INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Hierarchical Phone Recognition with Compositional Phonetics

Xinjian Li, Juncheng Li, Florian Metze, Alan W Black

Carnegie Mellon University, USA
xinjianl@cs.cmu.edu

Abstract

There is growing interest in building phone recognition systems for low-resource languages as the majority of languages do not have any writing systems. Phone recognition systems proposed so far typically derive their phone inventory from the training languages, therefore the derived inventory could only cover a limited number of phones existing in the world. It fails to recognize unseen phones in low-resource or zero-resource languages. In this work, we tackle this problem with a hierarchical model, in which we explicitly model three different entities in a hierarchical manner: phoneme, phone, and phonological articulatory attributes. In particular, we decompose phones into articulatory attributes and compute the phone embedding from the attribute embedding. The model would first predict the distribution over the phones using their embeddings, next, the language-independent phones are aggregated to the language-dependent phonemes and then optimized by the CTC loss. This compositional approach enables us to recognize phones even they do not appear in the training set. We evaluate our model on 47 unseen languages and find the proposed model outperforms baselines by 13.1% PER.

Index Terms: multilingual speech recognition, phone recognition, zero-shot learning, phonetics

1. Introduction

With the development of deep neural networks, there is growing interest in applying deep neural network models to speech recognition [1, 2, 3]. Those deep models, however, are restricted to languages with a large amount of training set such as English and Mandarin [4, 5], therefore, they are not available for most languages in the world. Additionally, the majority of the languages in the world have never been written [6], as a result, the only accessible speech recognition systems are phone recognition systems. Many works have focused on developing phone recognition systems for low-resource languages [7, 8, 9, 10]. However, most of them face the problem of the limited phone inventory. As the training languages typically consist of rich resource languages such as English and Mandarin, the training phone inventory usually consists of common phones available in European languages and East-Asian languages. This situation makes it hard to recognize unique phones in other language families. Another problem is the imbalanced phone distribution among the training set: some phones might appear frequently in many languages, but other phones might only occur in limited cases in one specific training language and therefore have much fewer training samples. This issue would cause the model to predict the first group more frequently and suppress the second group. Note that we distinguish the concept of *phone* and *phoneme* in this work [11]: *phone* represents the physical speech sound, it is the language-independent unit shared by all languages. In contrast, *phoneme* is the language-dependent unit, it is the smallest unit to distinguish meaning in a specific language. Phones and phonemes are highly related to each other

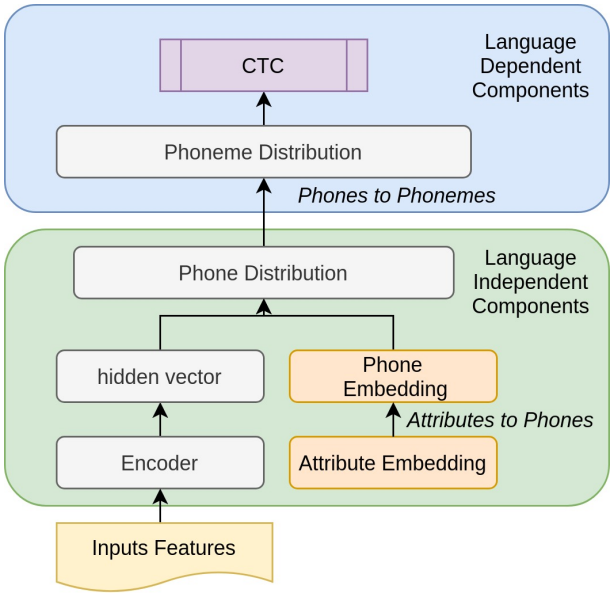

Figure 1: *The architecture of the hierarchical model. We first compose the phone embeddings from their attribute embeddings. Then we compute the phone distributions using the embeddings and the hidden vector from the encoder, Next, the language-independent phones are transformed into language-dependent phonemes with the allophone mappings, which would finally be optimized by the loss (CTC) function.*

and one phoneme might correspond to multiple phones (those phones are referred to as the *allophones*). For example, the phoneme /p/ in English have two actual phonetic realizations (allophones) [p] and [p^h].

