ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9415014

PHONE DISTRIBUTION ESTIMATION FOR LOW RESOURCE LANGUAGES

Xinjian Li; Juncheng Li; Jiali Yao; Alan W Black; Florian Metze;

Carnegie Mellon University

xinjianl@cs.cmu.edu

ABSTRACT

Phones are critical components in various computational linguistic fields, for example, phone distributions could be helpful in speech recognition and speech synthesis. Traditional approaches to estimate phone distributions typically involve G2P systems which are either manually designed by linguists or trained on large datasets. These prohibitive requirements make research on low resource languages extremely challenging. In this work, we propose a novel approach to estimate phone distributions by only requiring raw audio datasets: We first estimate the phone ranks by combining language-independent recognition results and Learning to Rank results. Next, we approximate the distribution with Expectation-Maximization by fitting *Yule distribution*. The results on 7 languages show the joint-model has better performance in both ranking estimation and distribution estimation tasks.

***Index Terms*—** phone distribution estimation, low resource languages, multilingual speech recognition, ranking models

1. INTRODUCTION

Phones are one of the fundamental elements widely used in traditional linguistic fields [1, 2] and computational linguistic fields [3, 4, 5, 6]. Phone distribution, which indicates how phones are distributed within its inventory in each language, has broad applications in both traditional and computational linguistic research. For instance, in the traditional linguistic fields, phone distributions are central components of phonetics, phonology, and typology [1, 7], they can be applied to estimate how sound changes in historical linguistics [2]. In the applied fields, they serve as a prior to transform between posterior and likelihood in speech recognition [4, 3]. Additionally, they are useful when creating phonetically balanced speech datasets to minimize human costs [8]. Therefore, estimating the phone distribution is an important task.

We note that the task of estimating phone distribution is related to, but different from the task of estimating the phone inventory. The latter focuses on identifying a set of the phoneme or phone inventory for the target language,

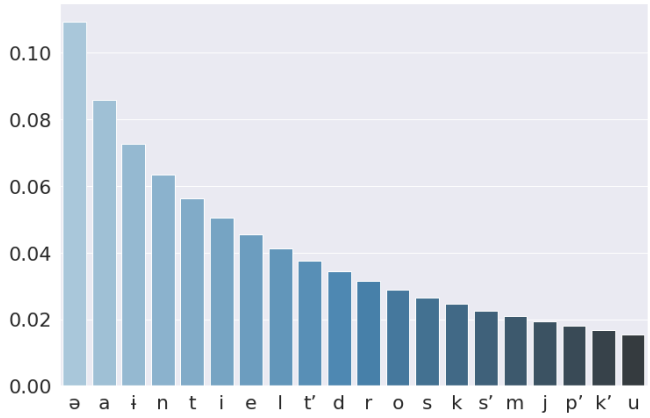

Fig. 1. An illustration of the phone distribution from the Amharic experiment. The horizontal axis shows the top ranked phones within the Amharic phone inventory, the vertical axis shows the estimated distributed frequencies.

whereas the former task is about estimating how phones are distributed within the fixed given phone inventory as illustrated in Figure.1, and hence the task of estimating phone distribution typically assumes the existence of predefined phone inventories. This assumption is reasonable in practice, because many languages including the low-resource ones have been studied carefully by phoneticians, and their phone inventory has been developed by those experts. For example, Phoible is a large phonological inventory that covers more than 2000 languages [9].

The phone distribution estimation task has not been fully explored so far. Although the phone distribution for rich-resource languages such as English and Mandarin could be easily estimated by applying a good G2P system to some text corpora [10], this is not the case for low-resource languages. Well-performing G2P systems are typically designed either manually by linguists or trained on large datasets [11], both of which are not usually available for most low-resource languages. Moreover, even a G2P system could be trained or transferred from other languages [12], the textual corpus itself might not be available as many low-resource languages do not possess writing systems. Therefore, estimating the phone dis-

tribution for low resource languages has been a challenging task.

