Hierarchical Multi-Task Transformers for Crosslingual Low Resource Phoneme Recognition

Kevin Glocker[1], **Munir Georges**[1,2]
[1]Technische Hochschule Ingolstadt, Research Institute AImotion Bavaria, Ingolstadt, Germany
[2]Intel Labs Germany
kevin.glocker@thi.de, munir.georges@thi.de

Abstract

This paper proposes a method for multilingual phoneme recognition in unseen, low resource languages. We propose a novel hierarchical multi-task classifier built on a hybrid convolution-transformer acoustic architecture where articulatory attribute and phoneme classifiers are optimized jointly.

The model was evaluated on a subset of 24 languages from the Mozilla Common Voice corpus. We found that when using regular multi-task learning, negative transfer effects occurred between attribute and phoneme classifiers. They were reduced by the hierarchical architecture. When evaluating zero-shot crosslingual transfer on a data set with 95 languages, our hierarchical multi-task classifier achieves an absolute PER improvement of 2.78% compared to a phoneme-only baseline.

1 Introduction

While many highly effective architectures for speech recognition have been introduced in recent years, most require large amounts of language-specific training data. However, for a substantial portion of the worlds languages, only few or no annotated speech recordings are available for training or fine-tuning. To leverage the accuracy of end-to-end architectures, systems intended for low-resource ASR are often (pre-)trained on large multilingual corpora from mostly high-resource languages such as in Xu et al. (2021), who fine-tune a multilingually pretrained wav2vec 2.0 model for the crosslingual transfer task. They are either fine-tuned on low resource languages as evaluated by, e.g., Siminyu et al. (2021) or directly applied zero-shot, as outlined by Li et al. (2021a).

Several systems have been introduced that use articulatory attribute systems developed by linguists to improve phoneme recognition performance. In such systems, attributes are primarily used as an input in the form of trainable embeddings for each attribute individually as proposed by, e.g., Li et al. (2021a) or for feature vectors as in, e.g., Zhu et al. (2021), or using signature matrices as described by, e.g, Li et al. (2020). In contrast, Lee et al. (2019) applied multi-task learning to train separate articulatory feature classifiers and triphone states using shared layers for Mandarin at the same time with a TDNN architecture on forced alignments.

In this work, a multilingual phoneme recognition architecture is introduced. It is derived from a similar architecture applied to computer assisted pronunciation training in Mandarin (Glocker, 2021). Hierarchical multi-task learning is used to learn jointly to classify articulatory attributes and phonemes with an additional direct connection between the attribute and the phoneme classifier.

The proposed acoustic model for phoneme recognition is introduced in Section 2. The system is then evaluated in Section 3 in the high resource and zero-shot crosslingual settings. Afterwards, results are discussed and the paper concluded in Section 4.

2 Crosslingual Phoneme Recognition

Section 2.1 describes the hybrid transformer-acoustic model for encoding frame sequence. The hierarchical multi-task classifier for articulatory attributes and phonemes is introduced in Section 2.2.

2.1 Transformer Acoustic Model

A hybrid convolution and transformer encoder model is used for acoustic sequence modeling as shown in Figure 1. The architecture and hyperparameter choices are derived from the transformer model introduced by Synnaeve et al. (2019). First, the audio is resampled to 16kHz. 40 dimensional MFCC features using 25ms frames with a stride of 10ms are extracted. The features are then passed into two GLU-activated convolution layers to encode local context, with a kernel size of three and 512 and 400 channels respectively. Each convolution layer is preceded by layer normalization and

