Autosegmental Neural Nets: Should Phones and Tones be Synchronous or Asynchronous?

Jialu Li, Mark Hasegawa-Johnson

Department of Electrical and Computer Engineering & Beckman Institute, University of Illinois

{jialuli3, jhasegaw}@illinois.edu

Abstract

Phones, the segmental units of the International Phonetic Alphabet (IPA), are used for lexical distinctions in most human languages; Tones, the suprasegmental units of the IPA, are used in perhaps 70%. Many previous studies have explored cross-lingual adaptation of automatic speech recognition (ASR) phone models, but few have explored the multilingual and cross-lingual transfer of synchronization between phones and tones. In this paper, we test four Connectionist Temporal Classification (CTC)-based acoustic models, differing in the degree of synchrony they impose between phones and tones. Models are trained and tested multilingually in three languages, then adapted and tested cross-lingually in a fourth. Both synchronous and asynchronous models are effective in both multilingual and cross-lingual settings. Synchronous models achieve lower error rate in the joint phone+tone tier, but asynchronous training results in lower tone error rate.

Index Terms: tones, asynchronous training of tones and phones, under-resourced languages, IPA, CTC

1. Introduction

In most of the world's languages (possibly as many as 70% [1]), the meaning of a word depends on both phones and tones. Phones are called segmental because their acoustic cues occur in and around discrete temporal segments. Tones are called suprasegmental because each tone may be aligned over one or more segments. In Mandarin, for example, tones are canonically synchronized with the vowel and coda consonant of each syllable [2], but may influence the onset of the following syllable [3], and may be adopted, in an apparently rule-driven manner, as the pitch of a following neutral-tone syllable [4]. Similar rightward spreading occurs in many, but not all, tone languages [5]. In many languages, rightward spreading of a tone is not blocked by intervening vowels, consonants, or even syllables, but only by the intervention of another tone, suggesting that tones and phones are "autosegmental" (communicated as loosely-related segmentations of the time axis) [6, 7].

Most hidden Markov model-based (HMM-based) ASR in non-tonal languages uses one HMM per phone or triphone [8]. HMM-based ASR for tonal languages, by contrast, may use one HMM per final [9], per complete syllable [10], or per sequence of two to three syllables [11] so that the canonical domain of the lexical tone can be learned by the HMM. Localizing lexical tone on the vowel of each syllable is possible in a deep neural network (DNN)-HMM hybrid, apparently because the DNN captures sufficient acoustic context [12, 13].

End-to-end neural ASR, trained using CTC [14], can sidestep the tone-to-phone alignment problem by generating characters, rather than phones, as the output [15]. In a CTC system with character outputs, however, it is difficult to share data for multilingual [16] or cross-lingual [17] ASR. Proposed solu-

tions have included separate softmax tiers for the character set of each language [18, 19, 20], or the generation of phone strings instead of characters as the output of the CTC [21, 22, 23], or the use of both methods, in a multi-task learning framework, with one output tier generating phones, while another generates characters [24].

Mixed tones and phones using CTC have been demonstrated for Mandarin [25] and for two under-resourced tonal languages [26], but there have been few studies (if any) about multilingual or cross-lingual modeling of tone-marked-phones using CTC. Different tone languages seem to lend themselves to different temporal domains for the tone, e.g., Mandarin benefits from tone-marked finals [27] or syllables [25, 28], while ASR in other tone languages has used tone-marked vowels [26]. There is also some disagreement about whether tones and phones should be modeled jointly, or separately. For example, monolingual systems trained for the under-resourced tonal languages Na and Chatino found that both phone error rate (PER) and tone error rate (TER) are lower, in a CTC-based recognizer, if the phones and tones are modeled together, rather than separately, unless the recognizer is trained using at least 120 minutes of training data per language [26]. With at least 120 minutes of data, the results were mixed: the joint system gave lower TER but higher PER in Na, but the opposite result in Chatino.