In this work, we propose a novel hierarchical model to tackle the two problems stated above. While most traditional works tend to consider each phone as the basic independent building block, we further decompose phones into their components: phonological articulatory attributes. For instance, the phone [a] can be characterized as a *open front unrounded vowel* where each word (e.g: *open*) can be seen as its attribute. We assign each attribute an *attribute embedding* to encode its information, then the *phone embedding* can be constructed by summing up its corresponding attribute embeddings. Those embedding would be fine-tuned during the training process. With those embeddings, we can build the recognition model as illustrated in Figure.1: the encoder (BLSTM) first receives the input features and generates hidden vectors. We take the inner product of the phone embedding and the hidden vector to compute the phone distributions. Then the phone distribution is mapped to phonemes in each language using the allophone mappings. Finally, the phoneme distribution is optimized by the loss (CTC)

http://dx.doi.org/10.21437/Interspeech.2021-1803

function. This approach enables us to solve the aforementioned two problems: first, the phones are no longer independent units, they are interconnected by articulatory attributes shared with each other. Even for a new phone which we have not encountered so far, we can decompose it into existing attributes and then compute its embedding as well. Therefore, this model has the ability to handle unseen phones. Furthermore, this model would suffer less from the imbalanced phone distribution problem as we are optimizing the attribute embeddings instead of the phones themselves: even the rare phones would be fully trained through their attributes shared with other frequent phones. We apply our models to 47 unseen languages and the results indicate that our model improves the average PER (Phone Error Rate) by 13.1%.

2. Related Work

Training speech recognition systems for low-resource languages remains a challenge due to the limited supervised training set. One common approach is to train multilingual models on languages with rich supervised resources and then transfer its knowledge to new languages [12, 13, 14]. Another promising method proposed recently is to use the unsupervised approach to pretrain the encoder with a large amount of unsupervised dataset. The pretrained encoder can be fine-tuned to the target language with a limited size of training set [15, 16, 17].

Despite the success of those models, they still rely on the supervised set for the target language and could not be applied to any unseen languages. In particular, the unseen language might contain unseen phones which are not available in the training languages. One solution is to use language-independent phones instead of the language-dependent phonemes or subwords [18]. During the training phase, the model can first predict the distribution over the language-independent phones, then it transforms the distribution into the language-dependent phonemes to be optimized (as most of the training set is typically available in the form of phonemes). As the language-independent units are shared by all languages, this model can be applied to unseen languages without any training set for the target language. The only information required for such a model is the phone inventory, which is easy to obtain as PHOIBLE has published the inventory containing more than 2000 languages [19]. However, even this model cannot solve the issues of unseen phones as the available phone inventory is limited to the phones covered by the training languages. Additionally, it suffers from the problem of the imbalanced phone distribution we mentioned above.

One potential approach to overcome those two problems is to use phonological articulatory attributes. The articulatory attributes are well-defined by the linguists and most phones can be reduced to a list of discrete articulatory attributes [11]. By learning the representations over the articulatory attributes, we can associate any unseen phones with well-known attributes and therefore be available to use those phones during inference. Note that applying articulatory attributes to speech recognition tasks is not a new idea. To name a few, it has been applied to improve robustness under the noisy environment [20], improve performance for multilingual speech recognition [21], doing phoneme clustering for unwritten languages [22]. However, most works do not apply them to predict unseen phones. One work has applied a similar idea to recognize unseen phones as ours [23]. This work, however, does not distinguish between phones and phonemes, it constructs the language-dependent phonemes directly from the articulatory attributes. We find this model would not be properly trained when the number of training language increases because more languages would bring more phone-phoneme inconsistencies.

3. Approach

3.1. Compositional Phonetics

In this work, we introduce the approach of *compositional phonetics*, where we decompose phones into a list of phonological articulatory attributes. Each attribute has been assigned a fixed length of embedding which we refer to as the *attribute embedding*, those embeddings are first randomly initialized and get fine-tuned together with other parameters during the training process. By using those attribute embeddings, each phone can also be assigned an embedding by linearly composing the embedding from their attributes. Formally, consider a set of phones P, for each phone $p \in P$, we could determine a list of its attributes A_p. For each attribute in the list $a \in A_p$, we could assign an attribute embedding $e_a \in \mathbb{R}^n$ where n is the hidden size of the model. Then, the phone embedding $e_p \in \mathbb{R}^n$ can be computed by aggregating its attribute embeddings.

