

UNIVERSAL PHONE RECOGNITION WITH A MULTILINGUAL ALLOPHONE SYSTEM

[†]Xinjian Li [†]Siddharth Dalmia [†]Juncheng Li
[°]Matthew Lee [△]Patrick Littell [□]Jiali Yao [†]Antonios Anastasopoulos
[†]David R. Mortensen [†]Graham Neubig [†]Alan W Black [†]Florian Metze

[†]Carnegie Mellon University; [°]SIL International;
[△]National Research Council Canada; [□]ByteDance AI Lab

xinjianl@cs.cmu.edu

ABSTRACT

Multilingual models can improve language processing, particularly for low resource situations, by sharing parameters across languages. Multilingual acoustic models, however, generally ignore the difference between phonemes (sounds that can support lexical contrasts in a *particular* language) and their corresponding phones (the sounds that are actually spoken, which are language independent). This can lead to performance degradation when combining a variety of training languages, as identically annotated phonemes can actually correspond to several different underlying phonetic realizations. In this work, we propose a joint model of both language-independent phone and language-dependent phoneme distributions. In multilingual ASR experiments over 11 languages, we find that this model improves testing performance by 2% phoneme error rate absolute in low-resource conditions. Additionally, because we are explicitly modeling language-independent phones, we can build a (nearly-)universal phone recognizer that, when combined with the PHOIBLE [1] large, manually curated database of phone inventories, can be customized into 2,000 language dependent recognizers. Experiments on two low-resourced indigenous languages, Inuktitut and Tusom, show that our recognizer achieves phone accuracy improvements of more than 17%, moving a step closer to speech recognition for all languages in the world.¹

Index Terms— multilingual speech recognition, universal phone recognition, phonology

1. INTRODUCTION

There is an increasing interest in building speech tools benefiting low-resource languages, specifically multilingual models that can improve low-resource recognition using rich resources available in other languages like English and Mandarin. One standard tool for recognition in low resource languages is multilingual acoustic modeling [2]. Acoustic models are generally trained on parallel data of speech waveforms and *phoneme* transcriptions. Importantly, phonemes are perceptual units of sound that closely correlate with, but do not exactly correspond to the actual sounds that are spoken, *phones*. An example of this is shown in Figure 1, which demonstrates two English words that share the same phoneme /p/, but different in the actual phonetic realizations [p] and [p^h]. *Allophones*,

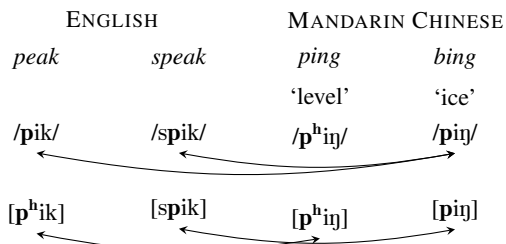


Fig. 1: Words, phonemes (slashes), and phones (square brackets).

the sets of phones that correspond to a particular phoneme, are language specific; distinctions that are important in some languages are not important in others.

Most multilingual acoustic models simply use existing phoneme transcriptions as-is, taking the union of the phoneme sets to be shared by all training languages [3, 4, 5, 6, 7]. The assumption is reasonable under some circumstances as phoneme names are typically associated with their most common or least marked allophone. However, this is obviously an over-simplistic view: in Figure 1, for example, this would mean that all training in English would assign the phones [p] and [p^h] to phoneme /p/. This is detrimental if we want to recognize Mandarin Chinese, for instance, where the two phones are corresponding to two distinctive phonemes /p/ and /p^h/.

In this paper, we propose a novel method for multilingual recognition based on phonetic annotation to tackle this problem: *Allosaurus* (**al**lophone system of **au**tomatic **re**cognition for **u**niversal speech). Our method incorporates knowledge of phonology into the multilingual model through an *allophone layer*, which associates a universal narrow phone set with the phonemes that appear in the transcription of each language. Our model first computes the phone distribution using a standard ASR encoder, then the allophone layer maps the phone distribution into the phoneme distribution for each language. This model can be trained end-to-end using only standard phonemic transcriptions and an allophone list created by phoneticians. The allophone layer is first initialized with the allophone list, then is further optimized during the training process. We demonstrate that accounting for the phoneme-phone mismatch in this way improves the accuracy of multilingual acoustic modeling by 2.0% phoneme error rate in low-resource conditions.

