Version of Record: https://www.sciencedirect.com/science/article/pii/S0885230822000067
Manuscript_fd5ddd1cdd686cd823edb30d77648a8d

Discovering Phonetic Inventories with Crosslingual Automatic Speech Recognition

Piotr Żelasko[a,b], Siyuan Feng[c], Laureano Moro Velázquez[a], Ali Abavisani[d], Saurabhchand Bhati[a], Odette Scharenborg[c], Mark Hasegawa-Johnson[d], Najim Dehak[a,b]

[a] *Center of Language and Speech Processing, The Johns Hopkins University, 3400 North Charles Street, Baltimore, 21218, MD, USA*
[b] *Human Language Technology Center of Excellence, The Johns Hopkins University, 810 Wyman Park Drive, Baltimore, 21218, MD, USA*
[c] *Multimedia Computing Group, Delft University of Technology, Van Mourik Broekmanweg 6, Delft, 2628 XE, The Netherlands*
[d] *Department of Electrical and Computer Engineering, University of Illinois, 405 N Mathews, Urbana, 61801, IL, USA*

Abstract

The high cost of data acquisition makes Automatic Speech Recognition (ASR) model training problematic for most existing languages, including languages that do not even have a written script, or for which the phone inventories remain unknown. Past works explored multilingual training, transfer learning, as well as zero-shot learning in order to build ASR systems for these low-resource languages. While it has been shown that the pooling of resources from multiple languages is helpful, we have not yet seen a successful application of an ASR model to a language unseen during training. A crucial step in the adaptation of ASR from seen to unseen languages is the creation of the phone inventory of the unseen language.

The ultimate goal of our work is to build the phone inventory of a language unseen during training in an unsupervised way without any knowledge about the language. In this paper, we 1) investigate the influence of different factors (i.e., model architecture, phonotactic model, type of speech representation) on phone recognition in an unknown language; 2) provide an analysis of which phones transfer well across languages and which do not in order to understand the limitations of and areas for further improvement for automatic phone inventory creation; and 3) present different methods to build a phone inventory of an unseen language in an unsupervised way. To that

end, we conducted mono-, multi-, and crosslingual experiments on a set of 13 phonetically diverse languages and several in-depth analyses. We found a number of universal phone tokens (IPA symbols) that are well-recognized cross-linguistically. Through a detailed analysis of results, we conclude that unique sounds, similar sounds, and tone languages remain a major challenge for phonetic inventory discovery.

Keywords: phone inventory, ASR, speech recognition, multilingual, crosslingual, zero-shot, phone recognition, speech representation
Funding: This work was funded by NSF grant number 1910319.

---

1. Introduction

Automatic Speech Recognition (ASR) is one of the technologies deployed on a massive scale with the highest impact of the 21^{st} century. It has achieved enormous advancement during the past fifteen years [1], enabling automatic transcription, translation, and the ubiquitous appearance of digital assistants, among other applications. Combined with spoken language understanding, ASR has the potential to automate numerous processes that require spoken communication, such as requests for support and knowledge. Typically, the development of ASR systems requires hundreds to thousands of hours of speech recordings and their correspondent transcriptions. Unfortunately, only a small number of the world's languages is sufficiently resourced to build speech processing systems, while the amount of transcribed speech data for most of the 7,000 spoken languages in the world [2] is very limited or nonexistent [3]. This results in digital divides [4] that can only be fixed with large investments in speech resources, if considering only traditional ways to develop ASR models for new languages.

One method that is often used to reduce the data requirements for under-resourced languages is to build word-level ASR out of phone models. The word "phone" was coined in [5] as an abbreviation for "phonetic symbol," defined in [6] as an element of a phonetic transcription corresponding to at most one phoneme, whose boundary times in the acoustic signal can be reliably identified using automatic forced alignment. Such language-dependent ASR segment inventories may be expressed using the language-independent symbols of the IPA [7], and their set union defines a language-independent phone inventory, which may be trained using multilingual data [8]; alternatively, language-dependent phone models may be trained using far less data than

language-dependent word models, because the number of phones in a language is far fewer than the number of words [9]. In order to use phone-based acoustic models, however, it is necessary to discover the phone inventory of the unseen language. Past research has addressed this problem and proposed knowledge transfer from high-resource to low-resource languages using ASR adaptation technologies. For instance, studies in [8, 10, 11, 12, 13] propose multilingual (i.e., using training material from multiple languages including that of the target language) or crosslingual (i.e., using material from multiple languages excluding that of the target language) ASR systems, trained with several languages and targeting a low-resource target language. The reasoning behind the use of models that are trained with data from other languages is that sounds in different languages share phonetic representations due to similarities in articulatory and acoustic patterns. The phones of the target language are then assumed to be part of one or more of the training languages. However, this is not always the case, in which case those phones in the target language cannot be found. In some cases when a phone is not found in the training languages, it may be a subtle variant of some corresponding phones in the training languages, that could be used instead to form a phone system.

Most past research reports performance on the task of a phone or word transcription task in the unseen, target language. In this work, we go one step further and focus on the task of automatic, unsupervised phone inventory creation without any knowledge about the target language. To the best of our knowledge, we are the first to do so.

Below, we present our main aim and research questions. We will refer to them throughout the paper like this: [**MA**].

Main Aim [MA]. Our main research question is: Can the phone inventory of an unseen language be discovered in an unsupervised manner employing cross-lingual ASR systems trained with several languages? Previous research reports improvements in terms of Phone Error Rate (PER) or Word Error Rate (WER) when going from monolingual (low-resource languages) to multilingual training [14, 15, 12, 13, 16, 15, 11]. Here, in addition to reporting PER, we investigate various factors that influence crosslingual ASR and automatic phone inventory creation for an unseen language, and investigate what the ASR models are learning in order to understand the limitations of and areas for further improvement for building ASR systems for low-resource languages, and particularly the creation of the phone inventory of an unseen language.

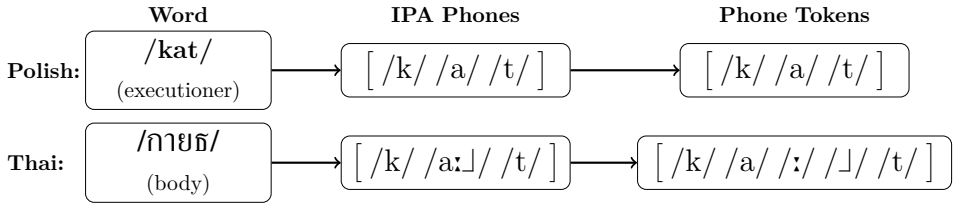


Figure 1: An overview of the basic phonetic units used in our study – phones, which include modifiers such as tones as a part of the symbol, and phone tokens, where the modifiers are modelled separately. Both variants leverage IPA symbols.