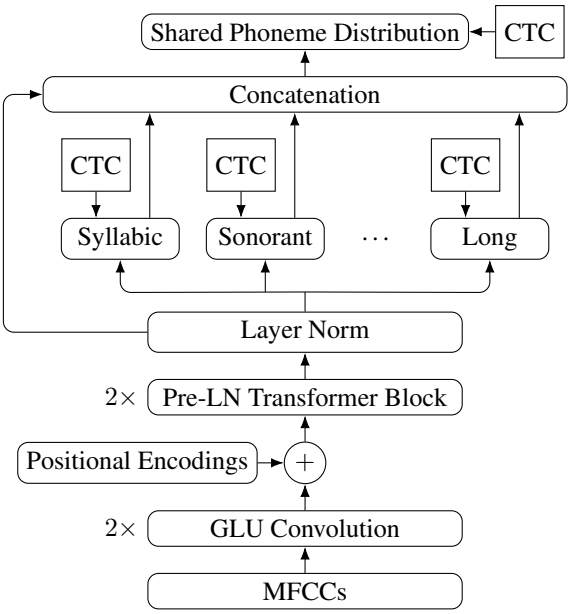

Figure 1: Illustration of the hybrid convolutional transformer phoneme recognition model with the hierarchical connections between attribute and phoneme classifiers

followed by a dropout layer for regularization. A stride of 2 is used in the second GLU layer, increasing the receptive field of the model to 5 frames while keeping the output lengths shorter than the length of phoneme sequences for CTC.

Sinusoidal positional encodings as proposed by Vaswani et al. (2017) are added to the output representations of the convolution layers. The sequence is passed through a shallow 2-layer transformer. In the transformer, Pre-LN transformer blocks are used without warmup as proposed by Xiong et al. (2020). Feedforward layers with a hidden size of 2048 and 4 attention heads are used motivated by Vaswani et al. (2017). The dropout rate is 0.2.

2.2 Hierarchical Multi-Task Classifiers

In contrast to previous work (Lee et al., 2019), classifiers are not trained completely independently but are connected in a hierarchical structure. Cascading information between tasks has also previously been successfully applied to jointly optimizing NLP tasks at different "levels" such as POS and dependency parsing (Crawshaw, 2020).

In the hierarchy, both the attribute and phoneme classifiers take the normalized output of the transformer acoustic model as an input. In addition to the acoustic representation, the phoneme classifier receives a concatenation of the probability distributions from each articulatory attribute classifier. More specifically, for each time step t given a set of attribute classifier logits A_t, the transformer hidden vector h_t, and the weights and biases of the phoneme projection layer W and b, the phoneme logits p_t are computed as follows:

$$v_t = \left(\bigoplus_{a \in A_t} \text{softmax}(a)\right) \oplus h_t \quad (1)$$

$$p_t = W^T v_t + b \quad (2)$$

Each classification layer is then independently but simultaneously optimized using connectionist temporal classification (CTC; Graves et al. (2006)). For consistency, articulatory attribute vectors are directly mapped to each phoneme without merging repetitions. As a result, there is always a 1:1 correspondence between attribute and feature labels at training time. While the attribute and phoneme classifiers form a flat hierarchy in this work, the hierarchical structure generalizes to any directed acyclic graph representing phonetic feature structures.

3 Evaluation

We evaluated the proposed hierarchical multi-task transformer with two experiments.

(1) In the "Multi-Task" variant, regular multi-task learning is used where attribute probabilities are not used as inputs to the phoneme classifier.

(2) In the "Phonemes Only" model, only the phoneme classifier is used and attribute information is only applied to phoneme mapping at test time.

Batch sizes are set dynamically for efficiency until the product of the batch and frame sequence dimensions reaches 320,000. The Adam optimizer (Kingma and Ba, 2015) was used for training with $\beta_1 = 0.9$ and $\beta_2 = 0.98$ as in Vaswani et al. (2017). A learning rate of 0.001 is used. The training was stopped once the average validation set losses did not decrease for more than 3 epochs.

The transformer acoustic model was implemented in the PyTorch framework (Paszke et al., 2019), using Torchaudio (Yang et al., 2021) for Audio processing and feature extraction.

The data sets for training and evaluation are described in Section 3.1. Section 3.2 presents and analyses the results for phoneme and attribute classification for high and low resource languages.