Furthermore, the architecture simultaneously makes it possible to perform *universal phone recognition*. Previous approaches can-

¹A web demo is available at <https://www.dictate.app>, the pre-trained model will be released at <https://github.com/xinjianl/allosaurus>

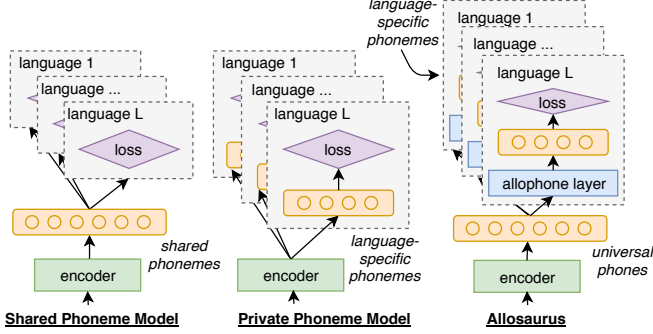


Fig. 2: Traditional approaches predict phonemes directly, either for all languages (left) or separately for each language (middle). On the contrary, our approach (right) predicts over a shared phone inventory, then maps into language-specific phonemes with an allophone layer.

not perform phone recognition in a universal fashion as they depend on language-specific phonemes, as illustrated with the previous example of English not distinguishing /p/ and /p^h/ as required in Mandarin. In contrast, because our approach allows recognition of phones directly, it already has learned to make these fine-grained distinctions. Taking advantage of this fact, we incorporate a large phone inventory database collected by linguists, PHOIBLE [1], and demonstrate that our phone recognizer can be customized to recognize over 2000 languages without any training data in the languages themselves. By evaluating the recognizer with completely unseen testing languages, we found that our recognizer achieves 17% better performance absolute compared with the traditional approach.

2. RELATED WORK

While some recent work in multilingual ASR focuses on end-to-end models to directly predict graphemes [8, 9], most systems still depend on phonetically inspired acoustic models. Multilingual acoustic models fall into two groups. The first group, *shared phoneme models*, creates a shared phoneme inventory of all phonemes from all training languages [3, 4, 5, 6, 7, 10]. The second group, *private phoneme models*, treats phonemes from each language as completely different classes performs phoneme classification separately for each language [11, 2, 12]. However, these two groups have their own respective drawbacks: the first group fails to consider the disconnect between the phonemes across languages while the second group completely ignores cross-lingual phonetic associations and is not applicable to recognition of new languages. In contrast, our approach solves both of these problems by taking into account allophones with phone-phoneme mappings.

There have been some attempts to apply phone recognizers to low resource languages. For example, English recognizers have been applied to align transcription corpora of an endangered language [13], facilitate language documentation [14], identify languages with language models [15], and perform linguistic annotation [16]. However, these approaches depend heavily on training data in the language of interest and their specific phonemic transcriptions. Our approach, on the other hand, abstracts away the dependency to phonemes by applying the allophone transformations.

3. APPROACH

3.1. Phone-Phoneme Annotation

Suppose there are $|L|$ training languages, and each language L_i has its own phoneme inventory Q_i which can be easily obtained by enumerating the phonemes appearing in its annotated training data. Most traditional multilingual approaches handle inventories at the phoneme level, and create a *shared phoneme inventory* Q_{sha} by taking union of the phoneme sets:

$$Q_{\text{sha}} = \bigcup_{1 \leq i \leq |L|} Q_i \quad (1)$$

In contrast, our method distinguishes phonemes from their phone realizations. We have linguists annotate each phoneme $q \in Q_i$ with its corresponding allophone set P_q^i , where each phone $p \in P_q^i$ is a realization of q in language L_i . Merging these sets for all languages, we obtain the *universal phone inventory* P_{uni} .

$$P_{\text{uni}} = \bigcup_{1 \leq i \leq |L|} \bigcup_{q \in Q_i} P_q^i \quad (2)$$

Additionally, we obtain a *signature matrix* $S^i = \{0, 1\}^{|Q_i| \times |P_{\text{uni}}|}$ describing the association of phone and phonemes in each language L_i : Suppose the phoneme $q \in Q_i$ has the row index j where $1 \leq j \leq |Q_i|$, phone $p \in P_{\text{uni}}$ has the column index k where $1 \leq k \leq |P_{\text{uni}}|$, if the p is a realization of q , then (j, k) cell of the S^i has a value of 1, otherwise it is assigned 0.