Research Question 1 [RQ1]. Firstly, we investigate what types of speech representations to use to train our models. Phonemes are language-dependent sound categories whose substitution may change the meaning of a word [17]. Phones are the acoustic realizations of phonemes whose boundary times in the acoustic signal can be reliably identified using automatic forced alignment. The number of non-phonemic contrasts in the ASR segment inventory is small in some languages, e.g., a typical American English ASR distinguishes a few allophones of /t/, such as [ɾ] and [t] [18]. In other ASR segment inventories, the number of suprasegmental contrasts may be quite large, e.g., a typical ASR segment inventory for Mandarin Chinese distinguishes among all possible tonally-marked syllable finals [19]: [a˥], [a˧˥], [a˨˩˧], [a˥˩], and the neutral-tone vowel [a] are all distinct. Such language-dependent ASR segment inventories may be expressed using the language-independent symbols of the IPA [7], and their set union defines a universal (language-independent) phone inventory. We define the goal of our research to be the discovery of the language-dependent ASR segment inventory as a subset of a larger, multilingual phone inventory. Defining our goal in this way, however, creates some ambiguity: if Mandarin ASR uses [a˨˩˧] as an element in its standard ASR segment inventory, but our phonetic inventory discovery system discovers only the phoneme /a/, should that be considered a correct, incorrect, or partially correct result? The ambiguity is solved, in this paper, by defining two levels of language-independent annotation. Our *phone* inventory is the strict set union of the language-dependent ASR segment inventories, expressed in IPA notation. Our *phone token* inventory is the set of all unicode characters used to express the phone inventory. Thus for example, the phone inventory includes the distinct phones {..., a, a˥, a˧˥, a˨˩˧, a˥˩, ...}, while the phone token inventory includes the distinct

phone tokens {..., a, ◌̩, ◌̪, ◌̺, ◌̻, ...}. We distinguish two types of phone tokens: symbols that describe acoustic segments (base phones such as [a]), and symbols that modify the base phone in order to produce a different phone (diacritics, including secondary articulations, lengthening, lexical stress, and lexical tone, such as [◌̻]). Diacritics such as the high tone symbol [◌̻] have no sound unless coupled with a base phone such as [a], but perhaps crosslingual ASR can be learned more effectively by a system that decomposes [a◌̻] into its component phone tokens, i.e., a system where [a] and [◌̻] are modelled separately (Figure 1). Therefore, we aim to answer the question: Can phone-based or phone-token-based acoustic models, trained using data from several languages coded with the same International Phonetic Alphabet (IPA) symbols, be used to detect the presence of a phone coded with the same IPA symbol(s) in the ASR segment inventory of a different language?

[RQ2]. Secondly, we know that the speech models learn phonetic representations [20], however, it is unclear: Which phones' representations are the most, and the least, amenable to crosslingual transfer? We also look for the factors that may help in explaining the findings. Using several in-depth analyses we answer these questions.

[RQ3]. Thirdly, it is not possible to have a complete picture on how multilingual representations can be used in zero-shot scenarios without studying to what degree the mismatch of phonotactics[1] between training and target languages affect the overall phone recognition results of the target language. Consequently, our third research question is: How do phonotactic constraints influence phone recognition in multilingual and crosslingual scenarios? Is this influence also visible in phonetic inventory discovery performance?

Proposed solution. To answer these questions, we trained two architectures with different properties, i.e., an End-to-End (E2E) phone-level ASR and a hybrid DNN-HMM ASR, using training data from 13 languages from 8 language families and with vastly different phonetic properties. This experimental protocol is based on the preliminary exploration we have performed before – in [21], we looked into E2E ASR phonetic representation transfer across languages, and in [22] we investigated the effect of phonotactics on zero-shot ASR transfer.

Hybrid ASR systems explicitly factorize the acoustic and language mod-

[1]Phonotactics define restrictions in a language on the permissible combinations of phonemes.

els into independent modules, which allows us to quantify the influence of the language phonotactics on phone recognition performance in multilingual and zero-shot scenarios ([**RQ3**]). To measure the effect of phonotactic information, phonotactic LMs of different capacities were trained and compared alongside the hybrid ASR. In our ASR experiments, we used both IPA phones and phone tokens as our speech representations ([**RQ1**]). In our experiments, we use three scenarios: monolingual, as a baseline; multilingual, to investigate the usefulness of pooling all languages together; and crosslingual, or zero-shot ASR. In each scenario, we evaluate the degree to which the confusion matrix rows corresponding to different phones, phone tokens, or classes of phones are helped or harmed by multilingual or cross-lingual modeling ([**RQ2**]). Finally, in order to address the **Main Aim** [**MA**] of the paper, we employed the best E2E and hybrid architectures in a crosslingual scenario, and propose and compare several methods for automatic phone inventory discovery.

2. Related work

Phonetic inventory discovery. The language-dependent categorical perception of phonemes was strongly argued in a series of thought experiments by linguists including Trubetzkoy [23] and Sapir [24]. While trying to define the phoneme inventory of the native American language Chitimacha, Swadesh provided one of the clearest published definitions of a phoneme: "The phoneme is the smallest potential unit of difference between similar words recognizable as different to the native. Given a correct native word, the replacement of one or more phonemes by other phonemes (capable of occurring in the same position) results in a native word other than that intended, or a native-like nonsense word" [17]. During the following twenty years, the Swadesh principle was used to propose phoneme inventories for hundreds of languages, including English [25], Russian [26], Amahuaca [27], Cashibo [28], Huasteco [29], Totonaco [30], and Zoque [31]. The PHOIBLE database [32] lists 3020 phoneme inventories that have been published as descriptions of 2186 languages.

There is a great deal of research on the automatic discovery of phone inventories for downstream technological tasks. Four types of system evaluation are prevalent. First, systems participating in the zero resource speech recognition challenges [33], including [34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48] are evaluated using an ABX task similar to the Swadesh

minimal-pairs test [17], but are not asked to propose a phone inventory whose speech sound clusters match to any human-annotated ground truth. Second, unsupervised phone discovery systems may be evaluated as components of zero-resource keyword spotting [49, 50, 51, 52, 53, 54, 55]. Third, unsupervised phone discovery systems may be evaluated for the purpose of developing automatic speech synthesis in a zero-resource language [56, 57, 58, 59, 60]. Fourth, unsupervised phone discovery can be used as a component in unsupervised lexicon discovery [61, 62, 63] or unsupervised speech-to-text [64]. Unsupervised speech-to-text may be evaluated on the basis of word error rate or phone error rate, but has not, to our knowledge, been evaluated in the task of phone inventory discovery.