3.1 Datasets

For training and evaluation in the high resource setting, version 10.0 of the Mozilla Common Voice corpus was used, which contains crowdsourced recordings of sentences. Each sentence is tokenized

using Stanza (Qi et al., 2020), after which punctuation is removed and each token is transcribed into phonemes using Epitran (Mortensen et al., 2018). Finally, the transcriptions are segmented according to the IPA segments available in the Panphon database (Mortensen et al., 2016) for the phoneme inventory extracted from the training data for each language. The 24 articulatory attributes from Panphon are used for creating and supervising the attribute classifiers. The multilingual training set was constructed from at most 15,000 sentences from the training sets of 24 languages from Common Voice, for which both a tokenization and a grapheme-to-phoneme model is available. The original development and test sets were used unchanged.

The first release[1] of the multilingual corpus published by Li et al. (2021b) is used for evaluating zero-shot transfer in this work as in Li et al. (2021a). It provides 5,509 validated utterances with phoneme transcriptions for 95 low-resources languages from five continents. Since recordings for Czech, Dutch, Maltese, Hindi and Hungarian are also included in the training data, they are removed from the test data before computing the averages.

To handle different inventories and OOV phonemes in the test languages, phonemes predicted by the model are mapped to each target inventory using the hamming distance between attribute vectors. This corresponds to the "tr2tgt" approach introduced by Xu et al. (2021). For the UCLA Phonetic Corpus, the included inventory files are used for this mapping even if they include a phoneme that doesn't appear in a transcription.

3.2 Experiments

The overall performance on phoneme and articulatory attribute detection on Common Voice can be seen in Table 1. In addition to the phoneme error rate (PER), the attribute error rate (AER) is computed for each attribute individually and then averaged over all attributes. The hierarchical multi-task model reaches lower PER and average AER than regular multi-task learning in both the high and low resource setting. The regular multi-task model also performs worse than the phoneme only baseline. This shows, that negative transfer effects are stronger without the hierarchical connection.

Compared to the "Phonemes Only" model, the hierarchical model performs almost identically in the high-resource setting. However, as shown in Table 2, there is an improvement to the unseen low-resource languages from the UCLA Phonetic Corpus. In contrast, the regular multi-task model also yields higher PERs in this setting.

Architecture	%PER	%AER
Phonemes Only	48.96	–
Multi-Task	52.19	19.43
Hierarchical Multi-Task	49.11	17.99

Table 1: Average phoneme and attribute error rates for the Common Voice subset representing the high resource setting

Architecture	%PER	%AER
Phonemes Only	74.77	–
Multi-Task	75.28	34.14
Hierarchical Multi-Task	71.99	30.25

Table 2: Average phoneme and attribute error rates for the UCLA Phonetic Corpus representing the low resource setting

Figure 2 shows the phoneme and average attribute error rates for the Common Voice test sets of the languages used for training. The variance of PERs between languages is high ($\sigma^2 = 135.03$). On the attribute level, the variance of the AER between languages is much less pronounced ($\sigma^2 = 15.61$) and lower AER doesn't correlate with higher PER ($r^2 = 0.016$). For instance, the PER is highest for Arabic and Vietnamese even though their AER are among the lowest in the test set.

Since the AER was improved most consistently across languages through the hierarchical architecture, research into better modeling the connection between articulatory attributes and phonemes could lead to larger PER improvements in future work.

For Arabic and Urdu, a contributing factor might be Epitran not transcribing short vowels since they are not present in their orthography (Mortensen et al., 2018). For Vietnamese, the higher PER is likely due to it being the second-lowest resource language in the training data with only 2259 validated utterances and one of only two tonal languages alongside Thai.