3.2. Allophone Layer

As mentioned in Section 2, traditional multilingual models can be divided into two groups. The first group, *shared phoneme models* (Figure 2 left), predicts phoneme distributions over the shared phoneme inventory Q_{sha} . The second group, *private phoneme models* (Figure 2 middle), on the other hand, shares a common encoder but computes distribution over private phoneme inventory Q_i for each language L_i . These approaches handle phonemes directly with no concept of underlying phones.

In contrast our proposed approach, *Allosaurus*, (Figure 2 right), comprises a language independent encoder and phone predictor, and a language dependent allophone layer and a loss function associated with each language. The encoder first produces the distribution $h \in \mathbb{R}^{|P_{\text{uni}}|}$ over the universal phone inventory P_{uni} , then the allophone layer transforms h into phoneme distribution $g^i \in \mathbb{R}^{|Q_i|}$ of each language. The allophone layer uses a trainable allophone matrix $W^i \in \mathbb{R}^{|Q_i| \times |P_{\text{uni}}|}$ to describe allophones in the similar way as S^i . The allophone matrix W^i is first initialized with S^i , and is allowed to be optimized during the training process, but we add an L2 penalty to penalize divergence from the original signature matrix S^i . The allophone layer computes its logit distribution g^i by finding the most likely allophone realization in P_{uni} with maxpooling.

$$g_j^i = \max(\{w_{j,k}^i \cdot h_k; 1 \leq k \leq |P_{\text{uni}}|\}), \quad (3)$$

where $g_j^i \in \mathbb{R}$ is the logit of j -th phoneme in g^i for language L_i , $w_{j,k}^i \in \mathbb{R}$ is the (j, k) cell of the allophone matrix W^i , $h_k \in \mathbb{R}$ is the logit of k -th phone in h . Intuitively, if the j -th phoneme has the k -th phone as an allophone, $w_{j,k}^i$ would be near 1, otherwise $w_{j,k}^i$ would be near 0. Therefore, the phoneme logit of g_j^i is decided by the largest allophone logit h_k . The phoneme distribution g^i is further fed into the loss function. This method for phoneme prediction can be used with any underlying multilingual ASR system. Here we

Table 1: Results of three models’ phoneme error rate performance on 11 languages. The top-half shows the results trained with all training datasets. The bottom-half shows the low-resource results in which only 1k utterances are used for training from each dataset.

		Amh	Eng	Ger	Ita	Jap	Man	Rus	Spa	Tag	Tur	Vie	Average
Full	Shared Phoneme PER	78.4	71.7	71.6	62.9	65.9	76.5	76.9	62.6	74.1	76.6	82.7	73.8
	Private Phoneme PER	37.1	22.4	17.6	26.2	17.6	17.9	21.3	18.5	47.6	35.8	56.5	25.6
	Allosaurus PER	36.0	20.5	18.8	23.7	23.8	17.0	26.3	19.4	57.4	35.3	57.3	25.0
Low	Shared Phoneme PER	80.4	73.3	74.3	72.2	77.1	83.0	83.2	72.8	84.8	84.4	84.5	78.4
	Private Phoneme PER	55.4	50.6	41.9	31.6	36.8	37.0	47.9	36.7	62.3	54.5	73.6	43.8
	Allosaurus PER	54.8	47.0	41.5	37.4	40.5	33.4	45.0	35.9	70.1	53.6	72.5	41.8

Table 2: Training corpora and size in utterances for each language. Models are trained and tested with 12 rich resource languages (top) and 2 low resource unseen languages (bottom).

Language	Corpora	Utt.
English	voxforge, Tedlium [17], Switchboard [18]	1148k
Japanese	Japanese CSJ [19]	440k
Mandarin	Hkust [20], openSLR [21, 22]	377k
Tagalog	IARPA-babel106b-v0.2g	93k
Turkish	IARPA-babel105b-v0.4	82k
Vietnamese	IARPA-babel107b-v0.7	79k
German	voxforge	40k
Spanish	LDC2002S25	32k
Amharic	openSLR25 [23]	10k
Italian	voxforge	10k
Russian	voxforge	8k
Inuktitut	private	1k
Tusom	private	1k

specifically optimize the parameters by minimizing CTC loss [24] for all training languages, with the addition of regularization of the allophone layer controlled by hyperparameter α .

$$\mathcal{L} = \sum_{1 \leq i \leq |L|} (\mathcal{L}_{ctc}^i + \alpha \|W^i - S^i\|_2^2). \quad (4)$$

3.3. Universal Phone Recognition

Not only does the allophone layer abstract away from the language-specific phonemes, which contributes to the improvement in the multilingual acoustic modeling, the model also gives us the capability to predict universal phones themselves. This has rarely been attempted in previous work. By applying the greedy decoding strategy over the phone distribution h , we can obtain a phone sequence in which all phones P_{uni} in the training languages are candidates. When combined with a large training languages sets, our universal inventory is expected to cover most common narrow phones appearing in many languages in the world, which we show in the experiment section.

