Towards Induction of Structured Phoneme Inventories

Alexander Gutkin
Google, UK
agutkin@google.com

Martin Jansche*
Amazon, UK
jansche@amazon.com

Lucy Skidmore
University of Sheffield, UK
lskidmore1@sheffield.ac.uk

1 Introduction

Phonological typology is an important branch of linguistic typology concerned with the study of the distribution and behavior of sounds in world's languages (Gordon, 2016). The cross-linguistic typological databases, such as PHOIBLE by Moran and McCloy (2019b), provide the crucial tools for drawing typological generalizations within the field. In addition to facilitating development of probabilistic models of phonological typology (Cotterell and Eisner, 2017; Ahn and Mortensen, 2019), such resources were also shown to positively influence the downstream multilingual NLP (Littell et al., 2017), speech (Li et al., 2020) and language documentation tasks (Anastasopoulos, 2019).

This short paper provides an overview of our completed and ongoing experiments with phonological representations utilizing phonological typology databases in speech processing tasks. Despite the increasing popularity of end-to-end approaches to automatic speech recognition (Moritz et al., 2019), text-to-speech (Chen et al., 2019) and speech-to-speech translation (Tjandra et al., 2019), there are still plenty of scenarios where integration of accurate phonological knowledge is crucial, or at least beneficial, including the end-to-end approaches themselves as recently demonstrated by Salesky and Black (2020).

In particular, we describe two strains of research motivated by the need to scale the development of speech technologies that still require phonological representations to low-resource languages and dialects. We first overview the framework for analysis of cross-lingual consistency of phonological features in multilingual phoneme inventories derived from cross-lingual typological databases. We then offer a sketch of a method that may serve as a potential building block in the future phoneme inventory induction system and the central role the phonological typology plays in this approach.

2 Multilingual Phoneme Inventories

The phoneme was originally defined as a theoretical abstraction that applies language-internally. Using phonemes and their succinct distinctive feature (DF) encodings in cross-linguistic settings, the practice going back to Dalsgaard and Andersen (1992), raises an important question: given a multilingual phoneme inventory derived from a typological database, such as PHOIBLE or PANPHON (Mortensen et al., 2016), it is not clear a priori whether all the DFs will be useful or even valid. If DF representations were phonetic rather than phonemic, and acoustic rather than articulatory, one would expect a close correspondence between DFs and the acoustic signal. In practical multilingual applications, however, the representations are often guided by purely phonemic considerations because of the availability of phonemic inventories and transcriptions. Such approaches are often outperformed by the models striving for more phonetic realism, such as the allophonic models of Mortensen et al. (2020).

We followed a simple method: to consider a phonemic contrast to be consistent or robust across languages, it needs to be easily predictable on held-out languages in a binary classification task (Johny et al., 2019). An instance of this problem consists of a span of a speech signal (e.g., a vowel in surrounding context) and a positive or negative label (e.g., front vowel vs. back vowel). A classifier is trained on a multi-speaker, multi-language dataset withholding one or more languages, which are later used for evaluation. For cases where cross-linguistic consistency did not hold, we extended this method by additionally grounding the representation on the contextual phonological knowl-

* This work was done while the author was at Google, prior to joining Amazon.

edge given as DFs, excluding the contrast itself (Skidmore and Gutkin, 2020). For our experiments we used a set of languages from Dravidian, Indo-European and Malayo-Polynesian families with the phoneme inventories derived from PHOIBLE.

Overall, our findings are mixed. A specially designed experiment for predicting contrasts in unvoiced labial consonants between Bengali and Spanish produced consistent and cross-lingually robust predictions (Johny et al., 2019), also for a variety of auditory representations (Gutkin, 2020), despite the conflicting status of some of the allophones of the phonemes in the experiment. Similarly robust were contrasts between front and back vowels, as well as the vowel height and continuant manner of articulation distinctions (Skidmore and Gutkin, 2020). The negative results include the cross-lingual prediction of retroflex consonants between the language families: retroflex predictor trained on Dravidian languages fails to reliably predict retroflex consonants in Bengali, conversely the predictor trained on Indo-Aryan languages is not reliable for Malayalam. Similarly inconsistent, but less disappointing, was the detection of aspiration. Furthermore, inclusion of other contrasts as contextual input features did not lead to significant improvements in predicting these hard cases.

One of the motivations behind our investigations describe above is a research question that still remains open: Can the above methodology be used for analyzing the cross-linguistic quality of the existing phoneme inventories given the data? For example, among the Malayo-Polynesian languages we considered, only Javanese has retroflex consonants, which it acquired through loanwords from Indo-Aryan or Dravidian languages (Ogloblin, 2006). PHOIBLE contains three different phoneme inventories for Javanese, the largest of which `GM 1675` represents retroflex plosives, while the smallest `UPSID 380` (Moran and McCloy, 2019a) omits them. Which of the two representations is more likely to be a better fit in a multilingual pronunciation model?

