Phonemic transcription of low-resource tonal languages

Oliver Adams, Trevor Cohn, Graham Neubig, Alexis Michaud

Phonemic Transcription of Low-Resource Tonal Languages

Oliver Adams,[♠] Trevor Cohn,[♠] Graham Neubig,[♡] Alexis Michaud[♣]
[♠]Computing and Information Systems, The University of Melbourne, Australia
[♡]Language Technologies Institute, Carnegie Mellon University, USA
[♣]CNRS-LACITO, National Center for Scientific Research, France
oadams@student.unimelb.edu.au, tcohn@unimelb.edu.au,
gneubig@cs.cmu.edu, alexis.michaud@cnrs.fr

Abstract

Transcription of speech is an important part of language documentation, and 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 to 150 minutes. 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.

1 Introduction

Language documentation involves eliciting speech from native speakers, and transcription of these rich cultural and linguistic resources is an integral part of the language documentation process. However, transcription is very slow: it often takes a linguist between 30 minutes to 2 hours to transcribe and translate 1 minute of speech, depending on the transcriber's familiarity with the language and the difficulty of the content. This is a bottleneck in the standard 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).

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 leverage of linguists. The data collected in this semi-automated workflow can then be used as training data for further re-

Oliver Adams, Trevor Cohn, Graham Neubig and Alexis Michaud. 2017. Phonemic Transcription of Low-Resource Tonal Languages. In *Proceedings of Australasian Language Technology Association Workshop*, pages 53–60.

finement 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 resource-scarce language documentation setting. 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 drastically reducing the data requirements in this way, we make the use of automatic transcription technology more feasible in a language documentation setting.

We evaluate this approach on two tonal languages, Yongning Na and Eastern Chatino (Cruz and Woodbury, 2006; Michaud, 2017). 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. Eastern Chatino, spoken in Oaxaca, Mexico, has a richer tone set but both languages have extensive morphotonology. 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 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 (for details on the situation for Eastern Chatino, see Cruz (2011, 18-23)).

Though a significant amount of Chatino speech has been transcribed (Chatino Language Documentation Project, 2017), its rich tone system and opposing location on the globe 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 is available online.[1]

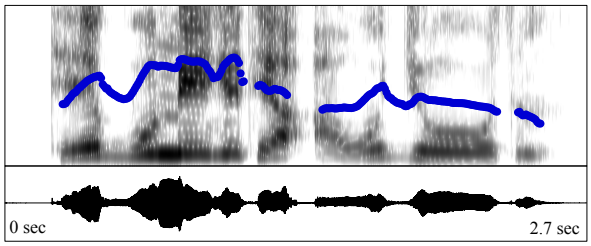
2 Model

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 features 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.

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. Figure 1 presents an example sentence from the Na corpus described in §3.1, along with an example of these three objectives.

[1] https://github.com/oadams/mam



/tʰiʌ̀ ∣ goꜜmiꜜ-dʑoꜛ ∣ tʰiʌ̀, ɑ˩ʁoꜜ dʑo˩ tsɯ˩ ∣ mɣ̩˩. ∣/
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 ɣ̩
2.	ʌ̀ ꜜ ꜜ ꛁ ʌ̀ ˩ ꜜ ˩ ˩
3.	tʰ i ʌ̀ g o ꜜ m i ꜜ d ʑ o ꛁ tʰ i ʌ̀ ɑ ˩ ʁ o ꜜ dʑ o ˩ t s ɯ ˩ m ɣ̩ ˩

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

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 150 minutes 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[2] (`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 important 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.

3.1 Data

We explore application of the model to the Na corpus 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 the recordings have been phonemically transcribed, and we used up to 149 minutes of this for training, 24 minutes for validation and 23 minutes for testing. The total number of phoneme and tone labels used for automatic transcription was 78 and 7 respectively.

For Chatino, we used data of Ćavar et al. (2016) from the GORILLA language archive for Eastern Chatino of San Juan Quiahije, Oaxaca, Mexico 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, 8 minutes for validation and 7 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 or tones such as tone group delimiters and hyphens connecting syllables within words.

4 Quantitative Results

Figure 2 shows the phoneme and tone error rates for Na and Chatino.