This paper performs Multilingual and Cross-lingual recognition of phones and tones using end-to-end neural networks trained using CTC. Trained recognizers are tested on languages within the training set (Multilingual test), and adapted to a language with minimal adaptation data (Cross-lingual test). Four different systems are tested. The first system generates phone-marked tones as its output, similar to the joint transcription model of [26]. The second system has two separate output tiers, similar to the phones and characters of [24], but containing, instead, phones and tones. The third system combines the first two, with three output tiers. The fourth system is similar to the third, but standardizes tones across languages by forcing every tone, in every language, to have exactly two pitch targets.

The paper is organized as follows: Section 2 introduces the CTC-based acoustic models and cross-language adaptation methods in detail. Section 3 describes datasets and experimental methods. Section 4 provides results for each system, followed by analysis of multilingual and cross-lingual phone and tone modeling effects. Finally, Section 5 concludes.

2. Model

Four different end-to-end multilingual ASR systems were trained, using a CTC training criterion (Figure 1). All four systems used language-independent encoder networks (bLSTM×3+one fully-connected layer), followed by a language-dependent softmax layer. All four systems were designed to learn a mapping from acoustic inputs (x =
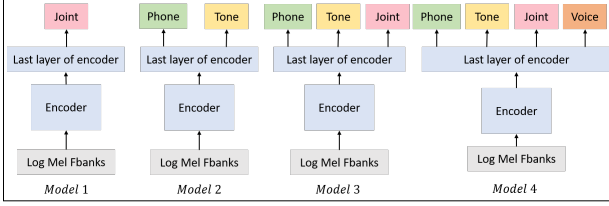
Figure 1: *CTC-based acoustic model with different multi-task learning tiers.*

$[x_1, \ldots, x_N]$) to sequences of phonetic label outputs arranged in one or more tiers. Let the reference label sequence in the tth tier be $\pi^{(t)} = [\pi_1^{(t)}, \ldots, \pi_m^{(t)}]$, where $\pi_j^{(t)} \in \mathcal{A}^{(t,l)}$ is a symbol in the tth sub-alphabet of the lth training language, and the length m is different in different tiers. Our systems are trained using a standard CTC training criterion [14] in each tier, which can be written as

$$\mathcal{L}^{(t)} = - \operatorname*{logsumexp}_{y^{(t)}:\mathcal{B}(y^{(t)})=\pi^{(t)}} \sum_{i=1}^{n} \log p(y_i^{(t)}|x_i), \qquad (1)$$

where $y^{(t)} = [y_1^{(t)}, \ldots, y_n^{(t)}]$, $y_i^{(t)} \in \mathcal{A}^{(t,l)} \cup \{b\}$, b is the blank symbol, and the operation $\mathcal{B}(y^{(t)})$ eliminates sequential duplicates, then eliminates blanks [14]. The four systems shown in Figure 1 differ only in the number and alphabets of their output tiers.

Model 1 has one output tier per training language, whose alphabet includes all consonant phonemes and all tone-marked vowel phonemes of the language. For example, the Mandarin-language softmax layer contains five variants of the vowel [a]: the vowel with neutral tone, and the vowel with four different lexical tones, as shown in the first column of Table 1.

Model 2 has two output tiers: phones and tones. The alphabet of the phone tier in each language is the set of its segmental phonemes. The alphabet of the tone tier is a universal phonetic tone inventory described in Section 2.2. Model 3 has three output tiers, exactly equal to the joint tier of model 1, and the phone and tone tiers of model 2. Model 4 standardizes the tone transcripts across languages, and extracts a separate voice quality tier, as described in Section 2.2. All four systems are trained using a multi-task loss function with equal weights for each tier, i.e., $\mathcal{L} = \sum_t \mathcal{L}^{(t)}$.