Comparison of an automatically discovered phone inventory to a ground-truth, human annotated phone inventory has been performed in certain limited ways. Kempton et al. [65] force-aligned human-annotated fine phonetic transcriptions to audio in an under-resourced language (Kua-nsi), and then evaluated the degree to which the force-aligned transcriptions could be clustered to compute the true phoneme inventory of the language. Results indicated the need for further improvements for any particular phonetic class, e.g., active articulator was correctly classified with only 0.745 Area Under the Curve (AUC). Following up on their study, several researchers have attempted unsupervised clustering of acoustic spectra using Bayesian [66], variational [67] or hidden Markov models [68, 69], in order to directly estimate the phone inventory of an unknown language. The cardinality of the resulting pseudo-phone set is usually ten times larger than the cardinality of the true phone inventory, but by associating each pseudo-phone to its most highly overlapped ground-truth phone, it is possible to evaluate the pseudo-phone inventory using metrics such as F1 and normalized mutual information (NMI, [70]). Finally, several recent studies have attempted a combination of Kempton's approach with the unsupervised clustering approach: cross-lingual ASR is used to annotate the articulatory features of an unknown language, which are then clustered to form unsupervised phone-like units [56, 57]. To our knowledge, only two of these papers [71, 72] directly evaluated phone inventory NMI or F1; using oracle cluster combination strategies that are standard in the field of unsupervised phone discovery, [72] achieved F1=64.14% for cross-lingual automatic phone inventory estimation.

Crosslingual ASR. We call an ASR system that may recognize more than one language as *multilingual*; when the target language is not seen during training, then the system becomes *crosslingual*. Multilingual ASR mod-

els have been a point of interest even before Deep Neural Network (DNN) became the state-of-the-art acoustic modelling technique [8]. Pooling multiple languages into the training set – with or without architectural changes in the neural network-based acoustic models – is known to be effective for improving ASR [8, 73, 74, 11, 14, 12, 75, 16, 76, 77, 13]. Note that both monolingual and multilingual systems can be applied crosslingually, although multilingual systems can be expected to provide a more universal phonetic representation [21]. Two major challenges specific to crosslingual ASR are *mismatched outputs*, as both graphemic and phonemic symbols are usually different across languages; and *mismatched inputs*, as the new languages might consist of sounds unseen in the training languages. Past works address this challenge with attempts to adapt the training vocabularies to new languages [78, 79, 80], convert the lexicons of other languages to the target language [81, 51], or extend the vocabulary and fine-tune on the target language's data [82, 83, 84]. Another recent direction is pre-training the models to learn word-level acoustic representations [85, 86, 87]. Apart from our proposed ASR systems, any ASR system attempting to recognize language-independent (i.e., phonetic) units, such as the multilingual allophone approach proposed in [13], would be suitable for performing phonetic inventory discovery.

Analysis of learned representations. Most of the mentioned studies do not perform a detailed analysis of which phones, manner and place of articulation classes, or language families can benefit more from multilingual and crosslingual training. This knowledge could be essential to build better ASR systems for languages with low or no resources. Finally, to our knowledge, no study performs IPA phone inventory discovery of unknown languages by exclusively employing cross-lingual ASR systems, without additional acoustic clustering based on audio in the target language. By studying this problem, we are able to make statements about the degree to which phonemes and phones may be approximated by universal, language-independent categories.

3. Methods

In this section we present the details of our approach to phonetic inventory discovery. We start with a brief revision of the architectures of two ASR system types used in our study: the hybrid ASR in Section 3.1, and the end-

to-end ASR in Section 3.2. We use two ASR toolkits: Kaldi[2] and ESPnet[3]. In Section 3.3 we propose a novel method of discovering which phones are similar to each other across languages, based on clustering and projection of the ASR phonetic confusion matrix. Finally, we describe our method of discovering phonetic inventories from ASR transcripts in Section 3.4.

3.1. Hybrid DNN-HMM ASR

A hybrid ASR system allow us to inspect the impact of phonotactics on the crosslingual system performance and phonetic inventory discovery [**RQ3**]. The hybrid system consists of three independent modules. The first is the Acoustic Model (AM), which predicts the posterior likelihood $p(\boldsymbol{X}|\boldsymbol{S})$ of phones or phone HMM states $\boldsymbol{S} = \{s_0, s_1, ...s_N\}$ given the acoustic feature frames $\boldsymbol{X} = \{x_0, x_1, ..., x_N\}$. Modern acoustic models are usually implemented with deep neural networks. The second module is the Pronunciation Lexicon (PL) that maps sequences of phones into words. The third module is the Language Model (LM), which predicts the conditional likelihood $p(w_{i+1}|w_0, w_1, ...w_i)$ of the next word w_{i+1} given the previous words. All these components are incorporated together in a weighted finite-state transducer framework [88], and the most likely word sequence is retrieved using graph search methods, such as beam search. For more details, we refer the reader to [88] and the documentation of the Kaldi ASR toolkit[4].

The above description concerns word-level hybrid ASR systems. The modifications we apply to this framework are the following:

1. The PL is only applied during training to convert the word-level transcripts into phone sequences. During decoding, we are only interested in knowing phonetic transcripts.
2. For decoding, we replace the word-level LM with a phone-level LM when converting the AM posteriors into a sequence of phones. We further refer to the phone-level LM as the phonotactic LM, or the phonotactic model.

Note that we do not experiment with phone-token hybrid ASR systems [**RQ1**]. Hybrid ASR depends on high quality frame-level alignments during its training. We believe that the task of aligning symbols such as tone

[2] `https://github.com/kaldi-asr/kaldi`
[3] `https://github.com/espnet/espnet`
[4] https://kaldi-asr.org/doc/

modifiers is ill-defined, as they are a property of the base phone and not a separate sound that follows the base phone. However, the same limitation is not shared by another class of ASR systems we are about to discuss.

3.2. End-to-End ASR

In the recent years, another class of ASR systems, popularly called End-to-End (E2E), has grown as an alternative to the hybrid systems. These systems will help us investigate which output label type, phones or phone tokens, could be more beneficial for crosslingual ASR [**RQ1**].

Whereas the exact definition of E2E is not clear, typically this term describes models that perform graphemic transcription directly, i.e., they require neither an additional pronunciation lexicon, nor even an external language model to transcribe[5]. These systems are trained without using a forced or ground-truth alignment – often using a Connectionist Temporal Classification (CTC) loss, or an attention encoder-decoder architecture. The CTC objective is frame-synchronous (i.e., each frame is assigned a label) and maximizes the likelihood of observing the correct transcript given the acoustic frames over all possible alignments corresponding to the ground truth text [89].