Furthermore, this recognition protocol can take into account phone inventories that have already been created for many languages in the world by linguists. For example, PHOIBLE [1] is a database of phone inventories for more than 2000 languages and dialects, allowing our model to be applied to these languages with some degree of accuracy. If the phone inventory for language L_i is P_i , we can restrict the decoder to only produce phones in $P_i \cap P_{\text{uni}}$ by filtering out other phones. When the universal inventory P_{uni}

covers most frequent phones in the world, we could expect that $P_i \approx P_i \cap P_{\text{uni}}$.

4. EXPERIMENTS

4.1. Settings

As we are interested in creating a large universal phone inventory, we select a phonetically diverse set of 11 training languages as summarized on the top of Table 2. We include corpora from a variety of speech domains to make our model robust (e.g., read speech, spontaneous speech). 5% of the dataset is used as the test set, and the remaining data are used as the training set and the validation set. We also consider a low resource condition, where 1,000 random utterances are used from each corpus to train the model. As baselines, we compare with the previously-described *shared phoneme* and *private phoneme* models. All methods use the same encoder and features. Features are high-resolution 40 dimensional MFCCs extracted with Kaldi [25]. The encoder is a 6-layer stacked bidirectional LSTM with hidden size of 1024 in each layer. The regularization hyperparameter α is set to 10. Phonemes for training languages are assigned using the grapheme-to-phoneme tool Epitran [26]. For each phoneme in each language, phoneticians (mostly authors of this paper) create the allophone mappings.²

We evaluate using phoneme error rate for the training languages. Furthermore, we select two languages not included in the training data: Inuktitut and Tusom. These languages are indigenous languages with few training resources, representing a realistic scenario where our model is applied to entirely new languages, as may be done when ASR is used for documentation of endangered languages. The datasets of these two languages are transcribed with phones, and accordingly we use phone error rate rather than the phoneme error rate. While Allosaurus is able to predict phones in a natural way by decoding h , the two baselines could not predict phones directly. In this unseen language experiment, we assume phonemes predicted by the baselines correspond to phones of the same name.

4.2. Main Results

Table 1 demonstrates the performance of the baseline models and Allosaurus evaluated on 11 languages. The top half of the table summarizes the performance when trained with the full training set. The results suggests both the private phoneme model and the Allosaurus model outperforms the shared phoneme model significantly. The results of the shared phoneme model can be explained by the disagreement of phoneme assignments across languages. In contrast, the pri-

²This work has been accepted to LREC 2020 and its database is available at <https://github.com/dmort27/allovera>

Table 3: Statistics of the phone coverage mean (standard deviation) of areas. Phone coverage of language L_i is defined as $\frac{|P_{uni} \cap P_i|}{|P_i|}$

Area	# Language	Shared	Allosaurus
Africa	875	53% (13%)	84% (11%)
America	659	52% (14%)	81% (13%)
Asia	377	46% (15%)	79% (13%)
Pacific	152	59% (15%)	87% (12%)
Europe	92	35% (9.5%)	69% (13%)
All	2155	52% (15%)	82% (13%)

Table 4: Comparisons of phone error rates in two unseen languages

	Inuktitut	Tusom
Shared Phoneme PER	94.1	93.5
Best Private Phoneme PER	86.2	85.8
Allosaurus PER	84.1	77.3
Allosaurus+PHOIBLE PER	73.1	64.2

vate phoneme model handles this issue by using language specific phoneme layers. Our model also circumvents this issue by introducing the language-specific allophone layers. The bottom half of the Table 1 highlights the results when the training set of each language is limited as mentioned above. Unsurprisingly, limiting the amount of training data hurts accuracy across the board. While the private phoneme model and our model achieve similar results when using the full training set, our model outperforms the private phoneme model by 2.0% when training data is limited. This suggests that our model is better at sharing parameters across languages by using prior phonetic knowledge in this case, likely due to the fact that the private phoneme model needs to learn each phoneme predictor from scratch, while our model already has phone-phoneme mapping knowledge seeded by linguistically motivated annotations.