In contrast, phoneme recognition is the most accurate for the five romance languages including Spanish, Italian and Catalan. They likely benefit the most from the multilingual settings since they

[1] https://github.com/xinjli/ucla-phonetic-corpus/releases/tag/v1.0

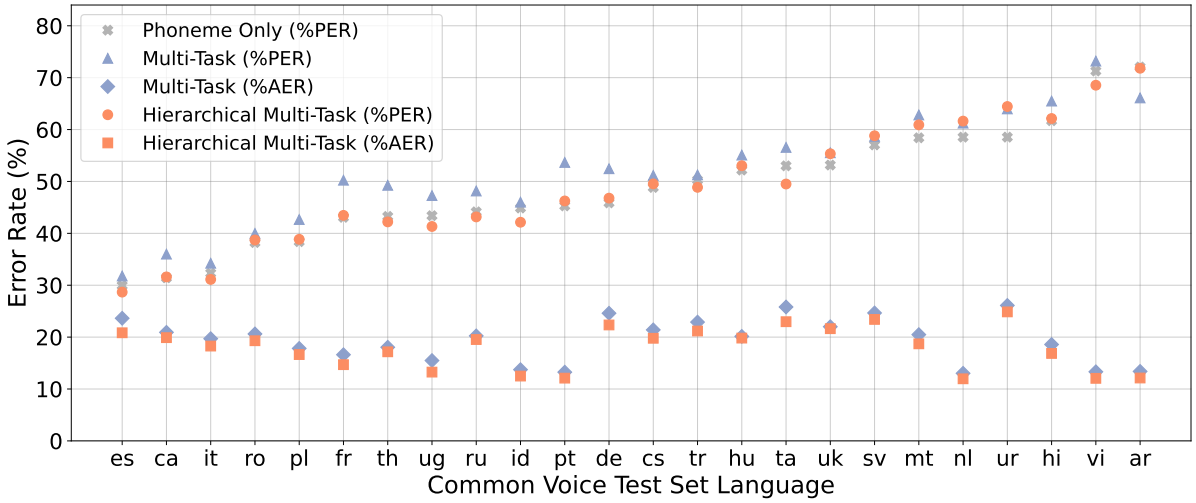

Figure 2: Phoneme Error Rates (PER) and the averages over all Attribute Error Rates (AER) on the test sets from Common Voice for the languages used for training

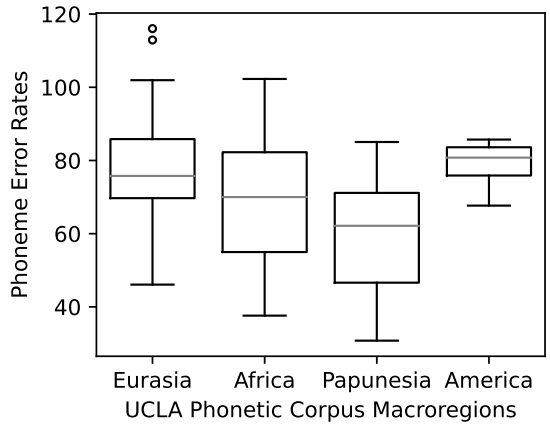

Figure 3: Phoneme Error Rates (PER) for the languages in the UCLA Phonetic Corpus grouped into macroregions accordig to Glottolog (Hammarström et al., 2022)

are closely related.

A possible explanation for the low correlation between AER and PER is, that the frame level probabilities tend to form single frame spikes when trained with CTC (Graves et al., 2006). Since CTC loss is computed for every classifier independently, spikes for attributes of the same phoneme sometimes occur on different frames. As a result, the phoneme classifier is likely to receive high blank probabilities from multiple attribute classifiers.

The crosslingual transfer results are further divided into macroregions in Figure 3 based on Glottolog (Hammarström et al., 2022). The model transfers best to the set of 10 languages from the "Papunesia" region, despite there being no languages from this region in the training set. In contrast, the model generalizes poorly to the four American languages. Some outliers with particularly high PER might also be caused by the noisy conditions under which some utterances were recorded (Li et al., 2021b).

4 Conclusion

A novel hierarchical multi-task architecture is presented and evaluated together with a hybrid convolution-transformer acoustic model for phoneme classification. In contrast to regular multi-task learning, the phoneme classifier receives attribute probabilities as additional inputs.

It tackles the crosslingual transfer task for phoneme recognition in low resource languages. For zero-shot classification in such languages, only their phoneme inventory is required.