The encoder-decoder attention model consists of two modules: an encoder, and a decoder with attention. The encoder first encodes the input frame sequence (possibly sub-sampling it first), and produces a sequence of intermediate representations that are used as decoder inputs. The decoder is autoregressive - it models the conditional probability of the next output token, conditioned on the previous output token and the intermediate representations provided by the encoder. The first decoding step is initialized with a special $[BOS]$ symbol (*beginning-of-sentence*), and it stops once the decoder recognizes a special $[EOS]$ symbol (*end-of-sentence*). For each prediction step, the encoder's activations are aggregated via a weighted sum, with the weights determined by the source-target attention mechanism [90]. The decoder's output sequence is typically much shorter than the vector of acoustic frames, making encoder-decoder a frame-asynchronous model. While the encoder-decoder attention model has the advantage of modelling the acoustics and language jointly, it is susceptible to misalignment due to

[5]Note that a fusion with an external language model can still improve E2E system's performance.

noise or convergence issues. Hence, [91] and [92] proposed to combine the CTC and encoder-decoder attention model to take advantage of both systems' capabilities. Going forward, in this work we will refer to this joint CTC and attention system as E2E. Note that whereas usually E2E means a system that performs graphemic transcription, we are specifically using the same approach to generate an IPA phonetic transcription, using units of either phone tokens (each decoder output generates one IPA character) or phones (each decoder output generates one or more IPA characters).

Note that the E2E system does not allow us to inspect the impact of the phonotactics **[RQ3]** as they are learned jointly with the acoustic representations (even when we do not use an external LM).

3.3. Visualizing phonetic confusions

We aim to learn which phones' representations are the most amenable to crosslingual transfer and which phones are hard to recognize crosslingually **[RQ2]**. To that end, we analyze the patterns in the zero-shot systems' phone confusions. A phone (token) confusion can be interpreted as: a given target-language phone (token) resembled a different source-language phone (token) more than it resembled the same source-language ("correct") phone (token). The cross-lingual confusion matrix therefore represents the degree to which each phone or phone token fails to transfer well across languages.

Our approach is visualized in Figure 2. We compute the edit distance alignments between the reference and hypothesized phone sequences and use them to estimate the phone confusion matrix. Unfortunately, the confusion matrix is quite large due to the size of the IPA symbol vocabulary, making a direct inspection of it hardly feasible. Instead, inspired by some of the methods of confusion matrix re-ordering [93], we transform it to the following two representations: a hierarchical clustering dendrogram, and a 2-D Uniform Manifold Approximation and Projection (UMAP) [94]. The dendrogram allows us to easily view specific phone token clusters, whereas the projection hints at the global structure of the confusion space and the relationships between different symbols and clusters.

The clustering and projection algorithms both depend on the estimation of a distance matrix with a carefully selected distance metric. In order to obtain meaningful visualisation, we propose to view the confusion matrix as a set of probability distributions – for each phone token (row), what is the probability of confusing it with another phone token (column). The intuition behind this approach is that when two symbols tend to be confused

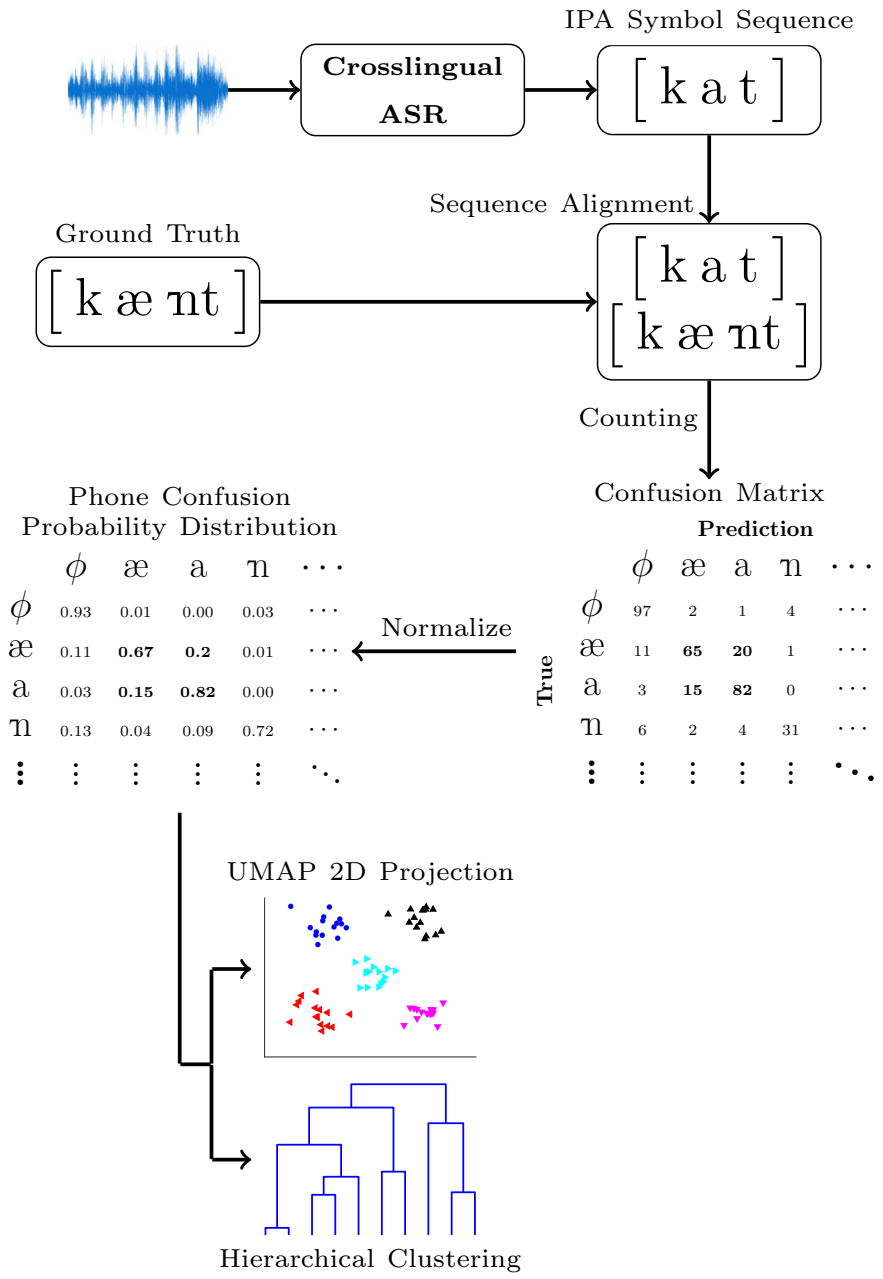


Figure 2: High-level workflow illustrating how the phone confusion matrix is obtained and mapped into phone similarity visualisations.