$$e_p = \sum_{a \in A_p} e_a \tag{1}$$

Suppose that the encoder computes the hidden vector $h \in \mathbb{R}^n$ for the current frame, we can obtain the logit l_p for this phone p by taking inner product

$$l_p = h^T e_p \tag{2}$$

Note that the embedding composition approach is not the only way to associate attributes and phones. A more simple idea used in [8] is to first compute the attribute logits $\mathbf{l}_A \in \mathbb{R}^{|A|}$ from the encoder, where $|A|$ is the size of entire attributes, then add up logits of corresponding attributes. We would refer to this model a linear model.

$$l_p = \sum_{a \in A_p} l_a \tag{3}$$

While the two approaches seem to be similar, we find that the embedding composition approach is more stable and typically leads to better performance. Our hypothesis is that the linear model encodes the hidden information with a small size of $|A|$, on the contrary, the embedding approach encodes the information with a much larger hidden size n and thus has better expressive power (in our experiment, $n = 640, |A| = 23$). Additionally, the embedding approach enables us to have a better understanding of the model through their embedding spaces.

Notice that while the potential number of phones is very large, the number of articulatory attributes is significantly smaller. In our estimation, we find PHOIBLE has listed more than 2000 unique phones across all registered languages [19], however, the articulatory phonological attributes are well-defined and we only consider 22 unique attributes (+1 ctc attribute) in this work. Even if a particular phone does not exist in the training set, we can still do the inference as we can easily compose its embedding from the known attribute embeddings.

3.2. Allophone Layer

The allophone layer is to transform the language-independent phone distributions into the language-dependent phoneme distributions. For the allophone layer, we follow the architecture

proposed in the previous work [18]. Suppose the current language is L, and its phoneme inventory is Q_L. For each phoneme in the inventory $q \in Q_L$, it has multiple allophones corresponding to it. Suppose the allophone set for q is P_q. Then each phone $p \in P_q$ is an allophone for q. The allophone layer computes the phoneme logits by selecting the max logits among its allophones.

$$l_q = \max\{l_p | p \in P_q\} \quad (4)$$

Finally, the phoneme distributions are fed into the CTC loss to be optimized [24]. CTC loss is selected as it has the conditional independence assumption, which reduces the dependency to the language modelings of the training languages, and thus make it easier to predict unseen phone sequence patterns.

4. Experiments

4.1. Settings

In this section, we describe our experiment in this work. We select 11 training languages as described in Table.1. Those languages are selected as they have large training sets and their phonology is well understood. We use Epitran to convert text into phoneme for each utterance in the text [25]. Each phoneme might correspond to several phones, those mapping rules are provided by Allovera [26]. Finally, we extract discrete phonological articulatory attributes from each narrow phone by using Panphon [27]. The tool supports 22 distinct features, we create two different attributes from each feature by considering whether that feature exists or not. For instance, `+syllabic` means it is a syllabic phone, `-syllabic` means it is not.

For the testing languages, we use a recently proposed dataset [28]. The dataset contains many small corpora from around 100 languages. Each corpus is phonetically annotated by linguists and manually aligned. We sort all corpus by their size and extract corpus whose size of utterances is larger than 50. The number of unique languages in this subset is 47, their ISO-639 id are abk, ady, afn, afr, agx, ajp, apc, ape, apw, asm, azb, bam, cbv, cpn, dan, ell, fin, guj, hau, haw, heb, hil, hin, hrv, hun, hye, ibb, ilo, isl, kan, kea, khm, klu, knn, lad, lav, lit, lug, mlt, mya, nan, nld, pam, pes, prs, wuu, yue.

For the evaluation, we compare 4 different acoustic models. The first one is the English phone recognition model which is a standard LSTM model trained using only English training sets. This model is used as a baseline to contrast language-dependent models and language-independent models. The second model is the Allosaurus model [18] whose architecture has an allophone layer mapping between phones and phonemes, it does not model any articulatory attributes and thus each phone is considered independent from each other. Those two models are open-sourced and available on Github.[1]. The other two models are hierarchical models we propose in this work. One hierarchical model is using a simple linear model mapping articulatory distributions into phone distributions. The other model is the main model we discuss in the previous section where we compose phone embeddings from the attribute embeddings and apply those embeddings to estimate distributions. All 4 models are using the same input feature and same encoder architecture: 40 dimension MFCCs and 5 layer bidirectional LSTM with 640 hidden size, the loss function are all CTC loss. The English model connects the encoder directly to the loss function, the Allosaurus model has an allophone layer between the encoder and loss function, the hierarchical models have the aforementioned compositional architecture.