Negative transfer effects observed in regular multi-task learning were reduced. When evaluated on the UCLA Phonetic Corpus, the proposed system yielded an absolute phoneme error rate reduction of 2.78% across 95 unseen languages compared to a phoneme-only baseline.

Future work may investigate the low correlation between AER and PER, and further analyse the cause of the high variance of PER between languages. In particular, we plan to investigate and improve the mapping between the shared phoneme inventory and language specific inventories to tackle these challenges. Furthermore, tones could be moved to their own layer in the hierarchy to better reflect their suprasegmental nature.

References

Michael Crawshaw. 2020. Multi-task learning with deep neural networks: A survey. *CoRR*, abs/2009.09796.

Kevin Glocker. 2021. Unsupervised end-to-end computer-assisted pronunciation training for mandarin. Master's thesis, Eberhard Karls Universität Tübingen.

Alex Graves, Santiago Fernández, Faustino J. Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. *Proceedings of the 23rd international conference on Machine learning*.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2022. glottolog/glottolog: Glottolog database 4.6.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Yueh-Ting Lee, Xuan-Bo Chen, Hung-Shin Lee, Jyh-Shing Roger Jang, and Hsin-Min Wang. 2019. Multi-task learning for acoustic modeling using articulatory attributes. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 855–861.

Xinjian Li, Siddharth Dalmia, David R. Mortensen, Juncheng Li, Alan W. Black, and Florian Metze. 2020. Towards zero-shot learning for automatic phonemic transcription. In *AAAI*.

Xinjian Li, Juncheng Li, Florian Metze, and Alan W. Black. 2021a. Hierarchical Phone Recognition with Compositional Phonetics. In *Proc. Interspeech 2021*, pages 2461–2465.

Xinjian Li, David R. Mortensen, Florian Metze, and Alan W Black. 2021b. Multilingual phonetic dataset for low resource speech recognition. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6958–6962.

David R. Mortensen, Siddharth Dalmia, and Patrick Littell. 2018. Epitran: Precision G2P for many languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori S. Levin. 2016. 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*, pages 3475–3484. ACL.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Kathleen Siminyu, Xinjian Li, Antonios Anastasopoulos, David R. Mortensen, Michael R. Marlo, and Graham Neubig. 2021. Phoneme Recognition Through Fine Tuning of Phonetic Representations: A Case Study on Luhya Language Varieties. In *Proc. Interspeech 2021*, pages 271–275.

Gabriel Synnaeve, Qiantong Xu, Jacob Kahn, Edouard Grave, Tatiana Likhomanenko, Vineel Pratap, Anuroop Sriram, Vitaliy Liptchinsky, and Ronan Collobert. 2019. End-to-end asr: from supervised to semi-supervised learning with modern architectures. *ArXiv*, abs/1911.08460.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *ArXiv*, abs/1706.03762.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. 2020. On layer normalization in the transformer architecture. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10524–10533. PMLR.

Qiantong Xu, Alexei Baevski, and Michael Auli. 2021. Simple and effective zero-shot cross-lingual phoneme recognition. *ArXiv*, abs/2109.11680.

Yao-Yuan Yang, Moto Hira, Zhaoheng Ni, Anjali Chourdia, Artyom Astafurov, Caroline Chen, Ching-Feng Yeh, Christian Puhrsch, David Pollack, Dmitriy Genzel, Donny Greenberg, Edward Z. Yang, Jason Lian, Jay Mahadeokar, Jeff Hwang, Ji Chen, Peter Goldsborough, Prabhat Roy, Sean Narenthiran, Shinji Watanabe, Soumith Chintala, Vincent Quenneville-Bélair, and Yangyang Shi. 2021. Torchaudio: Building blocks for audio and speech processing. *arXiv preprint arXiv:2110.15018*.

Chengrui Zhu, Keyu An, Huahuan Zheng, and Zhijian Ou. 2021. Multilingual and crosslingual speech

recognition using phonological-vector based phone embeddings. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1034–1041.