with the same set of classes, they must be similar to each other. We find it helpful to first zero out the confusion pairs that occurred less than 50 times, and then remove the phone symbols that have empty (all zero) rows or columns. Using this threshold, we reject about 5.3% of all confusions (452202 out of 8536116 symbols confused in the transcripts), which account for approx. 37% of all confusion types (e.g., we understand confusing [a] with [e] and [a] with [o] as a different confusion type; we discard 21168 out of 57120 confusion types). Without any thresholding, the patterns supported by most of the data are obfuscated by a long tail of noisy patterns after the occurrence counts are normalized. Then, we preprocess the confusion matrix with row-normalization, so that each row is converted from occurrence counts to a categorical probability distribution of confusing a ground truth phone P with another phone P'. We are going to interpret these distributions as feature vectors, describing the similarity between symbol pairs. Representing each phone as a probability distribution provides us a sound motivation to select the Jensen-Shannon divergence (JSD) as the distance metric.

3.4. Phonetic inventory discovery

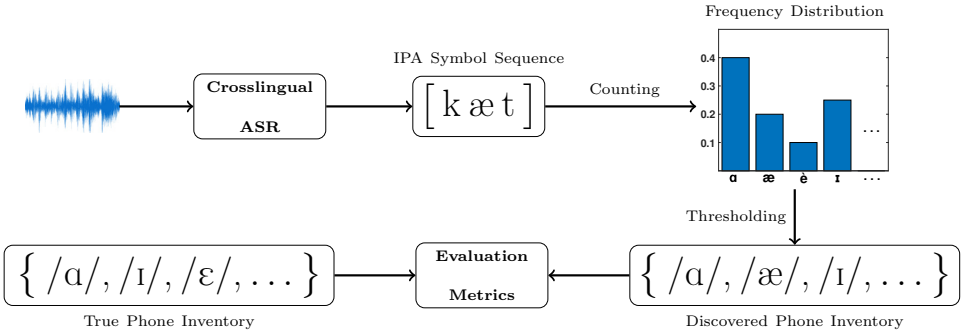


Figure 3: High-level workflow illustrating how a crosslingual (zero-shot) ASR is applied to phone inventory discovery of an unknown language.

The ultimate goal of our works is to automatically discover the phone inventories of an unknown language [**MA**]. Note that in this paper, we usually mention **phonetic inventories**. We use that term, in a slight abuse of terminology, to refer to any of the two different types of inventories – a **phone inventory** and a **phone token inventory**. The **phone inventory** of a language is the set of phones generated by a language-dependent Grapheme-

to-Phone (G2P). The **phone token inventory** is the set of IPA characters that make up those phones.

Frequency-based inventory discovery. To probe the zero-shot system's usefulness in phonetic inventory discovery, we propose a straightforward approach: we assume that all symbols that occur "frequently enough" in the zero-shot system's output for a given target language together constitute the target language's phonetic inventory. But what is "frequently enough"? Here, we simply count the number of occurrences of each symbol in the recognized transcripts of an unseen language and set a threshold, where any symbol with a frequency above the threshold is considered to be part of the language's phonetic inventory. To ensure that the approach is not strongly dependent on the amount of available test data, we convert the absolute occurrence counts into relative frequencies, dividing each count by the total number of recognized symbols for that language.

Phone inventory discovery cannot be performed using a phone-token ASR: there are many different ways in which the phone tokens generated by an ASR could be recombined into phones, and the different combination methods have different phone discovery accuracies. Both types of inventory discovery (phone inventory and phone token inventory) can be performed using phone-based ASR: the only difference between them is that, for phone token discovery, the phones in the ASR output are split into phone tokens prior to application of the process in Figure 3.

4. Experimental setup

This section presents the details of our experiments. We start by introducing the corpora and languages used in our study in Section 4.1. Section 4.2 describes the setup of our monolingual, multilingual and crosslingual ASR experiments. Finally, we describe the phonetic inventory experiments in Section 4.3.

4.1. Corpora

We selected 13 languages for this study which are shown in Table 1, identical to those used in our previous work [21, 22]. These languages were chosen for the diversity of their phoneme inventories. For instance, Bengali was chosen as a representative language with voiced aspirated stops, Javanese for its slack-voiced (breathy-voiced) stops, Zulu has clicks, and Mandarin, Cantonese, Vietnamese, Lao, and Thai are tone languages with very

Table 1: Speech data used in the experiments. Train and Eval columns are presented in the number of hours. Vow and Con indicate the number of distinct phones used to describe vowels and consonants, respectively. Uniq denotes the number of IPA phones only existing in that language. The symbol ‡ indicates tone languages.

Corpus	Language	Train	Eval	Vow	Con	Uniq
GlobalPhone	Czech	24.0	3.8	10	25	5
	French	22.8	2.0	19	26	9
	Spanish	11.5	1.2	6	24	4
	Mandarin‡	14.9	1.6	60	33	41
	Thai‡	22.9	0.4	108	35	17
Babel	Cantonese‡	126.6	17.7	110	25	88
	Bengali	54.5	9.8	19	33	13
	Vietnamese‡	78.2	10.9	204	25	130
	Lao‡	58.7	10.5	106	61	83
	Zulu‡	54.4	10.4	15	54	19
	Amharic	38.8	11.6	8	52	9
	Javanese	40.6	11.3	11	23	1
	Georgian	45.3	12.3	5	30	2

different tone inventories (Zulu also has lexical tones, but our grapheme-to-phoneme transducer for Zulu does not label the tones of each syllable). Table 1 reports the number of distinct phones in the ASR segment inventory for that language. Each phone consists of a base phone (a speech sound associated with a temporal segment), possibly modified by diacritics marking secondary articulation, lengthening, lexical stress, or lexical tone, e.g., the high-, rising-, dipping- and falling-tone /i/ in Mandarin are described by four different phones [i˥], [i˧˥], [i˨˩˧], and [i˥˩], respectively. Phones may be affricates if affricates are phonemic in the source language, e.g., the five Amharic phones /t/,/t'/,/ʃ/,/tʃ/,/tʃ'/ are coded using only three IPA symbols: [t], [ʃ], and ['']. Note that the high number of vowels in tone languages is due to different combinations of tone modifiers (e.g., Vietnamese and Cantonese each have six tones per vowel). We obtained the IPA phonetic transcripts by leveraging the LanguageNet Grapheme-to-Phone (G2P) models [95].