[1] eng2102 and uni2005 from https://github.com/xinjli/allosaurus

Table 1: *Training corpora and size in utterances for each language. Models are trained with 11 rich resource languages*

Language	**Corpora**	**Utt.**
English	voxforge, Tedlium [29], Switchboard [4]	1148k
Japanese	Japanese CSJ [30]	440k
Mandarin	Hkust [31], openSLR [32, 33]	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 [34]	10k
Italian	voxforge	10k
Russian	voxforge	8k

4.2. Results

Table.2 shows the main results of our experiment. For each model, we evaluate it across all 47 languages and take the average of their PER (phone error rate). In addition, we also show the percentage of errors of addition, deletion, and substitution. The table indicates that the English model has 72% PER, which is the worst phone error rate among all models. The result is expected as the English model could only recognize phones available in English but is not able to recognize any unseen phones in our testing languages. This also explains the high substitution error rate in English as it typically replaces unknown phones with English phones during inference. The Allosaurus model performs better than the English model as it is a language-independent model and could cover a larger phone inventory. It improves the substitution error rate from 45.6% to 37.8%. Both hierarchical models perform significantly better than the Allosaurus model. The linear model has 57.6% PER and the compositional model has 51.2% PER.

Table 2: *Average Performance of 47 testing languages for each model. The proposed Hierarchical model using embedding approach performs best. PER is the phone error rate, Add, Del, Sub denotes the addition, deletion and substitution errors. All numbers are shown in %*

Model	PER	Add	Del	Sub
English model	72.0	11.2	15.2	45.6
Allosaurus model	64.3	7.86	18.6	37.8
Hierarchical (linear)	57.6	7.87	13.6	36.1
Hierarchical (embedding)	51.2	3.4	18.9	28.8

To have a better understanding of performance across languages, Figure.2 shows the box plot of the 4 models. It is clear from the figure that each model has a very large variance: some languages perform better and other languages perform worse. By investigating the performance of each language, we find languages with better recording environments tend to obtain better scores, and languages with many background noise tend to score worse. We also compute the correlations across 4 models as shown in Figure.3. It demonstrates that Allosaurus model and both hierarchical models are highly correlated, but the En-

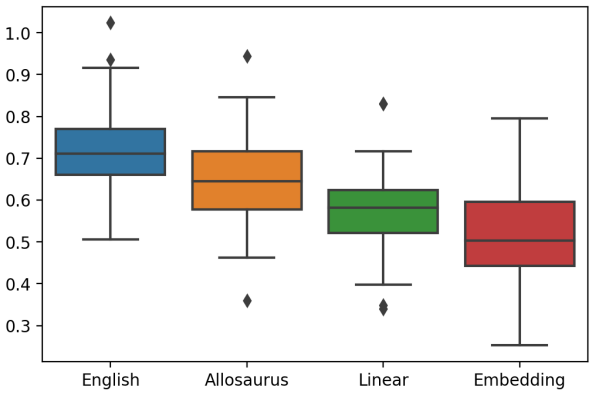

Figure 2: *The boxplot of performance distribution across all 47 languages for each model*

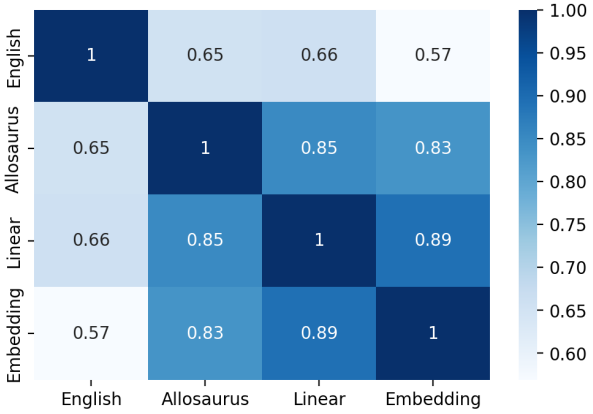

Figure 3: *Performance correlation between 4 models*

glish model is much less related. This is because the three models are language-independent models but the English model is language-dependent.