In this work, we propose a novel approach to estimate phone distribution only using the phone inventory and the raw audios of the target language, which are both collectible even for unwritten languages [13, 9]. We first apply a language-independent phone recognizer to count the occurrence of each phone. For the recognizable phones, their frequencies can be used as the empirical distribution towards the phone distribution. In this step, there would be some missing phones that could not be recognized by the model as they are not included in the recognizer's inventory. We estimate the missing phones' ranks based on a trained ranking model. Then the occurrences and ranks are combined to fit *Yule distribution* [14] with Expectation-Maximization algorithm. We apply our approach to generate phone distributions and evaluate our results on 7 languages in detail. The result shows that the joint model produces the best performance.

2. APPROACH

2.1. Phone Recognition

In this work, only the raw audios are used to estimate the phone distribution. To cover as many languages as possible, we use the CMU Wilderness dataset [15] which is a collection of Bible recordings from around 800 different languages. The phone inventory is obtained from Phoible [9]. We associate languages from two datasets with ISO639-3 code and extract 676 languages. To recognize phones of various languages, a language-independent phone recognition model is required because any language-dependent systems could only discover phones of that specific language [16]. Additionally, language-dependent systems could not distinguish allophones which might be crucial in other languages [17]. To handle this issue, we adopt a recently proposed language-independent phone recognition model as the recognition tool in this work [18].

One critical issue with most multilingual recognition models is that their phone coverage is hardly complete [19]. For example, our trained model could cover around 200 phones, whereas the Phoible inventory has around 2000 distinct phones. For each low resource language, we estimate that around 20% phone inventory is missing from the model, which prevents us from measuring those phones' distribution. Suppose the phone inventory for a specific language is I and the model's recognizable inventory is I_{rec}. The distribution of phones within the inventory $p \in I_{rec}$ can be estimated easily by counting its occurrences in the audio dataset. The core problem to solve in this work is to estimate phone distribution for the missing phone inventory $I_{miss} = \{p | p \in I, p \notin I_{rec}\}$. To tackle the missing phones, we break this problem into two steps: 1) We estimate *ranks* for all phones including the missing phones in each language's inventory. 2) We approximate the entire phone distribution with *Yule distribution* and fill the values for the missing phones.

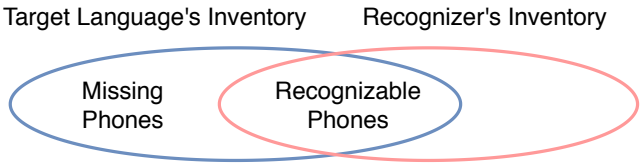

Fig. 2. Recognizable phones I_{rec} and missing phones I_{miss}

2.2. Phone Rank Estimation

The problem of phone rank estimation can be seen as a *Learning to Rank* problem, which is heavily used in the field of information retrieval to estimate ranks of documents [20]. In this work, we apply the same framework to estimate the phone ranking. In particular, we use the rank SVM approach to train a pair-wise model to assign ranks for all phones [21]. We use the articulatory feature as the feature to rank phones. For each phone, we use a fixed set of 37 articulatory feature templates extracted from Phoible, each template is encoded as a category feature. The ranking model could be easily trained with well-resource languages where actual phone ranks are available. Then the model is applied to low-resource languages: for each phone in its inventory I, a ranking score $S_{rank} \in R$ would be assigned.

Empirically, the ranks of phones are highly correlated with their occurrences. To quantify it, we introduce another score $S_{rec} \in [0, 1]$ which denotes the empirical score estimated from the recognition results: for every recognizable phone in I_{rec}, we use its frequency percentage in the entire dataset as the score S_{rec}. For other phones in I_{miss}, we assign 0 as they do not appear in the recognition. Finally, to consider both S_{rec} and S_{rank}, they are linearly combined to obtain a new score S_{joint}, which we use to sort and obtain the final ranks as the equation below. Note that the coefficient α here can be optimized using the training languages.