[2] 41 log Filterbank features along with their first and second derivatives

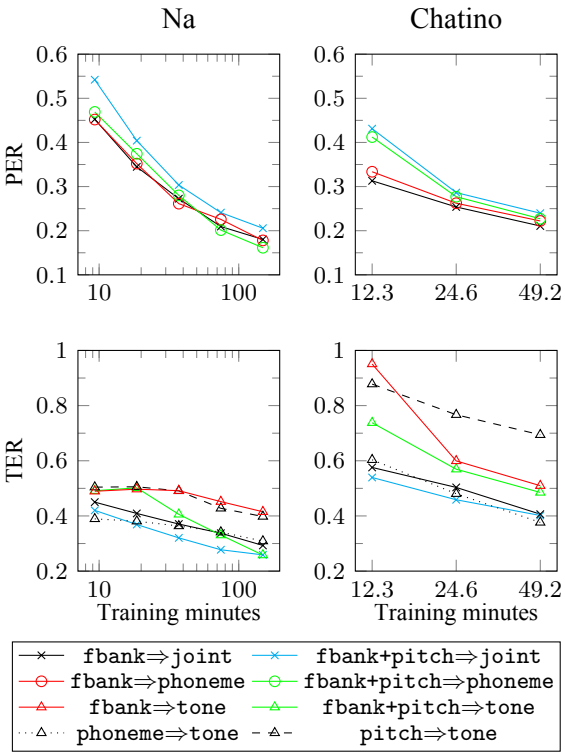


Figure 2: Phoneme error rate (PER) and tone error rate (TER) on test sets as training data is scaled for Na (left) and Chatino (right). The legend entries are formatted as ⟨input⟩ ⇒ ⟨output⟩ to indicate input features to the model and output target labels.

Error rate scaling Error rates decrease logarithmically with training data. The best methods reliably 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 these results suggesting how much linguists might need to transcribe before semi-automation can become part of their workflow.

In the case of phoneme-only prediction, use of pitch information does help reduce the PER, which is consistent with previous work (Metze et al., 2013).

Tonal modelling TER is always higher than PER for the same amount of training data, despite there being only 7 tone labels versus 78 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 the equal best performing 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 future work.

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 by using phoneme input features, since 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 outperform the best methods for separate phoneme and tone prediction, except in the case of Chatino tone prediction, if we discount `phoneme⇒tone`. 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.

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 more competitive in the case of Chatino, suggest-

	M	L	H	LH	MH
M	0	69.8	18.6	7.4	4.2
L	77	0	14.6	6.1	2.3
H	56.1	27.3	0	10.6	6.1
LH	38.6	31.8	25	0	4.5
MH	41.4	22.4	17.2	19	0

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

ing 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.

4.1 Error Types

Figure 3 shows the most common tone substitution mistakes for `fbank+pitch⇒joint` in the test set. Proportions were very similar for other methods. The most common tonal substitution errors were those between between M and L. Acoustically, M and L are neighbours; as mentioned above, in Na the same tone can be realised with a different pitch at different points in a sentence, leading to overlapping pitch ranges between these tones. Moreover, M and L tones were by far the most common tonal labels.

5 Qualitative Discussion

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

5.1 Recognition Errors

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. As a further example, the negation that is transcribed as /mõ˦/ in *Housebuilding2.290* instead of /mɤ˦/. This highlights that the vowel in that syllable is probably nazalised, and acoustically unlike the average /ɤ/ vowel for lexical words. The extent to which a word's morphosyntactic category influences the way it is pronounced 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.

5.2 Benefits for the Linguist

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

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, who is engrossed in a holistic process involving interpretation, translation, anno-

tation, and communication with the language consultant. 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.

Typographical errors and the transcriber's mindset Transcriptions are made during fieldwork with a language consultant and are difficult to correct down the line based only on auditory impression when the consultant is not available. However, such 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 entirely 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.

Speed Assessing automatic transcription's influence on the speed of the overall language documentation process is beyond the scope of this paper and is left to future work. Language documentation is a holistic process. Beyond phonemic transcription, documentation of Na involves other work that happens in parallel: translating, discussing with a native speaker, copying out new words into the Na dictionary, and being constantly on the lookout for new and unexpected linguistic phenomena. Further complicating this, the linguist'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 linguist 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 ASR 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 now available for subsequent Na acoustic modelling training.

As a first example of output by incorporating automatic transcription into the Yongning Na documentation workflow, transcription of the recording *Housebuilding* was completed using automatic transcription as a canvas; this document is now available online.[3]

6 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 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. 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. There is still much room for modelling improvement, including incorporation of linguistic constraints into the architecture for more accurate transcriptions.

References

Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. 2014. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication* 56:85–100.

Marc Brunelle, Daryl Chow, and Thụy Nhã Uyên Nguyễn. 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*. University of Glasgow, Glasgow, pages 1–5.

Lukáš Burget, Petr Schwarz, Mohit Agarwal, Pinar Akyazi, Kai Feng, Arnab Ghoshal, Ondřej Glembek, Nagendra Goel, Martin Karafiát, Daniel Povey, 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*. IEEE, pages 4334–4337.

[3] http://lacito.vjf.cnrs.fr/pangloss/corpus/show_text_en.php?id=crdo-NRU_F4_HOUSEBUILDING2_SOUND&idref=crdo-NRU_F4_HOUSEBUILDING2

Małgorzata E. Ćavar, Damir Cavar, and Hilaria Cruz. 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.