Next, we investigate the most common errors of the embedding model. Table.3 shows the top 3 errors and their occurrences across the dataset. The statistics indicate that the most common error is the deletion of phone [s]. Our hypothesis is that our model might have some difficulties in recognizing unvoiced sounds. For example, [s] is an unvoiced fricative consonant and [t] is an unvoiced plosive consonant. We find those unvoiced sounds typically have some characteristic patterns in the high frequency regions of spectrograms. However, our training set contains many 8k frequency audios and therefore the resolution of our model is restricted to 4k due to the Nyquist sampling theorem. Those deletion errors might be overcome by using high resolution audio corpus in the future. Another major errors come from the substitution errors, they have longer tails than the other two errors. The table suggests that most common substitution errors come from ambiguous vowels.

4.3. Analysis of Embeddings

During the training process, we also obtain the embeddings of both articulatory attributes and phones. The attribute embeddings do not have much patterns in them as they are mostly independently from each other. However, the phone embeddings

Table 3: *Most frequent errors in the Hierarchical model (embedding), the left side in the tuple is the error and the right side is its total occurrences in the test set. In the substitution row, the phone on the left side is the reference and the phone on the right side is the hypothesis*

Types	Most Common Errors
Add	([i], 104), ([a], 53), ([m], 47)
Del	([s], 247), ([a], 238), ([t], 221)
Sub	([a] -> [ə], 122), ([u] -> [o], 109), ([a] -> [ɑ], 104)

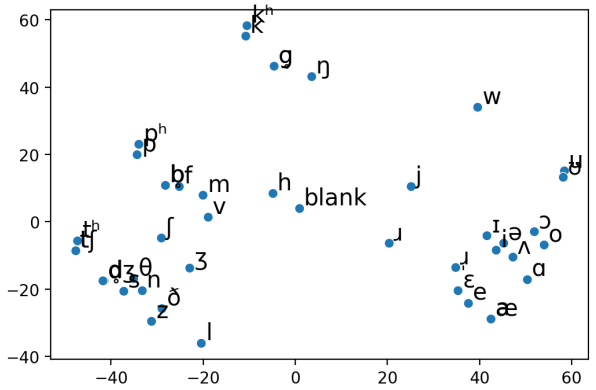

Figure 4: *PCA projected embeddings for all phones available in English. The embeddings are from the Hierarchical (embedding) model.*

have several interesting patterns. Figure.4 shows the embeddings of English phones. The embeddings originally have 640 dimension and get reduced to 2 dimension by PCA. There are several interesting things we can observe in the figure. First, there are a couple of clusters in the graph. The easiest one to identify is the vowel cluster at the right bottom corner. We have vowels such as [a], [o], [u] clustered together. This provides another reason for the substitution error: the embeddings of those phones are near to each other, therefore it is easy to confuse them with each other. On the top of the figure, we have the plosive velar group: [k] and [g]. [ng] near them is another velar consonant. Furthermore, we could find several word2vec like relations (e.g: king - queen = man - woman) in the figure. For example, for voiced and unvoiced sounds,

$$e([k]) - e([g]) = e([p]) - e([b]) \tag{5}$$

Similarly, for aspirated and unaspirated sounds, we find following relations:

$$e([p^{\mathrm{h}}]) - e([p]) = e([k^{\mathrm{h}}]) - e([k]) \tag{6}$$

5. Conclusion

In this work, we propose the hierarchical model for low-resource phone recognition where we explicitly model three different entities: phoneme, phone and articulatory attributes. We test the model on 47 unseen languages and the result demonstrates that our approach achieves 13.1% PER better than the baseline model. The model will be integrated into the Allosaurus repository for more researchers to explore phone recognition systems.

6. References

[1] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International conference on machine learning*, 2016, pp. 173–182.

[2] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina *et al.*, "State-of-the-art speech recognition with sequence-to-sequence models," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4774–4778.

[3] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 4960–4964.

[4] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, vol. 1. IEEE, 1992, pp. 517–520.

[5] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 5206–5210.

[6] F. Coulmas, *Written and unwritten language*, ser. Key Topics in Sociolinguistics. Cambridge University Press, 2013, p. 39–59.