2.1. Cross-lingual phone transfer

Two types of experiments were performed in this study: Multilingual ASR (training and testing on different speech data from the same set of training languages), and Cross-lingual ASR (the fully trained model is adapted using limited data, then tested on the adaptation language). In order to initialize the fully-connected layers for the adaptation language, we adopted a strategy similar to [17, 29, 18, 30], based on knowledge-based cross-lingual mapping of IPA [31] symbols. The softmax layer of the adaptation language is initialized as follows: denote the dense layer weight matrix as $W^{(t,l)} \in \mathbf{R}^{(1+|\mathcal{A}^{(t,l)}|)\times d}$, where $\mathcal{A}^{(t,l)}$ is the alphabet of tier t in language l, 1 is the blank character, and d is the dimension of the hidden layer. For a target phone k in the adaptation language, if k exists in any training language, then the average over the corresponding entries of all training language weight matrices is used to initialize the adaptation language. If phone k exists in no training language, then

it is initialized, if possible, using a phone k' that is equal to k plus a diacritic, e.g., the phone [a] could be initialized by [aː]. If there is no such extension, then finally, k is initialized by a phone k' that is most similar according to the consonant or vowel features of the IPA chart [31], e.g., the vowel [ɤ] could be initialized by [o]. Similarly, for the joint tier, the closest phone is first located, then among the candidate tone-marked versions of that phone, the one with the closest tone is located, e.g., [uː˩˧] could be initialized using [uː˥˩], and [ɑʊ˩˧] could be initialized using [ɑʊ˩˥]. Once the closest phones among training languages have been identified, then the corresponding weights of the adaptation language are initialized as

$$W_k^{(t,l)} = \begin{cases} \frac{\sum_{l'=1}^{n} W_k^{(t,l')}}{\sum_{l'=1}^{n} \mathbb{1}_{k \in \mathcal{A}^{(t,l')}}} & \text{if } \sum_{l'=1}^{n} \mathbb{1}_{k \in \mathcal{A}^{(t,l')}} \neq 0 \\ \frac{\sum_{l'=1}^{n} W_{k'}^{(t,l')}}{\sum_{l'=1}^{n} \mathbb{1}_{k' \in \mathcal{A}^{(t,l')}}} & \text{else} \end{cases}$$

$$(2)$$

where n is the number of training languages, $W_k^{(t,l)}$ is the weight vector for phone k in tier t of language l, $\mathcal{A}^{(t,l)}$ is the corresponding alphabet, and $\mathbb{1}$ denotes the identity function.

2.2. Cross-lingual tone transfer

Lexical tone is suprasegmental: it is not necessarily time-aligned with any single phone segment. Standard IPA transcription methods list a lexical tone as a sequence of tone targets following the vowel, but it is not clear that synchronizing the tones in this way helps ASR. In order to explore possible asynchrony between tones and phones, Models 2, 3, and 4 use separate tone tier outputs. In order to further reduce synchronization requirements, the alphabet of the tone tier is not linked to the particular lexical tone inventory of each language: instead, the alphabet of this tier is language-independent, and consists of the five distinct IPA tone targets (extra high(˥), high(˦), mid(˧), low(˨), and extra low(˩)), the symbol ⟨neutral⟩ as a placeholder for a syllable with neutral or unmarked tone, and a ⟨boundary⟩ symbol to mark syllable boundaries. Models 2 and 3, but not Model 4, augment this alphabet with the symbols [ʔ] and [h], in order to correctly label the glottalized and breathy tones of Vietnamese.

Model 4 attempts some degree of cross-language standardization, in both the length and content of the tone targets in each syllable. Tone-tier training transcripts for Model 4 were normalized prior to training and testing, so that each syllable corresponds to exactly three characters: two tones, and a syllable boundary. Lexical tones that are canonically transcribed with three IPA symbols, like Mandarin tone 3 (˨˩˦), were truncated (˨˩). Tones that are usually transcribed with one target, including neutral tones and, e.g., Mandarin tone 1 (˥), were reduplicated (˥˥). Voice quality symbols in the canonical tone descriptions of Vietnamese ([ʔ] and [h]) were moved to a new voice quality tier, as were the corresponding phone segments in Lao. In order to maintain structure in the voice-quality transcripts, each syllable received at least one voice quality marker: either [ʔ], or [h], or a new modal-voicing symbol, ⟨modal⟩. The resulting tier alphabets for Model 4 are $\mathcal{A}^{(tone)} = \{˨, ˦, ˧, ˥, \langle\text{neutral}\rangle, \langle\text{boundary}\rangle\}$ and $\mathcal{A}^{(voice)} = \{ʔ, h, \langle\text{boundary}\rangle, \langle\text{modal}\rangle\}$.