4.3. Universal Phone Recognition Results

In addition to the improvements over low resource settings, our model enables us to predict (nearly-)universal phone distributions. By merging phone inventories from all of our languages, we obtain a shared inventory of 187 phones. First, we assess how close this inventory gets to covering the languages registered in PHOIBLE. The Allosaurus column in Table 3 summarizes the phone coverage of our model, split into different geographic areas. The phone coverage in each cell represents the mean and standard deviation for each category. As the table suggests, our model has a promising phone coverage over all areas consistently. On average, it has 82% mean phone coverage and 12.8% standard deviation over all PHOIBLE languages. Furthermore, by comparing our model with the baseline model in which we merged all the phoneme inventories from the corpus as-is, we significantly improve the phone coverage by 30%. Additionally, the standard deviation shows that our model covers phones more consistently than the baseline model.

Next, we actually evaluate the model with respect to its ability to recognize phones. Table 5 shows a decoded English example. The utterance contains three English phonemes /p/ in word *people* and *speak*. The underlying allophones, however, are [p^h] and [p] as mentioned in Section 1. While the original English training set

Table 5: An English example from switchboard in which Allosaurus could distinguish [p^h] and [p] for phoneme /p/

Model	Phones
Utterance	the quebec people that that speak french
Annotation	/ð ə k w ə b ɛ k p i p ə l . s p i k f ɪ ɛ n tʃ/
Allosaurus	[ð ə x o b ə k ɛ p ^h i θ o : l . s p ɪ k f ɪ ɛ n d]

Table 6: A qualitative example from Inuktitut dataset

Model	Phones
Ground Truth	[i l i t s i l : i]
Allosaurus	[e l e p r i l : e]
Allosaurus+PHOIBLE	[i l i t i l : i]

annotates those two words with the same phoneme /p/, Allosaurus is able to predict different allophones by leveraging knowledge from other languages (e.g: Mandarin). We also note that Allosaurus is still not perfect: it fails to recognize the second /p/ in “people.”

Additionally we also investigate unseen languages on the Inuktitut and Tusom datasets. The results are summarized in the Table 4. As the result show, the shared phoneme model can hardly recognize any phonemes in these two languages, with more than 90.0% phone error rate on both datasets. Next, we try all 11 private phoneme models from the training datasets and use the one with the lowest phoneme error rate. Unsurprisingly, this also can not achieve satisfying results on both datasets, as none of our 11 languages is similar to Inuktitut and Tusom; they both have over 85.0% phone error rate. On the other hand, the proposed Allosaurus model achieves 84.1% phone error rate on Inuktitut and 77.3% phone error rate on Tusom, a significant drop. When combined with the PHOIBLE inventory, the error rates are further improved to 73.1% and 64.2% respectively, which shows 17% improvements on average over the shared phoneme baseline. Table 6 shows one qualitative example from Inuktitut data. It suggests that simply applying Allosaurus could capture some aspects of the target phonemes, but it still made many errors especially substitution errors between [e] and [i]. The reason is Allosaurus has a much broader phone search space (187 phones), it might be difficult to distinguish similar phones (e.g: both [e] and [i] are front vowels, but [e] is a close vowel and [i] is a close-mid vowel). We find those substitution errors account for the majority of errors in the test sets. Those confusing phones, however, might be solved when combined with an appropriate inventory such as PHOIBLE. The last row suggests that Allosaurus could fix those substitution errors as [e] does not exist in Inuktitut’s inventory.

5. CONCLUSION

In this work, we propose *Allosaurus*, which considers the relationship between phones and phonemes in multilingual acoustic modeling. It improves significantly the phone recognition accuracy over unseen languages by 17%.

6. ACKNOWLEDGMENT

This work is supported by NSF awards ACI-1548562 and 1761548. We would like to thank Alexis Michaud, Steven Abney, Hilaria Cruz and other participants of the LTLDR workshop for their feedback.

