great advantage of being able to carry out transcription independent of consultants (as in the case of Hilaria Cruz for SJQ Chatino), native language orthographies are for the most part very young and for this reason, there are few people who are trained to perform these tasks. The transcription is thus overwhelmingly handled by few overworked native linguists, which has led to repetitive stress injuries from excessive typing.

Speed Assessing automatic transcription's influence on the speed of the overall language documentation process will require time. Language documentation is a holistic process. Beyond phonemic transcription, documentation of Na involves other work that happens in parallel: translating, copying out new words into the Na dictionary, and being constantly on the lookout for new and unexpected linguistic phenomena.

In the case of Na, this all takes place in the context of discussions with a native speaker linguist. Further complicating this, Michaud's proficiency of the language and speed of transcription is dynamic, improving over time. This makes comparisons difficult.

From this preliminary experiment, the efficiency of the transcription in the Na workflow was perceived to be improved, but the benefits lie primarily in the advantages of providing a transcript faithful to the recording, and allowing the linguist to minimize manual entry, focusing on correction and enrichment of the transcribed document.

The snowball effect More data collection means more training data for better automatic transcription performance. The process of improving the acoustic model by training on such semi-automatic transcriptions has begun, with the freshly transcribed *Housebuilding2* used in this investigation having already been incorporated into subsequent Na acoustic modelling training. In the current set-up, this has involved sending new transcriptions between the linguist and computer-science for re-training, though it's conceivable this process could be automated at the linguist's end.

Reviewing transcriptions A goal of Michaud in the Na documentation process is carefully groomed transcriptions. As stated earlier, conventions for transcribing a newly documented language are not static but change.

The process of using cross-validation to review transcriptions for Na is now in its early stages. In this process, some errors in transcription have been noted that arose from the workflow: A form of respeaking that took place in the documentation has had some minor influence on the transcription. Sometimes the consultant would be requested to respeak a few seconds of speech for the greatest clarity, or Michaud would respeak. In both cases, substitutions of one acceptable variant for another can happen, such as replacing "this" for "that" when they are semantically equivalent. One instance is in *Buried Alive 2*, sentence #123 (Michaud and Latami, 2017b), where Persephone predicted /tʰɣ̣˧/ in the 5th tone group, while the manual transcription has /ʈʂʰɯ˧/ on the basis of a subtly distinct respeaking.

In 33 of 207 transcriptions in the Chatino validation set, comparison of the model hypothesis with the reference transcription helped the linguist to spot errors in the reference transcription (eg. Table 5). Since the Na narratives and Chatino read speech are substantially different, this suggests cross-language generality in the potential for automation to help highlight potential inconsistencies in the manual transcription, as well as aiding in the transcription of untranscribed resources.

7. Conclusion

We have presented the results of applying a CTC-based LSTM model to the task of phoneme and tone transcription in a resource-scarce context: that of a newly documented language. Beyond comparing the effects of a various training inputs and objectives on the phoneme and tone error rates, we reported on the application of this method to linguistic documentation of Yongning Na. Its applicability as a first-pass transcription is very encouraging, and it has now been incorporated into the workflow for transcribing hitherto untranscribed speech as well as reviewing existing transcriptions. Our results give an idea of the amount of speech other linguists might aspire to transcribe in order to bootstrap this process: as little as 30 minutes in order to obtain a sub-30% phoneme error rate as a starting point, with further improvements to come as more data is transcribed in the semi-automated workflow.

8. Acknowledgements

We are very grateful for support from NSF Award 1464553 *Language Induction Meets Language Documentation*.

9. Bibliographical References

Adams, O., Cohn, T., Neubig, G., and Michaud, A. (2017). Phonemic transcription of low-resource tonal languages. In *Australasian Language Technology Association Workshop 2017*, pages 53–60.

Besacier, L., Barnard, E., Karpov, A., and Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85–100.

Brunelle, M., Chow, D., and Nguyễn, T. N. U. (2015). Effects of lexical frequency and lexical category on the duration of Vietnamese syllables. In The Scottish Consortium for ICPhS 2015, editor, *Proceedings of 18th International Congress of Phonetic Sciences*, pages 1–5, Glasgow. University of Glasgow.

Burget, L., Schwarz, P., Agarwal, M., Akyazi, P., Feng, K., Ghoshal, A., Glembek, O., Goel, N., Karafiát, M., Povey, D., and Others. (2010). Multilingual acoustic modeling for speech recognition based on subspace Gaussian mixture models. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 4334–4337. IEEE.

Ćavar, M. E., Cavar, D., and Cruz, H. (2016). Endangered Language Documentation: Bootstrapping a Chatino Speech Corpus, Forced Aligner, ASR. In *LREC*, pages 4004–4011.

Chatino Language Documentation Project. (2017). Chatino Language Documentation Project Collection.

Collobert, R., Weston, J., and Karlen, M. (2011). Natural Language Processing (almost) from Scratch. 1:1–34.

Cruz, E. and Woodbury, T. (2006). El sandhi de los tonos en el Chatino de Quiahije. In *Las memorias del Congreso de Idiomas Indígenas de Latinoamérica-II*. Archive of the Indigenous Languages of Latin America.

Cruz, E. and Woodbury, T. (2014). Finding a way into a family of tone languages: The story and methods of the Chatino Language Documentation Project. *Language Documentation and Conservation*, 8:490–524.

Cruz, E. (2011). *Phonology, tone and the functions of tone in San Juan Quiahije Chatino*. Ph.D., University of Texas at Austin, Austin.