3. Experimental methods

Sources of data, and quantities used for training, development, and test sets are listed in Table 2. In order to test Cross-lingual

Table 1: *Lexical tones and glottal phones that are part of the phoneme inventories of the four tonal languages used in this study: Mandarin, Cantonese, Vietnamese, and Lao.*

Man	Descriptions	Can	Descriptions	Viet	Descriptions	Lao	Descriptions
a˥	high	a˥	high	a˧	mid	a˨	low
a˧˥	mid-rising	a˧	mid	a˧˥	mid-rising	a˧	mid
a˨˩˦	low-dipping	a˨	low	a˧˨ʔ	mid-falling, glottalized	a˥	high
a˥˩	falling	a˧˥	mid-rising	a˨˩h	mid-falling, breathy	a˩˧	low-rising
		a˨˩	low-falling	a˨˩˨	low-falling	a˥˧	high-falling
		a˨˧	low-rising	a˧ʔ˥	mid-rising, stopped	a˧˩	mid-falling
						ʔ	glottal stop
						h	glottal fricative

Table 2: *Sources of data, and quantities (hours) used for Multilingual and Cross-lingual training, development, and testing.*

Setting	Language	Source	Train	Dev	Test
Multi	Mandarin	HUB4-NE	26.50	1.46	1.43
	Cantonese	BABEL	31.08	1.56	1.84
	Vietnamese	BABEL	18.24	1.54	1.39
Cross	Lao	BABEL	1	1	1

ASR, the Lao dataset was artificially restricted to just 1 hour for adaptation, 1 hour for development, and 1 hour for testing.

BABEL speech corpora consist of conversational and scripted data for each language; we used scripted data only because of its better audio quality. We found that conversational speech data often contains noise and long silences.

All experiments were performed using extracted 40-dimensional log Mel filterbank features, computed using the python speech features library [32], with a 25ms Hamming window and 10ms shift. Each feature dimension was Z-normalized per speaker. One additional experiment was performed with Model 1, in which its input feature vector was augmented by a fundamental frequency measurement (F0), because F0 has been shown to reduce ASR error rates for tonal languages [33, 34]. F0 was extracted from the same 25ms windowed frame, converted from Hertz to Mel scale, then appended to the 40-dimensional log Mel features. Model 1 was chosen for augmentation because it gave the lowest joint error rate in the Cross-lingual train/test condition, as described in Section 4.

IPA phone transcripts were created for each language using the LanguageNet Grapheme-to-Phoneme (G2P) transducers [35] implemented in Phonetisaurus [36] to generate IPA-based phonetic transcripts for each utterance. Vowels usually have tones associated with them, and consonants often don't have tones associated with them. We extracted each phone and its corresponding tone letters respectively as described in subsection 2.1 and subsection 2.2 to prepare for multitask learning in the acoustic modeling.

Models were implemented using the eXtensible Neural Machine Translation toolkit [37]. Three layers of pyramidal Bi-directional Long-Short Term Memory (pBLSTM) are used as the encoder. The hidden dimension of the fully-connected layer is $d = 512$; the input and hidden dimensions of the LSTM layer are 1024 and 256. The optimizer is Adadelta, with a learning rate of 0.004, and with early stopping using the development set to choose the best model. Decoding used a beam search with language modeling to obtain the best results on test set;

Table 3: *Phone error rates (PER), tone error rates (TER), joint phone and tone error rates (JER), and voice quality error rates (VER) in the Multilingual (trained and tested on different speech data from the same three languages), Cross-lingual (adapted using one hour), and Monolingual (trained using one hour) settings, in percent. M1+F0=Model 1 with both Mel filterbank and F0 input features. Lowest number in each column is bold.*