Speech data for these languages were obtained from the GlobalPhone [96] and IARPA Babel corpora. GlobalPhone has a small number of recording

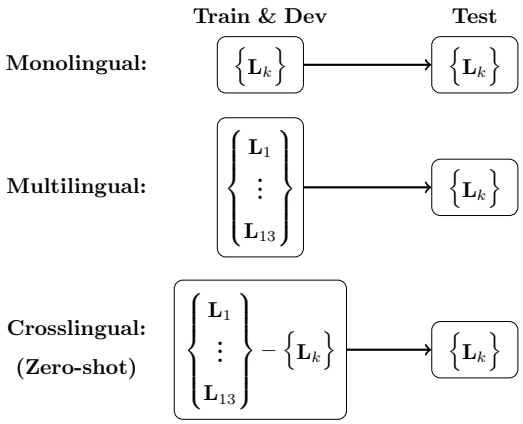


Figure 4: Three ASR experimental scenarios in our study: monolingual, multilingual, and crosslingual. L_k indicates k-th language in our experimental dataset.

hours for each language - about 10 to 25 hours - and represents a simpler scenario with limited noise and reverberation. On the other hand, Babel languages have more training data, ranging from 40 to 126 hours, but these databases are significantly more challenging due to the recordings having been made in a larger variety of acoustic environments. We used standard train, development, and evaluation splits for Babel. For GlobalPhone languages, whenever the documentation provides standard split information, we use these – otherwise, we chose the first 20 speakers' recordings for development and evaluation sets (10 speakers each), and train on the rest. Table 1 gives an overview of the number of hours of training and test (Eval) material for each language.

4.2. ASR experiments

The ultimate aim of our work is to discover phonetic inventories [**MA**]. In order to be able to asses the strengths and weaknesses of the crosslingual ASR approach, we also need baseline systems that have learned a robust phonetic representation of the target language. They will serve as oracles for evaluating phonetic inventory discovery. Hence, our ASR experiments are divided into three scenarios (mimicking our previous setups [21, 22]; see also Figure 4):

1. monolingual (mono) – 13 different ASR systems are trained and tested each on the same language;

2. multilingual (multi) – a single ASR is trained and tested on all languages; and
3. crosslingual (cross) – 13 different ASR systems are each trained on 12 languages and tested on a held-out language. We also refer to this system as a **zero-shot** ASR, as the target language is never seen in training.

E2E ASR experiments. We used an E2E ASR system whose inputs were combined filterbank and pitch features [97] with a frame length of 25ms and a frame shift of 10ms. The ASR was trained with joint CTC and attention objectives [98], employing the ESPnet toolkit [99]. The setup is based on the system described in [100] that uses a transformer architecture. Transformer-based models have been shown to produce state-of-the-art results on several Natural Language Processing (NLP) tasks [101, 102], and has recently been adopted in speech processing [100]. We use 12 transformer layers for the encoder and 6 for the decoder. Each transformer layer consists of 4 attention heads with a dimension of 256 and 2048-dimensional feed-forward layer. Prior to the transformer layers, the encoder uses 2 2D convolutional layers with a stride of 2 each to sub-sample the signal with a factor of 4. We use label smoothing with a factor of 0.1. To accommodate the transformer's $O(n^2)$ memory requirements, we discard all training and test utterances which exceed 30 seconds of speech (about 1.7% of utterances).

Two types of E2E experiments were carried out, one using IPA phones as speech representation (E2E P) and one which used phone tokens as speech representation (E2E PT) **[RQ1]**. The E2E P output alphabet is the union of IPA phones in all training languages, hence it cannot recognize novel target-language phones in a crosslingual setting. The E2E PT output alphabet is the union of all unique unicode characters representing IPA base phones or diacritics in any training language. It is not always clear how we should segment the output of the E2E PT system in order to measure phone error rate. We take this phenomenon into account during the evaluation, by measuring **Phone Token Error Rate (PTER)** rather than Phone Error Rate (PER). PTER treats every modifier symbol as a separate token. That means that if the reference transcript has a vowel with a tone symbol, and the ASR recognized the vowel correctly but omitted the tone, we would count this as

one correct token and one deletion instead of a single substitution [6].

Hybrid ASR experiments. The hybrid ASR was implemented using the Kaldi toolkit [103]. It consisted of a Factorized Time-Delay Neural Network (TDNNF) AM [104], with 12 layers, where the hidden layers had a dimension of 1024, and bottleneck dimension of 128. Resnet-style bypass connections were made between each pair of consecutive TDNNF layers. This TDNNF AM was trained with the Lattice-Free Maximum Mutual Information (LF-MMI) criterion [105] for 4 epochs. The frame-level phone alignments used as supervision for the AM model training were obtained by forced-alignment with a GMM-HMM AM trained beforehand. The input features were 40-dimension high-resolution MFCCs, 3-dimension pitch features [97] and 100-dimension i-vectors [106]. We used the frame length of 25ms and frame shift of 10ms. In TDNNF training, we used Kaldi's Wall Street Journal recipe[7] standard hyperparameters without further tuning in our experiments. An identical TDNNF structure was used for each of the experiments in this study.

The LM in the hybrid ASR system was an n-gram phonotactic LM estimated using the SRILM toolkit [107]. IPA phonetic transcripts converted from the GlobalPhone and Babel training graphemic transcripts are used to train the phonotactic LM. We implemented uni-gram, bi-gram, and tri-gram LMs to investigate the effect of imposing different amounts of phonotactic constraint on a hybrid ASR system in recognizing phone sequences. To study the systems behavior under an upper bound of phonotactic constraint, we also decoded using a word-level tri-gram LM and converted the word lattices to phone lattices before scoring.

The hybrid system allows us to combine AM and LM trained on a different mix of languages. We leverage that possibility by combining a monolingual target language model with crosslingual AM, which shows the effect of adding oracle-level phonotactic knowledge to an otherwise zero-shot system [**RQ3**]. This demi-oracle experiment helps us to better understand the degree of

[6]Note that we introduced PTER in our best effort to make P and PT systems comparable. The only alternative we see is imposing a decoding constraint on the E2E PT system that forces it to choose paths that only contain allowed phone token sequences that correspond to phones in an inventory known up-front. That is however equivalent to introducing a language model and FST decoding. We decided not to pursue LM integration with E2E models for the sake of avoiding additional complexity.