$$S_{joint} = S_{rec} + \alpha S_{rank} \tag{1}$$

2.3. Phone Distribution Estimation

The next step is to estimate the full distribution for all phones I, especially the distribution for the missing phones I_{miss}. It is known that the phone distribution can be modeled with *Yule distribution* [10]. The phone distribution with rank r is defined as

$$P(r; a, b) = \left(\sum_{i=1}^{|I|} \frac{a^i}{i^b} \right)^{-1} \left(\frac{a^r}{r^b} \right) \tag{2}$$

where the distribution is determined by two parameters $a, b \in R$. Our goal here is to estimate the full distribution

Language	Recognition Ranking Model	Estimated Ranking Model	Joint-Ranking Model
Amharic	0.818(***)	0.716(***)	0.785(***)
Cebuano	0.161(-,-,-)	0.631(***)	0.579 (-,**)
Ilocano	0.068(-,-,-)	0.645(***)	0.489 (-,-,*)
Kurmanji	-0.080(-,-,-)	0.342(-,-,*)	0.340 (-,-,*)
Swahili	0.788(***)	0.684(***)	0.774(***)
Tagalog	0.841(***)	0.695(***)	0.770(***)
Zulu	0.768(***)	0.576(***)	0.646(***)
Average	0.490(*)	0.612(***)	**0.626(***)**

Table 1. Results of the ranking evaluations on 7 languages. Three ranking models are compared using Spearman's ρ where a higher value shows better performance. It indicates that the joint-ranking model performs best on average. All numbers are shown with its statistical significance: (-,-,*) $p \leq 0.05$ (-,**) $p \leq 0.005$ (***) $p \leq 0.0005$

including the missing phones using *Yule distribution*. To estimate the missing phone's distribution, the parameters a, b should be specified. However, they are dependent on the missing phone's distribution. This is a typical problem that can be effectively solved by the Expectation-Maximization algorithm [22], where the missing distributions are latent variables to be estimated, a, b are parameters to be optimized. In the E-step, we estimate the missing phone distribution from parameter a, b, in the M-step, we optimize the parameter a, b with the full distribution including the missing phone distribution. After its convergence, both parameters a, b and the full distribution of inventory I are obtained.

3. EXPERIMENT

In this work, 10 languages are selected as the training languages: English, Mandarin, German, French, Italian, Javanese, Kazakh, Russian, Spanish, Turkish, Vietnamese. The 10 languages are used to train both the language-independent phone recognizer and the ranking model. Those languages are selected as training languages because they have a large amount of speech data to train the recognizer and good G2P systems to estimate actual phone distributions. Additionally, 7 different languages are used as the testing languages: Amharic, Cebuano, Ilocano, Kurmanji, Swahili, Tagalog, Zulu. These are selected from various language families and they have reasonable G2P systems to extract distribution as the golden dataset [11]. For the recognizer training, we follow the approach and dataset in the original work [18]. For the ranking model, we optimize the parameter α with all training languages and use $\alpha = 20$ in this work.

For each testing language, two golden datasets are prepared: the rank dataset and distribution dataset. The distribution dataset contains the golden phone distribution which is estimated by applying G2P system to the entire text dataset for each language (which is a subset of the speech training corpus). Then we obtain the rank dataset by only keeping the ranks for phones.

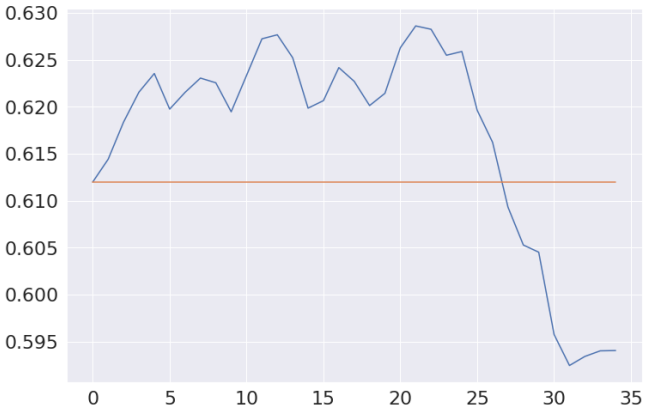

Fig. 3. The effect of using different α in the phone ranking estimation. The horizontal axis is the different values of α, the vertical axis is the Spearman's ρ in the joint-ranking model. The blue line moving up and down is the joint-ranking model with different α, the straight orange line is the estimated ranking model.