		Multilingual			Cross	Mono
		Man	Can	Viet	Lao	Lao
JER	Model1	55.73	45.95	53.45	54.36	83.81
	Model3	61.07	45.91	53.37	69.32	**82.36**
	Model4	60.22	46.13	53.49	81.72	84.93
	M1+F0	**55.35**	**40.31**	**48.91**	**53.26**	-
PER	Model2	59.88	47.02	55.51	**57.69**	**90.05**
	Model3	52.59	**39.97**	49.69	60.88	90.53
	Model4	**51.60**	40.34	**49.04**	77.97	90.74
TER	Model2	58.32	43.80	48.05	**44.34**	**79.01**
	Model3	62.34	39.19	44.59	46.88	82.52
	Model4	**52.09**	**39.02**	**33.91**	68.04	92.53
VER	Model4	-	-	37.08	75.11	90.42

beam width is 25 and the language modeling coefficient is 0.1.

For Cross-lingual adaptation, softmax output layers for Lao were (1) initialized as described in Section 2.1, (2) retrained without updating the adapted encoder, then (3) fine-tuned together with the encoder until convergence.

A Monolingual system was trained and tested as a baseline for Cross-lingual adaptation in Lao. The Monolingual baseline used the same architectures as the Cross-lingual systems, but when trained on 1 hour of data using the same number of parameters as the Cross-lingual system, it failed to converge. In order to achieve error rates below 100%, therefore, the parameter count of the Monolingual system was greatly reduced. The input and hidden dimensions of the LSTM layers, and the hidden layer dimension of the fully-connected layer, are reduced as necessary to minimize development-set error rates, resulting in dimensions of 2–4 nodes each.

4. Results and discussion

Table 3 shows error rates of all four models, and of Model 1 with both Mel filterbank and F0 inputs (M1+F0). Three experimental settings are distinguished: Multilingual (trained using 75.82 hours of Mandarin, Cantonese, and Vietnamese, tested

on a different 4.66 hours in the same languages), Cross-lingual (adapted using 1 hour of Lao, tested using 1 hour of Lao), and Monolingual (trained using 1 hour, tested using 1 hour). Cross-lingual training is better than Monolingual training, for all Models, and for all error metrics, but closer analysis reveals striking differences between the different test conditions. **MULTILINGUAL:** Joint phone+tone error rate (JER) is either lowest for Model 1 (Mandarin and Lao) or roughly comparable across Models 1, 3, and 4 (Cantonese and Vietnamese), but in all four languages, JER is significantly reduced by adding F0 to the input feature vector (M1+F0). Phone error rate (PER) and tone error rate (TER) are much worse in Model 2 than in Models 3 and 4. For Mandarin, JER, PER and TER are relatively higher for all four models. This is perhaps due to the noisy speech collected while reporters were interviewing outdoor and some code-switching utterances that the models failed to generate correct phonemes. **MONOLINGUAL:** TER is lowest in Model 2, suggesting that lexical tones in Lao may be best learned in isolation (without the joint tier), and JER is lowest in Model 3, suggesting that the joint phones+tones tier may be best learned in combination with the tones-only tier. **CROSS-LINGUAL:** the smaller the model is, the better, within the limits of the optimized-parameter-count systems shown in Table 3. JER is lowest in Model 1, while PER and TER are lowest in Model 2.

Model 4 has the lowest TER in the Multilingual setting, but its superiority may be caused by its lower cardinality: as described in Section 2.2, the tone tier of Model 4 has an output alphabet with only 6 output symbols (plus blank), while those of Models 2 and 3 both have output alphabets containing 9 output symbols (plus blank). Even if the superiority of Model 4 is discounted, however, the key finding of the TER section in Table 3 is unchanged. Model 3 has lower TER than Model 2 in the Multilingual case, but not in Monolingual or Cross-lingual. The key finding remains, therefore, that multi-task training of the tone tier together with a joint tier improves TER in the Multilingual setting, but not in the Monolingual or Cross-lingual settings.