[7]`wsj/s5/local/chain/tuning/run_tdnn_1g.sh`

universality of AM representations. We also decode using multilingual AM and monolingual LM to understand the effect of the phonotactic model on an AM that has learned the target language phonetic representation.

4.3. Phonetic inventory discovery

Our phonetic inventory discovery method is based on counting the phone tokens or phones, in crosslingual ASR transcripts. To convert the occurrence counts into frequencies, we divide the number of observed phone tokens by the total number of phone tokens in the transcripts. We checked the range of relative frequency threshold values between 1e-4 and 1e-2 and select the ones that give best performance with crosslingual ASR systems. We report the final threshold values in section 5.3.x The optimal threshold selection might be somewhat language-dependent: in languages with a large number of phones (e.g., Cantonese) the relative phone frequencies will necessarily be lower (in expectation) than in languages with a smaller phone vocabulary. As we don't know the inventory size when discovering inventories, we select the threshold based on the best overall performance across the 13 languages.

We test four crosslingual systems: the E2E P system; E2E PT system; and two hybrid systems with crosslingual LMs – uni-gram (weaker phonotactic constraint) and tri-gram (stronger phonotactic constraint). To put the crosslingual ASR performance in perspective, we also check the performance of four other systems that have either full or partial oracle knowledge about the target language. The first three are the best performing systems of each kind: multilingual E2E P, multilingual E2E PT, and the monolingual hybrid system with a tri-gram word-level LM. Additionally, we test the hybrid system with a crosslingual AM, but using a monolingual LM trained on the target language. The last system allows us to probe the significance of knowing the target language's phonotactics in the task of phone inventory discovery – since in actual applications we would never know them *a priori*, it also reveals the limitations of this phone discovery method, as ASR depends on language model for accurate recognition.

We perform the phonetic inventory discovery for all thirteen languages presented in Section 4.1 [**MA**].

5. Results

In this section we present the experimental results. We start with monolingual, multilingual and crosslingual (zero-shot) ASR in Section 5.1. Then,

we cast more insight on what the crosslingual system has learned by visualising its phonetic confusion matrix in Section 5.2. Finally, we show the results for phonetic inventory discovery in Section 5.3.

5.1. ASR experiments

To provide the basis for answering [**RQ1**] and [**RQ3**], we collect the results from all our ASR experiments in Table 2. Each row represents the aggregated scores over the 13 test languages. For each system we show PTER, and for the phone-level systems also the PER. Additionally, we provide a breakdown of the percentage of insertion, deletion, and substitution errors in PTER. Note that the E2E PT and hybrid P results are based on our previous experiments in [21] and in [22], but our previous publications did not aggregate scores across all languages. E2E P results have not been previously published.

Below, we present the high-level insights that we learned from performing the ASR experiments.

Phones are generally preferable to phone tokens in mono- and multilingual models [RQ1]. In order to understand whether phones or phone tokens are preferable, we have to analyse their performance across mono-, multi-, and crosslingual scenarios. We first compare the E2E systems indicated with P and PT in Table 2. In the group of systems that have seen the target language during training (*mono*, *multi*), we observe that in terms of PTER, the best overall system is *multi* E2E P, closely followed by *multi* E2E PT. The relative improvement of E2E P vs E2E PT system is consistent when we compare the *mono* variants of the E2E P and E2E PT systems. We conclude that the E2E P system outperforms the E2E PT system. Possibly, the lower number of errors of the E2E P system is due to a larger constraint on its output token space – the number of recognized symbol combinations is restricted due to larger vocabulary units. Therefore, it seems that phone-level modeling is beneficial when the test language is in-domain.

Phone tokens might be generally preferable to phones in crosslingual models [RQ1]. Turning to the zero-shot systems section of Table 2, we see that the phone-token E2E scored the best at PTER across all crosslingual systems, with phone E2E as a second and hybrid as a close third. We find that the PTER is consistently smaller in the phone-token system than in the phone system for all but two languages: Mandarin and Bengali. Whereas the error differences are small (81.9% vs 83.2% and 83.6%), this result suggests that the flexibility provided by smaller vocabulary units is helpful in

System	AM	LM	PER	PTER	INS%	DEL%	SUB%
Systems with knowledge of the target language							
E2E P	multi	-	31.8	29.3	19.2	29.0	51.8
E2E PT	multi	-	-	32.4	17.8	31.7	50.5
Hybrid P	mono	wtg-mono	33.6	44.1	26.6	28.8	44.6
	multi	wtg-mono	37.2	48.8	26.8	26.3	47.0
		wtg-multi	37.3	48.2	27.7	24.9	47.5
	mono	tg-mono	38.1	36.7	11.3	44.4	44.3
	multi	tg-mono	41.0	39.9	12.9	38.8	48.3
	mono	bg-mono	43.5	41.0	8.5	52.1	39.5
	multi	tg-multi	43.5	40.9	9.7	44.9	45.4
E2E P	mono	-	43.9	39.4	16.1	28.2	55.7
Hybrid P	multi	bg-mono	45.4	43.4	9.1	45.9	45.0
E2E PT	mono	-	-	47.1	14.7	29.3	56.0
Hybrid P	multi	bg-multi	51.9	49.4	4.2	63.2	32.7
	mono	ug-mono	54.9	52.6	3.9	70.3	25.8
	multi	ug-multi	63.9	62.6	1.9	77.5	20.6
		ug-mono	64.0	52.6	3.1	72.4	24.5
	cross	tg-mono	72.9	74.2	2.3	68.2	29.5
		ug-mono	80.9	82.0	1.1	79.6	19.3
Zero-shot systems							
E2E PT	cross	-	-	81.9	7.9	21.8	70.3
E2E P	cross	-	82.1	83.2	9.7	20.9	69.4
Hybrid P	cross	bg-cross	83.2	83.6	1.8	59.4	38.8
		ug-cross	85.3	84.6	1.2	72.5	26.3
		tg-cross	86.1	93.6	5.2	40.0	54.7

Table 2: Comparison of all ASR systems trained in this study. The results are aggregated across all 13 languages. For E2E systems, P stands for phone symbols as recognition units, and PT for phone tokens (individual IPA symbols). For hybrid systems, ug, bg, and tg indicate (phone) uni-gram, bi-gram, or tri-gram LM; and wtg stands for word tri-gram LM. For each system, we provide the phone token error type distribution (insertion, deletion, substitution) as the percentage of the total number of phone token errors. We do not report PER for PT systems, as it cannot be computed in a straight-forward manner due to ambiguities in decoding phone tokens into phones.

crosslingual recognition.