7. REFERENCES

- [1] Steven Moran and Daniel McCloy, Eds., *PHOIBLE 2.0*, Max Planck Institute for the Science of Human History, Jena, 2019.
- [2] Siddharth Dalmia, Ramon Sanabria, Florian Metze, and Alan W Black, “Sequence-based multi-lingual low resource speech recognition,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4909–4913.
- [3] Hui Lin, Li Deng, Dong Yu, Yi-fan Gong, Alex Acero, and Chin-Hui Lee, “A study on multilingual acoustic modeling for large vocabulary asr,” in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 4333–4336.
- [4] Paul S. Cohen, Satyanarayana Dharanipragada, Jerneja Zganec Gros, Michael Daniel Monkowski, Chalapathy Neti, Salim Roukos, and Todd Ward, “Towards a universal speech recognizer for multiple languages,” in *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*. IEEE, 1997, pp. 591–598.
- [5] Tanja Schultz and Alex Waibel, “Language-independent and language-adaptive acoustic modeling for speech recognition,” *Speech Communication*, vol. 35, no. 1-2, pp. 31–51, 2001.
- [6] Tanja Schultz and Alex Waibel, “Fast bootstrapping of lvsr systems with multilingual phoneme sets,” in *Fifth European Conference on Speech Communication and Technology*, 1997.
- [7] Xinjian Li, Siddharth Dalmia, David R Mortensen, Juncheng Li, Alan W Black, and Florian Metze, “Towards zero-shot learning for automatic phonemic transcription,” in *Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020.
- [8] Shinji Watanabe, Takaaki Hori, and John R Hershey, “Language independent end-to-end architecture for joint language identification and speech recognition,” in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 265–271.
- [9] Shubham Toshniwal, Tara N Sainath, Ron J Weiss, Bo Li, Pedro Moreno, Eugene Weinstein, and Kanishka Rao, “Multilingual speech recognition with a single end-to-end model,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4904–4908.
- [10] Jessica AF Thompson, Marc Schönwiesner, Yoshua Bengio, and Daniel Willett, “How transferable are features in convolutional neural network acoustic models across languages?,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2827–2831.
- [11] Jui-Ting Huang, Jinyu Li, Dong Yu, Li Deng, and Yifan Gong, “Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7304–7308.
- [12] Xinjian Li, Siddharth Dalmia, Alan W Black, and Florian Metze, “Multilingual speech recognition with corpus relatedness sampling,” *Proc. Interspeech 2019*, pp. 2120–2124, 2019.
- [13] Christian DiCanio, Hosung Nam, Douglas H Whalen, H Timothy Bunnell, Jonathan D Amith, and Rey Castillo García, “Using automatic alignment to analyze endangered language data: Testing the viability of untrained alignment,” *The Journal of the Acoustical Society of America*, vol. 134, no. 3, pp. 2235–2246, 2013.
- [14] Alexis Michaud, Oliver Adams, Trevor Anthony Cohn, Graham Neubig, and Séverine Guillaume, “Integrating automatic transcription into the language documentation workflow: Experiments with na data and the persephone toolkit,” 2018.
- [15] Pavel Matejka, Petr Schwarz, Jan Cernocký, and Pavel Chytil, “Phonotactic language identification using high quality phoneme recognition,” in *Ninth European Conference on Speech Communication and Technology*, 2005.
- [16] Graham Neubig, Patrick Littell, Chian-Yu Chen, Jean Lee, Zirui Li, Yu-Hsiang Lin, and Yuyan Zhang, “Towards a general-purpose linguistic annotation backend,” *arXiv preprint arXiv:1812.05272*, 2018.
- [17] Anthony Rousseau, Paul Deléglise, and Yannick Esteve, “TED-LIUM: an automatic speech recognition dedicated corpus,” in *LREC*, 2012, pp. 125–129.
- [18] John J Godfrey, Edward C Holliman, and Jane McDaniel, “SWITCHBOARD: Telephone speech corpus for research and development,” in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*. IEEE, 1992, vol. 1, pp. 517–520.
- [19] Kikuo Maekawa, “Corpus of spontaneous japanese: Its design and evaluation,” in *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003.
- [20] Yi Liu, Pascale Fung, Yongsheng Yang, Christopher Cieri, Shudong Huang, and David Graff, “Hkust/mts: A very large scale mandarin telephone speech corpus,” in *Chinese Spoken Language Processing*, pp. 724–735. Springer, 2006.
- [21] Xingyu Na Bengu Wu Hao Zheng Hui Bu, Jiayu Du, “Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline,” in *Oriental COCOSDA 2017*, 2017, p. Submitted.
- [22] Zhiyong Zhang Dong Wang, Xuewei Zhang, “Thchs-30 : A free chinese speech corpus,” 2015.
- [23] Solomon Teferra Abate, Wolfgang Menzel, and Bairu Tafila, “An amharic speech corpus for large vocabulary continuous speech recognition,” in *INTERSPEECH-2005*, 2005.
- [24] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 369–376.
- [25] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., “The kald speech recognition toolkit,” in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011, number CONF.
- [26] David R Mortensen, Siddharth Dalmia, and Patrick Littell, “Epitran: Precision G2P for many languages,” in *LREC*, 2018.