[7] T. Schultz and A. Waibel, "Fast bootstrapping of lvcsr systems with multilingual phoneme sets," in *Fifth European Conference on Speech Communication and Technology*, 1997.

[8] X. Li, S. Dalmia, D. R. Mortensen, J. Li, A. W. Black, and F. Metze, "Towards zero-shot learning for automatic phonemic transcription," in *Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020.

[9] J. A. Thompson, M. Schönwiesner, Y. Bengio, and D. 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.

[10] X. Li, S. Dalmia, A. W. Black, and F. Metze, "Multilingual speech recognition with corpus relatedness sampling," *Proc. Interspeech 2019*, pp. 2120–2124, 2019.

[11] P. Ladefoged and K. Johnson, *A course in phonetics*. Nelson Education, 2014.

[12] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. 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.

[13] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean, "Multilingual acoustic models using distributed deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8619–8623.

[14] K. Veselỳ, M. Karafiát, F. Grézl, M. Janda, and E. Egorova, "The language-independent bottleneck features," in *Spoken Language Technology Workshop (SLT), 2012 IEEE*. IEEE, 2012, pp. 336–341.

[15] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised Pre-Training for Speech Recognition," in *Proc. Interspeech 2019*, 2019, pp. 3465–3469. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-1873

[16] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.

[17] M. Rivière, A. Joulin, P.-E. Mazaré, and E. Dupoux, "Unsupervised pretraining transfers well across languages," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7414–7418.

[18] 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 *ICASSP 2020*, 2020.

[19] S. Moran and D. McCloy, Eds., *PHOIBLE 2.0*. Jena: Max Planck Institute for the Science of Human History, 2019. [Online]. Available: https://phoible.org/

[20] K. Kirchhoff, "Combining articulatory and acoustic information for speech recognition in noisy and reverberant environments," in *Proc. ICSLP*, 1998.

[21] M. Müller, S. Stüker, and A. Waibel, "Towards improving low-resource speech recognition using articulatory and language features," in *Proc. IWSLT*, 2016.

[22] M. Müller, J. Franke, S. Stüker, and A. Waibel, "Improving phoneme set discovery for documenting unwritten languages," *Elektronische Sprachsignalverarbeitung (ESSV)*, vol. 2017, 2017.

[23] X. Li, S. Dalmia, D. R. Mortensen, J. Li, A. Black, and F. Metze, "Towards zero-shot learning for automatic phonemic transcription," in *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020.

[24] A. Graves, S. Fernández, F. Gomez, and J. 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] D. R. Mortensen, S. Dalmia, and P. Littell, "Epitran: Precision G2P for many languages." in *LREC*, 2018.

[26] D. R. Mortensen, X. Li, P. Littell, A. Michaud, S. Rijhwani, A. Anastasopoulos, A. W. Black, F. Metze, and G. Neubig, "AlloVera: A multilingual allophone database," in *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, 2020.

[27] D. R. Mortensen, P. Littell, A. Bharadwaj, K. Goyal, C. Dyer, and L. S. Levin, "Panphon: A resource for mapping IPA segments to articulatory feature vectors," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. ACL, 2016, pp. 3475–3484.

[28] X. Li, D. R. Mortensen, F. Metze, and A. W. Black, "Multilingual phonetic dataset for low resource speech recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6958–6962.

[29] A. Rousseau, P. Deléglise, and Y. Esteve, "TED-LIUM: an automatic speech recognition dedicated corpus." in *LREC*, 2012, pp. 125–129.

[30] K. Maekawa, "Corpus of spontaneous japanese: Its design and evaluation," in *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003.

[31] Y. Liu, P. Fung, Y. Yang, C. Cieri, S. Huang, and D. Graff, "Hkust/mts: A very large scale mandarin telephone speech corpus," in *Chinese Spoken Language Processing*. Springer, 2006, pp. 724–735.

[32] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*. IEEE, 2017, pp. 1–5.

[33] Z. Z. Dong Wang, Xuewei Zhang, "Thchs-30 : A free chinese speech corpus," 2015. [Online]. Available: http://arxiv.org/abs/1512.01882

[34] S. T. Abate, W. Menzel, and B. Tafila, "An amharic speech corpus for large vocabulary continuous speech recognition," in *INTERSPEECH-2005*, 2005.