Model 1's superior JER, in the Cross-lingual case, suggests an experiment in which PER and TER are measured using the phone and tone symbols produced by Model 1's joint output tier. Table 4 shows the PER and TER of phones and tones extracted from the joint output tiers of Models 1, 3, and 4. Table 4 also shows the consonant error rate (CoER) and vowel error rate (VoER) of consonants and vowels extracted from the joint tiers of Models 1, 3, and 4, and from the phone tier of Model 2. These error rates are computed by deleting all out-of-class symbols from both the reference and hypothesis transcripts, and then computing the string edit distance between reference and hypothesis (for example, CoER is computed by deleting all non-consonant symbols from both reference and hypothesis). This method usually gave PER and TER, for Models 3 and 4, that are worse than their corresponding results in Table 3. In order to facilitate comparison between the tables, therefore, the lowest PER or TER in each column of Table 4 is bold **only if** it is lower than the corresponding best entry in Table 3. As shown in Table 4, M1+F0 provides the lowest CoER and VoER in every language, but not always the best PER. Closer study shows that, without F0 inputs, Model 1 always provides the best consonant error rates, but not always the best vowel error rates. Tone behaves in a surprising manner. Without F0, none of the TER entries in Table 4 are lower than Table 3. Even with F0, the M1+F0 entry in Table 4 beats that of Table 3 for only one language. We conclude tentatively that consonants and vowels are

Table 4: *Consonant error rates (CoER), vowel error rates (VoER), phone error rates (PER), and tone error rates (TER) computed from joint tier in Models 1, 3 and 4 and phone tier in Model 2. Lowest number in each column is bold only if lower than the corresponding best result in Table 3.*

		Multilingual			Cross
		Man	Can	Viet	Lao
CoER	Model1	46.34	50.21	48.23	**39.43**
	Model2	64.22	68.41	52.94	72.95
	Model3	52.53	53.78	51.47	48.77
	Model4	52.00	55.37	48.84	60.77
	M1+F0	**46.12**	**46.61**	**45.19**	41.77
VoER	Model1	49.11	33.94	54.73	61.67
	Model2	53.00	39.75	64.04	**51.78**
	Model3	53.72	31.52	54.85	76.81
	Model4	56.84	32.25	65.43	91.46
	M1+F0	**48.93**	**28.33**	**50.00**	57.67
PER	Model1	**48.61**	41.12	56.33	50.98
	Model3	53.80	40.11	57.97	66.55
	Model4	55.65	40.30	59.39	77.78
	M1+F0	54.81	**34.72**	51.42	**48.62**
TER	Model1	55.35	43.02	51.42	68.31
	Model3	58.14	39.38	51.65	77.99
	Model4	54.79	40.16	53.08	91.83
	M1+F0	54.41	**37.80**	49.18	67.12

best recognized using an output tier that requires them to carry their tone markings (Model 1), but that tone is best recognized using a separate output tone tier (Model 4 in the Multilingual case, Model 2 in the case of Lao).

5. Conclusions

This experiment compared four methods for Multilingual and Cross-lingual CTC ASR of tones and phones. Cross-lingual results must be considered tentative, because only one language (Lao) was available as the target of Cross-lingual ASR; future work should repeat the Cross-lingual experiment for all four languages (or more), using a cross-validation training paradigm. Nevertheless, some results of this experiment seem very clear, and likely to be supported by future experimentation. Both synchronous (Model 1) and asynchronous (Models 2, 3, and 4) phones and tones can be adapted Cross-lingually, resulting in error rates far below those achieved by a Monolingual system trained on the same limited data. An output tier that requires tone-marking of every vowel results in lower joint error rates, as well as lower error rates for both consonants and vowels separately, than the systems that recognize phones and tones on separate output tiers. Conversely, tones are most accurately recognized using a system with separate phone and tone output tiers. The lowest tone error rates in the Multilingual case are provided by a multitask system with four output tiers (phone, tone, voice quality, and joint), while the lowest tone error rate for Cross-lingual ASR is provided by a system with two output tiers (phones and tones).