3.1. Phone Ranking Estimation

First, we evaluate our approach by only using phone rank scores. In particular, three models are compared: The recognition ranking model is based on the recognition score S_{rec}, the estimated ranking model uses the scores S_{rec} estimated from the ranking model, the joint-ranking model applies the combined score S_{joint}. For each model, we sort phones with their scores and then evaluate them using Spearman's ρ, which estimates the correlation between their ranks and golden ranks. The results are shown in Table.1. Overall, it shows that most ranks are statistically significantly correlated with the golden ranks. However, the recognition ranking model is unstable across languages: it shows a very

Language	Recognition Ranking Model	Estimated Ranking Model	Joint-Ranking Model
Amharic	0.440	0.281	0.225
Cebuano	1.333	0.397	0.368
Ilocano	1.924	0.455	0.366
Kurmanji	4.605	0.404	0.559
Swahili	0.386	0.345	0.242
Tagalog	0.100	0.282	0.231
Zulu	0.595	0.319	0.263
Average	1.340	0.354	**0.332**

Table 2. Results of the distribution evaluations on 7 languages. Three ranking models are compared using KL divergence with the golden dataset where a lower value indicates better performance.

high positive score on Tagalog but an even negative correlation on Kurmanji. This can be explained by the instability of the underlying recognition model because many phones are not recognized correctly. The estimated ranking model shows more stable results in all languages and significantly outperforms the recognition model by 0.12 on average. Additionally, by combining two scores from two models, our joint-ranking model achieves the best average 0.626 score and shows the most stable results across all languages. More detailed results show that the average performance of missing phones is 0.465 and recognizable phones' score is 0.655. The deviation is reasonable as the recognizable phones have more information from the empirical rankings.

Additionally, we investigate the effect of using different α in the joint ranking model. The results is illustrated in Figure.3. The joint ranking model with a small range of α performs better than the estimated ranking model. But the performance becomes worse very fast when α becomes larger. The reason might be the estimated ranking are much more stable than the recognition one. Using the empirical results can improve the ranks marginally, but relying too much on it could harm the performance.

3.2. Phone Distribution Estimation

Next, we evaluate the phone distribution estimated from those three models: In the recognition ranking model, we take the phone frequency as the distribution and assign all missing phones a fixed distribution (0.01) to avoid the 0 frequency issue. For the estimated ranking model, we use the ranks estimated from the ranking model and assign its distribution with a fixed *Yule distribution* whose parameters are estimated from training languages. In the joint ranking model, we estimate both missing phone distributions and parameters with EM algorithm. Three models are compared with the golden distribution using KL divergence. The results are demonstrated in Table.2. Similar to the results for phone ranking estimation, the recognition ranking model is unstable on phone distribution estimation as well: the Tagalog score is better than the other two models but the Kurmanji score is significantly worse than the two others. The Kurmanji recognition also affects the joint-ranking model's score through its score combination. Compared with the recognition ranking model, both the estimated ranking model and the joint-ranking model give consistent results across all languages. On average, the joint-ranking model performs the best among the three models.

4. DISCUSSION AND FUTURE WORK

While the joint-ranking model performs reasonably well in this work, we note there are several points that could be improved in the future. First, the joint-ranking model is easily affected by the recognition results as shown in the Kurmanji case. To make the result more robust in new languages, the recognition confidence score should be considered in the α parameter. Additionally, the training language for the ranking model is limited to a few language families which might not reflect distribution characteristics for other language families. For instance, phones heavily used in Romance languages might tend to have high ranks. Despite these potential improvements, however, this work paves the road to the future work in the phone distribution estimation.

5. CONCLUSION

In this work, we propose a novel approach to estimate phone distribution in low resource languages from only raw audio datasets. We combine the language-independent recognition model and the ranking model to estimate phone rank, then we optimize the distribution with EM algorithm. The results show that the joint-ranking model has the best performance in both ranking estimation and distribution estimation tasks.