Deng, L., Hinton, G., and Kingsbury, B. (2013). New types of deep neural network learning for speech recognition and related applications: an overview. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8599–8603. IEEE, may.

Do, T.-N.-D., Michaud, A., and Castelli, E. (2014). Towards the automatic processing of Yongning Na (Sino-Tibetan): developing a 'light' acoustic model of the target language and testing 'heavyweight' models from five national languages. In *4th International Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU 2014)*, pages 153–160, St Petersburg, Russia, may.

Feng, Y.-M., Xu, L., Zhou, N., Yang, G., and Yin, S.-K. (2012). Sine-wave speech recognition in a tonal language. *The Journal of the Acoustical Society of America*, 131(2):EL133–EL138.

Ghahremani, P., BabaAli, B., Povey, D., Riedhammer, K., Trmal, J., and Khudanpur, S. (2014). A pitch extraction algorithm tuned for automatic speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 2494–2498. IEEE.

Girshick, R. (2015). Fast R-CNN. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448. IEEE, dec.

Graves, A., Fernandez, S., Gomez, F., and Schmidhuber, J. (2006). Connectionist Temporal Classification : Labelling Unsegmented Sequence Data with Recurrent Neural Networks. *Proceedings of the 23rd international conference on Machine Learning*, pages 369–376.

Himmelmann, N. (2006). Language documentation: what is it and what is it good for? In J. Gippert, et al., editors, *Essentials of language documentation*, pages 1–30. de Gruyter, Berlin/New York.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., and Others. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE*, 29(6):82–97.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Jacobson, M., Michailovsky, B., and Lowe, J. B. (2001). Linguistic documents synchronizing sound and text. *Speech Communication*, 33:79–96.

Le, V.-B. and Besacier, L. (2009). Automatic speech recognition for under-resourced languages: application to Vietnamese language. *Audio, Speech, and Language Processing, IEEE Transactions on*, 17(8):1471–1482.

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

Metze, F., Sheikh, Z. A. W., Waibel, A., Gehring, J., Kilgour, K., Nguyen, Q. B., and Nguyen, V. H. (2013). Models of tone for tonal and non-tonal languages. *2013 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2013 - Proceedings*, pages 261–266.

Michailovsky, B., Mazaudon, M., Michaud, A., Guillaume, S., François, A., and Adamou, E. (2014). Documenting and researching endangered languages: the Pangloss Collection. *Language Documentation and Conservation*, 8:119–135.

Michaud, A. (2017). *Tone in Yongning Na: lexical tones and morphotonology*. Number 13 in Studies in Diversity Linguistics. Language Science Press, Berlin.

Mortensen, D. R., Littell, P., Bharadwaj, A., Goyal, K., Dyer, C., and Levin, L. (2016). PanPhon: A Resource for Mapping IPA Segments to Articulatory Feature Vectors. *Proceedings of COLING 2016, the 26th Interna-*

tional Conference on Computational Linguistics: Technical Papers, pages 3475–3484.

Müller, M., Stüker, S., and Waibel, A. (2017). Language Adaptive Multilingual CTC Speech Recognition. In Alexey Karpov, et al., editors, *Speech and Computer: 19th International Conference, SPECOM 2017, Hatfield, UK, September 12-16, 2017, Proceedings*, pages 473–482. Springer International Publishing, Cham.

Niebuhr, O. and Michaud, A. (2015). Speech data acquisition: the underestimated challenge. *KALIPHO - Kieler Arbeiten zur Linguistik und Phonetik*, 3:1–42.

Ramsundar, B., Kearnes, S., Riley, P., Webster, D., Konerding, D., and Pande, V. (2015). Massively Multitask Networks for Drug Discovery. feb.

Ruder, S. (2017). An Overview of Multi-Task Learning in Deep Neural Networks. jun.

Schuster, M. and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Gary F. Simons et al., editors. (2017). *Ethnologue: languages of the world*. SIL International, Dallas, twentieth edition edition.

Thieberger, N. (2016). Documentary linguistics: methodological challenges and innovatory responses. *Applied Linguistics*, 37(1):88–99.

Vu, N. T., Imseng, D., Povey, D., Motlicek, P., Schultz, T., and Bourlard, H. (2014). Multilingual deep neural network based acoustic modeling for rapid language adaptation. In *Proceedings of the 39th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7639–7643, Florence, Italy.

Xu, P. and Fung, P. (2013). Cross-lingual language modeling for low-resource speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 21(6):1134–1144.

Xu, H., Su, H., Ni, C., Xiao, X., Huang, H., Chng, E.-S., and Li, H. (2016). Semi-supervised and Cross-lingual Knowledge Transfer Learnings for DNN Hybrid Acoustic Models under Low-resource Conditions. In *Proceedings of the Annual Conference of the International Speech Communication Association, (INTERSPEECH)*, pages 1315–1319, San Francisco, USA.

10. Language Resource References

Cavar, D. and Cavar, M. and Cruz, H. (2016). *Chatino Speech Corpus Archive Dataset*. ELRA, ISLRN 557-415-504-956-6.

Michaud, A. and Latami, D. (2017a). *Housebuilding 2*. Pangloss Collection: http://lacito.vjf.cnrs.fr/pangloss/.

Michaud, A. and Latami, D. (2017b). *Yongning Na Corpus*. Pangloss Collection: http://lacito.vjf.cnrs.fr/pangloss/.