6. References

[1] M. Yip, *Tone*. Cambridge: Cambridge University Press, 2002.

[2] Y. Xu, "Consistency of tone-syllable alignment across different

syllable structures and speaking rates," *Phonetica*, vol. 55, no. 4, pp. 179–203, 1998.

[3] C. X. Xu, Y. Xu, and L.-S. Luo, "A pitch target approximation model for F0 contours in Mandarin," in *Proc. Interspeech*, 1999, pp. 2359–2362.

[4] M. Y. Chen, *Tone Sandhi Patterns Across Chinese Dialects*. Cambridge, UK: Cambridge University Press, 2000.

[5] L. M. Hyman and R. G. Schuh, "Universals of tone rules: Evidence from west Africa," *Linguistic Inquiry*, vol. 5, no. 1, pp. 81–115, 1974.

[6] J. A. Goldsmith, "Tone melodies and the autosegment," in *Proceedings of the 6th Conference on African Linguistics, Ohio State University Working Papers in Linguistics*. Columbus, OH: Ohio State University, 1975, pp. 135–147.

[7] ——, "Autosegmental phonology," Ph.D. dissertation, MIT, 1976.

[8] K.-F. Lee, "Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition," *IEEE Trans. on Acoustics, Speech, and Sig. Proc.*, vol. 38, 1990.

[9] C.-H. Lin, L.-S. Lee, and P.-Y. Ting, "A new framework for recognition of Mandarin syllables with tones using sub-syllabic units," in *Proc. ICASSP*, vol. II, 1993, pp. 227–230.

[10] T. Lee, W. Lau, Y. Wong, and P. Ching, "Using tone information in Cantonese continuous speech recognition," *ACM Transactions on Asian Language Information Processing*, vol. 1, no. 1, pp. 83–102, 2002.

[11] Y. Qian, T. Lee, and F. K. Soong, "Tone recognition in continuous Cantonese speech using supratone models," *J. Acoust. Soc. Am.*, vol. 121, no. 5, pp. 2936–2945, 2007.

[12] F. Metze, Z. A. W. Sheikh, A. Waibel, J. Gehring, K. Kilgour, Q. B. Nguyen, and V. H. Nguyen, "Models of tone for tonal and non-tonal languages," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, 2013, pp. 261–266.

[13] Van Huy Nguyen, Chi Mai Luong, and Tat Thang Vu, "Tonal phoneme based model for Vietnamese LVCSR," in *2015 International Conference Oriental COCOSDA held jointly with 2015 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, 2015, pp. 118–122.

[14] A. Graves, S. Fernández, and F. Gomez, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the International Conference on Machine Learning, ICML 2006*, 2006, pp. 369–376.

[15] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng, "Deep speech: Scaling up end-to-end speech recognition," arXiv, Tech. Rep. 1412.5567, 2014.

[16] K. Vesely, M. Karafiát, F. Grezl, M. Janda, and E. Egorova, "The language-independent bottleneck features," in *Proceedings of SLT*, 2012.

[17] W. Byrne, P. Beyerlein, J. M. Huerta, S. Khudanpur, B. Marthi, J. J. Morgan, N. Peterek, J. Picone, D. Vergyri, and W. Wang, "Towards language independent acoustic modeling," in *Proc. ICASSP*, 2000, pp. 1029–1032.

[18] S. Tong, P. N. Garner, and H. Bourlard, "Multilingual training and cross-lingual adaptation on CTC-based acoustic model," *Speech Communication*, vol. 104, pp. 39–46, 2018.

[19] J. Cho, M. K. Baskar, R. Li, M. Wiesner, S. H. Mallidi, N. Yalta, M. Karafiát, S. Watanabe, and T. Hori, "Multilingual sequence-to-sequence speech recognition: Architecture, transfer learning, and language modeling," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 521–527.

[20] H. Inaguma, J. Cho, M. K. Baskar, T. Kawahara, and S. Watanabe, "Transfer learning of language-independent end-to-end ASR with language model fusion," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6096–6100.