3 Towards Phonology Induction

Our ongoing research focuses on induction of phoneme inventories for languages for which the conventional resources required for speech model training are often missing, placing this task among other related work in zero-resource subword modeling problems (Baljekar et al., 2015; Lee et al., 2015; Chen et al., 2017). The current state-of-the-art unsupervised acoustic unit discovery approaches derive acoustic-phonetic (Hu et al., 2020; Morita and Koda, 2020; van Niekerk et al., 2020) and latent auditory-like (Ondel et al., 2019) representations, yet it is unclear how accurate these representations are from a typological standpoint.

Our initial investigations utilized "universal" multilingual phoneme recognizers. The results were disappointing primarily because the training data sparsity and the absence of language models to constrain the search often resulted even in unreliable recovery of the phoneme inventories for unseen dialects of the same language, e.g., determining the inventory of Argentinian Spanish having seen Castilian and Mexican Spanish. A more viable approach to this task is integrating language identification and phonological typology into the phoneme recognizer. Given an accurate language identification model, the PHOIBLE phoneme inventories belonging to the "closest" languages may be used to constrain the phonemic hypothesis space for the unseen language or dialect. We developed efficient search techniques for the lattices of this type (Jansche and Gutkin, 2019).

Currently we are revisiting the phonological contrast predictor methods described in the previous section to adapt them to phonology induction tasks. Several known approaches to phonemic feature detection in continuous speech are known, some are purely reliant on signal processing, while others are model-based (Frankel et al., 2007; Müller et al., 2017). At the simplest level, the output of such predictors represents speech as the parallel asynchronous streams of articulatory features. More sophisticated models that exploit the structure of articulation, feature geometry and other types of correlations between various features are also possible. In these approaches, the phonological cross-linguistic databases, such as PHOIBLE or PanPhon, provide the important source of truth not only for combining various features into phonemes, but also for determining which combinations of hypothesized phonemes are admissible given the existing phoneme inventories. Furthermore, additional phonological evidence provided by other typological resources, such as World Atlas of Language Structures (WALS) (Dryer and Haspelmath, 2013), can be integrated as well if it can be reliably extracted from the speech signal (Gutkin et al., 2018).

References

Emily Ahn and David Mortensen. 2019. Predicting Continuous Vowel Spaces in the Wilderness. In *Proc. of TyP-NLP: The First Workshop on Typology for Polyglot NLP*, pages 22–24, Florence, Italy. Association for Computational Linguistics.

Antonios Anastasopoulos. 2019. *Computational Tools for Endangered Language Documentation*. Ph.D. thesis, University of Notre Dame, Indiana.

Pallavi Baljekar, Sunayana Sitaram, Prasanna Kumar Muthukumar, and Alan W Black. 2015. Using articulatory features and inferred phonological segments in zero resource speech processing. In *Proc. Interspeech*, pages 3194–3198, Dresden, Germany. ISCA.

Wenda Chen, Mark Hasegawa-Johnson, Nancy F. Chen, and Boon Pang Lim. 2017. Mismatched Crowdsourcing from Multiple Annotator Languages for Recognizing Zero-Resourced Languages: A Nullspace Clustering Approach. In *Proc. Interspeech*, pages 2789–2793, Stockholm, Sweden. ISCA.

Yuan-Jui Chen, Tao Tu, Cheng chieh Yeh, and Hung-Yi Lee. 2019. End-to-End Text-to-Speech for Low-Resource Languages by Cross-Lingual Transfer Learning. In *Proc. Interspeech*, pages 2075–2079, Graz, Austria. ISCA.

Ryan Cotterell and Jason Eisner. 2017. Probabilistic typology: Deep generative models of vowel inventories. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1182–1192, Vancouver, Canada. Association for Computational Linguistics.

Paul Dalsgaard and Ove Andersen. 1992. Identification of Mono- and Poly-Phonemes Using Acoustic-Phonetic Features Derived by a Self-Organising Neural Network. In *Proc. 2nd International Conference on Spoken Language Processing (ICSLP)*, pages 547–550, Banff, Canada.

Matthew S. Dryer and Martin Haspelmath. 2013. WALS online. *Leipzig: Max Planck Institute for Evolutionary Anthropology*. `http://wals.info/`.

Joe Frankel, Mirjam Wester, and Simon King. 2007. Articulatory feature recognition using dynamic Bayesian networks. *Computer Speech & Language*, 21(4):620–640.