6. REFERENCES

[1] Colin Yallop and Janet Fletcher, "An introduction to phonetics and phonology," 2007.

[2] Hans Henrich Hock, *Principles of historical linguistics*, Walter de Gruyter, 2009.

[3] Dan Jurafsky, *Speech & language processing*, Pearson Education India, 2000.

[4] Yajie Miao, Mohammad Gowayyed, and Florian Metze, "Eesen: End-to-end speech recognition using deep rnn models and wfst-based decoding," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 167–174.

[5] Jose Sotelo, Soroush Mehri, Kundan Kumar, Joao Felipe Santos, Kyle Kastner, Aaron Courville, and Yoshua Bengio, "Char2wav: End-to-end speech synthesis," 2017.

[6] Sercan Ö Arık, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, Jonathan Raiman, et al., "Deep voice: Real-time neural text-to-speech," in *International Conference on Machine Learning*, 2017, pp. 195–204.

[7] Matthew K Gordon, *Phonological typology*, vol. 1, Oxford University Press, 2016.

[8] M Asunción Moreno Bilbao, D Poig, Antonio Bonafonte Cávez, Eduardo Lleida, Joaquim Llisterri, José Bernardo Mariño Acebal, and Climent Nadeu Camprubí, "Albayzin speech database: Design of the phonetic corpus," in *EUROSPEECH 1993: 3rd European Conference on Speech Communication and Technology: Berlin, Germany: September 22-25, 1993*. . EUROSPEECH, 1993, pp. 175–178.

[9] Steven Moran, Daniel McCloy, and Richard Wright, "Phoible online," 2014.

[10] Yuri Tambovtsev and Colin Martindale, "Phoneme frequencies follow a yule distribution," *SKASE Journal of Theoretical Linguistics*, vol. 4, no. 2, pp. 1–11, 2007.

[11] David R Mortensen, Siddharth Dalmia, and Patrick Littell, "Epitran: Precision g2p for many languages," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[12] Aliya Deri and Kevin Knight, "Grapheme-to-phoneme models for (almost) any language," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 399–408.

[13] Gilles Adda, Sebastian Stüker, Martine Adda-Decker, Odette Ambouroue, Laurent Besacier, David Blachon, Hélene Bonneau-Maynard, Pierre Godard, Fatima Hamlaoui, Dmitry Idiatov, et al., "Breaking the unwritten language barrier: The bulb project," *Procedia Computer Science*, vol. 81, pp. 8–14, 2016.

[14] George Udny Yule, "A mathematical theory of evolution, based on the conclusions of dr. j. c. willis, f.r.s," *Philosophical transactions of the Royal Society of London. Series B, containing papers of a biological character*, vol. 213, no. 402-410, pp. 21–87, 1925.

[15] Alan W Black, "Cmu wilderness multilingual speech dataset," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5971–5975.

[16] Xinjian Li, Siddharth Dalmia, Alan W Black, and Florian Metze, "Multilingual speech recognition with corpus relatedness sampling," *Proc. Interspeech 2019*, pp. 2120–2124, 2019.

[17] David Mortensen, Xinjian Li, Patrick Littell, Alexis Michaud, Shruti Rijhwani, Antonios Anastasopoulos, Alan Black, Florian Metze, and Graham Neubig, "Allovera: a multilingual allophone database," in *LREC 2020: 12th Language Resources and Evaluation Conference*, 2020.

[18] Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, David R Mortensen, Graham Neubig, Alan W Black, et al., "Universal phone recognition with a multilingual allophone system," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 8249–8253.

[19] Xinjian Li, Siddharth Dalmia, David Mortensen, Juncheng Li, Alan Black, and Florian Metze, "Towards zero-shot learning for automatic phonemic transcription," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 8261–8268.

[20] UP Cambridge, "Introduction to information retrieval," 2009.

[21] Thorsten Joachims, "Training linear svms in linear time," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006, pp. 217–226.

[22] Christopher M Bishop, *Pattern recognition and machine learning*, springer, 2006.