[21] D. He, X. Yang, B. P. Lim, Y. Liang, M. Hasegawa-Johnson, and D. Chen, "When CTC training meets acoustic landmarks," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5996–6000.

[22] H. Sak, A. Senior, K. Rao, O. İrsoy, A. Graves, F. Beaufays, and J. Schalkwyk, "Learning acoustic frame labeling for speech recognition with recurrent neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4280–4284.

[23] Y. Miao, M. Gowayyed, X. Na, T. Ko, F. Metze, and A. Waibel, "An empirical exploration of CTC acoustic models," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 2623–2627.

[24] O. Adams, M. Wiesner, S. Watanabe, and D. Yarowsky, "Massive multilingual adversarial speech recognition," in *2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 2019.

[25] Z. Qu, P. Haghani, E. Weinstein, and P. Moreno, "Syllable-based acoustic modeling with CTC-SMBR-LSTM," in *Proc. ICASSP*, 2017, pp. 173–177.

[26] O. Adams, T. Cohn, G. Neubig, and A. Michaud, "Phonemic transcription of low-resource tonal languages," in *Proceedings of Australasian Language Technology Association Workshop*, 2017, pp. 53–60.

[27] S. Zhang, M. Lei, Y. Liu, and W. Li, "Investigation of modeling units for Mandarin speech recognition using DFSMN-CTC-sMBR," in *Proc. ICASSP*, 2019, pp. 7085–7089.

[28] Y. Zhao, L. Dong, S. Xu, and B. Xu, "Syllable-based acoustic modeling with CTC for multi-scenarios Mandarin speech recognition," in *International Joint Conference on Neural Networks (IJCNN)*, 2018, pp. 1–8.

[29] O. Scharenborg, F. Ciannella, S. Palaskar, A. Black, F. Metze, L. Ondel, and M. Hasegawa-Johnson, "Building an ASR system for a low-resource language through the adaptation of a high-resource language ASR system: Preliminary results," in *Proc. Internat. Conference on Natural Language, Signal and Speech Processing (ICNLSSP)*, 2017, pp. 26–30.

[30] X. Li, S. Dalmia, J. Li, M. Lee, P. Littell, J. Yao, A. Anastasopoulos, D. R. Mortensen, G. Neubig, A. W. Black, and F. Metze, "Universal phone recognition with a multilingual allophone system," in *Proc. ICASSP*, 2020, pp. 8249–8253.

[31] International Phonetic Association, Ed., *Handbook of the International Phonetic Association*. Cambridge: Cambridge University Press, 1999.

[32] J. L. Huang and et al., "jameslyons/python_speech_features: release v0.6.1," January 2020, Web Download.

[33] X. Lei, M. Siu, M.-Y. Hwang, M. Ostendorf, and T. Lee, "Improved tone modeling for Mandarin broadcast news speech recognition," in *Interspeech*, 2006.

[34] S. Li, Y. Wang, L. Sun, , and L. Lee, "Improved tonal language speech recognition by integrating spectro-temporal evidence and pitch information with properly chosen tonal acoustic units," in *Interspeech*, 2011.

[35] M. Hasegawa-Johnson, "LanguageNet grapheme-to-phoneme transducers," 2020, downloaded 5/15/2020 from https://github.com/uiuc-sst/g2ps.

[36] J. Novak, P. Dixon, N. Minematsu, K. Hirose, C. Hori, and H. Kashioka, "Improving WFST-based G2P conversion with alignment constraints and RNNLM n-best rescoring," in *Interspeech*, 2012.

[37] G. Neubig, M. Sperber, X. Wang, M. Felix, A. Matthews, S. Padmanabhan, Y. Qi, D. S. Sachan, P. Arthur, P. Godard, J. Hewitt, R. Riad, and L. Wang, "XNMT: The extensible neural machine translation toolkit," in *Conference of the Association for Machine Translation in the Americas (AMTA) Open Source Software Showcase*, Boston, March 2018.