Matthew K Gordon. 2016. *Phonological Typology*. Oxford Surveys in Phonology and Phonetics. Oxford University Press.

Alexander Gutkin. 2020. Eidos: An open-source auditory periphery modeling toolkit and evaluation of cross-lingual phonemic contrasts. In *Proc. 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 9–20, Marseille, France. ELRA.

Alexander Gutkin, Tatiana Merkulova, and Martin Jansche. 2018. Predicting the Features of World Atlas of Language Structures from Speech. In *Proc. 6th International Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU)*, pages 248–252, Gurugram, India. ISCA.

Yushi Hu, Shane Settle, and Karen Livescu. 2020. Multilingual jointly trained acoustic and written word embeddings. *arXiv preprint arXiv:2006.14007*.

Martin Jansche and Alexander Gutkin. 2019. Sampling from Stochastic Finite Automata with Applications to CTC Decoding. In *Proc. Interspeech*, pages 2230–2234, Graz, Austria. ISCA.

Cibu Johny, Alexander Gutkin, and Martin Jansche. 2019. Cross-Lingual Consistency of Phonological Features: An Empirical Study. In *Proc. Interspeech*, pages 1741–1745, Graz, Austria. ISCA.

Chia-ying Lee, Timothy J O'Donnell, and James Glass. 2015. Unsupervised lexicon discovery from acoustic input. *Transactions of the Association for Computational Linguistics*, 3:389–403.

Xinjian Li, Siddharth Dalmia, Juncheng Li, Patrick Littell, Matthew Lee, Jiali Yao, Antonios Anastasopoulos, David Mortensen, Graham Neubig, Alan Black, and Florian Metze. 2020. Universal Phone Recognition with a Multilingual Allophone System. In *Proc. 45th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 8249–8253, Barcelona, Spain. IEEE.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Steven Moran and Daniel McCloy. 2019a. Javanese sound inventory from UCLA Phonological Segment Inventory Database (UPSID). In *PHOIBLE 2.0*. Max Planck Institute for the Science of Human History, Jena.

Steven Moran and Daniel McCloy, editors. 2019b. *PHOIBLE 2.0*. Max Planck Institute for the Science of Human History, Jena.

Takashi Morita and Hiroki Koda. 2020. Exploring TTS without T using Biologically/Psychologically Motivated Neural Network Modules (ZeroSpeech 2020). *arXiv preprint arXiv:2005.05487*.

N. Moritz, T. Hori, and J. L. Roux. 2019. Triggered Attention for End-to-end Speech Recognition. In *Proc. 44th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5666–5670, Brighton, United Kingdom. IEEE.

David R. Mortensen, Xinjian Li, Patrick Littell, Alexis Michaud, Shruti Rijhwani, Antonios Anastasopoulos, Alan W. Black, Florian Metze, and Graham Neubig. 2020. AlloVera: A Multilingual Allophone Database. In *Proc. 12th Language Resources and Evaluation Conference (LREC)*, pages 5329–5336, Marseille, France. ELRA.

David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori 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, Osaka, Japan. The COLING 2016 Organizing Committee.

Markus Müller, Jörg Franke, Alex Waibel, and Sebastian Stüker. 2017. Towards phoneme inventory discovery for documentation of unwritten languages. In *Proc. 42nd International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5200–5204, New Orleans, USA. IEEE.

Benjamin van Niekerk, Leanne Nortje, and Herman Kamper. 2020. Vector-quantized neural networks for acoustic unit discovery in the ZeroSpeech 2020 challenge. *arXiv preprint arXiv:2005.09409*.

Alexander Ogloblin. 2006. Javanese. In Alexander Adelaar and Niklaus Himmelman, editors, *The Austronesian Languages of Asia and Madagascar*, 1st edition, chapter 21, pages 590—594. Routledge.

Lucas Ondel, Ruizhi Li, Gregory Sell, and Hynek Hermansky. 2019. Deriving Spectro-Temporal Properties of Hearing from Speech Data. In *Proc. 44th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 411–415, Brighton, UK. IEEE.

Elizabeth Salesky and Alan W Black. 2020. Phone Features Improve Speech Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2388–2397. ACL.

Lucy Skidmore and Alexander Gutkin. 2020. Does A Priori Phonological Knowledge Improve Cross-Lingual Robustness of Phonemic Contrasts? In *Proc. 22nd International Conference on Speech and Computer (SPECOM)*, pages 530–543, St. Petersburg, Russia. Springer Nature.

Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. 2019. Speech-to-Speech Translation between Untranscribed Unknown Languages. In *Proc. Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 593–600. IEEE.