Ronan Collobert, Jason Weston, and Michael Karlen. 2011. Natural Language Processing (almost) from Scratch 1:1–34.

Emiliana Cruz. 2011. *Phonology, tone and the functions of tone in San Juan Quiahije Chatino*. Ph.D., University of Texas at Austin, Austin. http://hdl.handle.net/2152/ETD-UT-2011-08-4280.

Emiliana Cruz and Tony Woodbury. 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.

Li Deng, Geoffrey Hinton, and Brian Kingsbury. 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*. IEEE, pages 8599–8603. https://doi.org/10.1109/ICASSP.2013.6639344.

Thi-Ngoc-Diep Do, Alexis Michaud, and Eric Castelli. 2014. Towards the automatic processing of Yongning Na (Sino-Tibetan): developing a 'light' acoustic model of the target language and testing 'heavyweight' models from five national languages. In *4th International Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU 2014)*. St Petersburg, Russia, pages 153–160. https://halshs.archives-ouvertes.fr/halshs-00980431.

Yan-Mei Feng, Li Xu, Ning Zhou, Guang Yang, and Shan-Kai Yin. 2012. Sine-wave speech recognition in a tonal language. *The Journal of the Acoustical Society of America* 131(2):EL133–EL138. https://doi.org/10.1121/1.3670594.

Pegah Ghahremani, Bagher BabaAli, Daniel Povey, Korbinian Riedhammer, Jan Trmal, and Sanjeev Khudanpur. 2014. A pitch extraction algorithm tuned for automatic speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, pages 2494–2498.

Ross Girshick. 2015. Fast R-CNN. In *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, pages 1440–1448. https://doi.org/10.1109/ICCV.2015.169.

Alex Graves, Santiago Fernandez, Faustino Gomez, and Jurgen Schmidhuber. 2006. Connectionist Temporal Classification : Labelling Unsegmented Sequence Data with Recurrent Neural Networks. *Proceedings of the 23rd international conference on Machine Learning* pages 369–376. https://doi.org/10.1145/1143844.1143891.

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

Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, 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.

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

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

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

Florian Metze, Zaid A.W. W Sheikh, Alex Waibel, Jonas Gehring, Kevin Kilgour, Quoc Bao Nguyen, and Van Huy Nguyen. 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. https://doi.org/10.1109/ASRU.2013.6707740.

Boyd Michailovsky, Martine Mazaudon, Alexis Michaud, Séverine Guillaume, Alexandre François, and Evangelia Adamou. 2014. Documenting and researching endangered languages: the Pangloss Collection. *Language Documentation and Conservation* 8:119–135.

Alexis Michaud. 2017. *Tone in Yongning Na: lexical tones and morphotonology*. Number 13 in Studies in Diversity Linguistics. Language Science Press, Berlin. http://langsci-press.org/catalog/book/109.

David R Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori Levin. 2016. PanPhon: A Resource for Mapping IPA Segments to Articulatory Feature Vectors. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* pages 3475–3484. http://aclweb.org/anthology/C16-1328.

Markus Müller, Sebastian Stüker, and Alex Waibel. 2017. Language Adaptive Multilingual CTC Speech Recognition. In Alexey Karpov, Rodmonga Potapova, and Iosif Mporas, editors, *Speech and Computer: 19th International*

Conference, SPECOM 2017, Hatfield, UK, September 12-16, 2017, Proceedings, Springer International Publishing, Cham, pages 473–482. https://doi.org/10.1007/978-3-319-66429-3_47.

Bharath Ramsundar, Steven Kearnes, Patrick Riley, Dale Webster, David Konerding, and Vijay Pande. 2015. Massively Multitask Networks for Drug Discovery http://arxiv.org/abs/1502.02072.

Sebastian Ruder. 2017. An Overview of Multi-Task Learning in Deep Neural Networks http://arxiv.org/abs/1706.05098.

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

Gary F. Simons and Charles D. Fennig, editors. 2017. *Ethnologue: languages of the world*. SIL International, Dallas, twentieth edition edition. http://www.ethnologue.com.

Nick Thieberger. 2016. Documentary linguistics: methodological challenges and innovatory responses. *Applied Linguistics* 37(1):88–99. https://doi.org/10.1093/applin/amv076.

Ngoc Thang Vu, David Imseng, Daniel Povey, Petr Motlicek, Tanja Schultz, and Hervé Bourlard. 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)*. Florence, Italy, pages 7639–7643.

Haihua Xu, Hang Su, Chongjia Ni, Xiong Xiao, Hao Huang, Eng-Siong Chng, and Haizhou Li. 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)*. San Francisco, USA, pages 1315–1319.

Ping Xu and Pascale Fung. 2013. Cross-lingual language modeling for low-resource speech recognition. *IEEE Transactions on Audio, Speech and Language Processing* 21(6):1134–1144. https://doi.org/10.1109/TASL.2013.2244088.