Strong phonotactic model hurts crosslingual ASR [RQ3]. Ultimately, we want to understand the effect of phonotactics on phonetic inventory discovery. For this reason, we include and assert our previous finding from [22] here. The summary in Table 2 clearly shows that the *tg-cross* LM has the worst scores all the systems in this study. Interestingly, the E2E systems scored slightly better both in terms of PER and PTER, which suggests that the E2E systems might have a certain degree of robustness towards phonotactics mismatch. The best performance for *cross* AM is observed when we introduce target language phonotactic *tg-mono* LM – it approximates the lower bound of PER/PTER for a crosslingual model.

Language	Mono		Cross		Multi	
	PTER†	PER$^{\diamond}$	PTER†	PER$^{\diamond}$	PTER†	PER$^{\diamond}$
Czech	26.4	36.4	65.8	71.5	8.1	12.5
French	29.5	30.5	61.7	62.8	11	11.5
Spanish	27.6	35.8	76.3	69.1	9.1	11.6
Mandarin	46.3	40.9	85.9	80.3	17.2	18.1
Thai	46	39.1	76.2	115.8	18.1	18.8
Cantonese	36.6	30.6	76.9	78.7	29.8	28.3
Bengali	41.2	34.7	81.4	74.0	34.6	34.0
Vietnamese	52.3	37.4	75	78.6	35.4	30.0
Lao	58.6	37.4	77.9	74.2	34.1	31.3
Zulu	41.7	43.3	77.3	77.5	35.8	35.4
Amharic	52.1	51.9	75.2	74.0	33.9	33.2
Javanese	53	51.4	99.6	76.1	41	38.3
Georgian	43.9	43.9	76.4	77.5	32.4	30.4

Table 3: E2E recognition results for all 13 languages separately, for each of the 3 evaluation scenarios: Mono(lingual), Cross(lingual), Multi(lingual). Two systems are evaluated: phone-token system (†) for which we report PTER, and phone system (◊) for which we report PER. Besides the differences in label set construction, the systems are otherwise identical.

Phone tokens are preferable in most tone languages that we studied [RQ1]. We present a per-language, side-by-side comparison of E2E PT systems PTER and E2E P systems PER in Table 3. While PER and PTER cannot be directly compared, we can analyse whether they have the

same trends across different languages and different training scenarios. Most monolingual systems trained on tone languages tend to have lower PER than PTER (Mandarin, Thai, Cantonese, Vietnamese, Lao), indicating that the phone-based system is more adequate, while the reverse is true for some non-tonal language systems (Czech, French, Spanish). This suggests that IPA recognition of tone languages benefits from the additional constraint on the possible IPA symbol sequences in a low resource scenario. Looking at the differences between the multilingual E2E and monolingual E2E, phone systems generally follow the same trend as phone-token systems, but the improvements in PER (vs. the improvements in PTER) are relatively smaller. It seems that larger units (phones) impose a constraint on the network (i.e., an inductive bias) that helps monolingual systems perform better in many languages, reducing the opportunities for improvement with multilingual data.

Both phone and phone token crosslingual systems may confuse tonal and non-tonal vowels [RQ1]. Our thirteen test languages are naturally divided into four categories, according to two binary variables: (1) Globalphone corpus versus Babel corpus, and (2) tonal versus non-tonal. The Globalphone corpora (Czech, French, Spanish, Mandarin, and Thai) are read speech, recorded with little acoustic noise; the Babel corpora are a mixture of read and spontaneous speech, in a mixture of recording conditions. The tonal languages include those from the Sino-Tibetan, Kra-Dai, Austronesian, and Niger-Congo language families (Mandarin, Cantonese, Thai, Lao, Vietnamese, and Zulu), while the non-tonal languages include those from the Indo-European, Afro-Asiatic, Austronesian, and Kartvelian families (Czech, French, Spanish, Bengali, Amharic, Javanese, and Georgian). Among the Globalphone languages, Indo-European languages have considerably lower PTER and PER than tonal languages, under Monolingual, Cross-lingual, and Multilingual settings. Thai is a particular outlier, with 115.8% cross-lingual PER. *Post-hoc* evaluation of ASR transcripts suggests that the system incorrectly assumed that Thai is a non-tonal language, and exclusively recognized non-tonal phones; but note that the phone-token ASR did not make this mistake. Among the Babel languages, Javanese is an outlier, with 99.6% cross-lingual PTER. Analysis of the phonetic transcript output of the phone-token ASR showed that the phone-token ASR mistakenly treated Javanese as a tone language, appending tone modifiers to vowels, where there should be none; but note that the phone ASR did not make this mistake. We know of no reason why the system should choose to mis-identify the language families of these two languages under these two particular test conditions.

We can note only that we have one example of a phone-based ASR that deleted lexical tone cross-lingually, and one example of a phone-token ASR that inserted lexical tone cross-lingually.

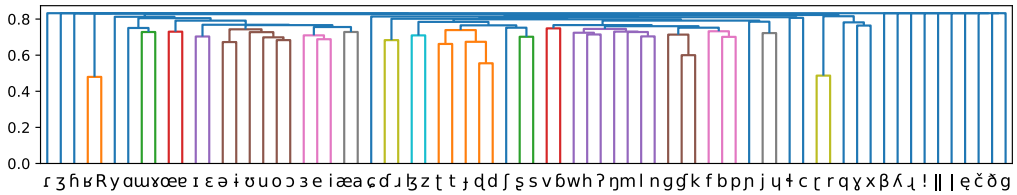
5.1.1. General remarks

Below, we present observations that we found interesting when analysing the ASR experimental results, but not directly related to our core research questions.

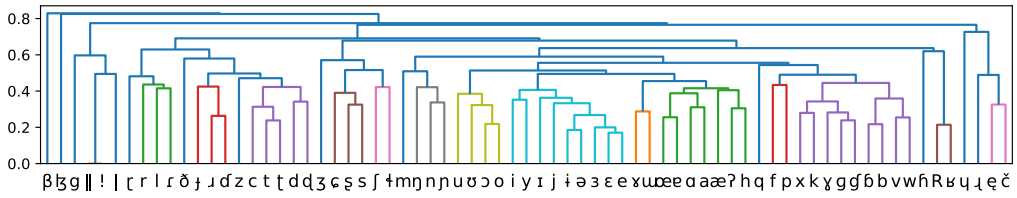
Monolingual hybrid system is more robust than multilingual hybrid system. We observe the best performance with *mono* AM and word-level *mono* LM. We expected that adding multilingual phonetic transcripts to n-gram LM training would not help (i.e., we expect that it acts as noise and "steals" the probability mass from target language n-grams during pruning). However, the acoustic model's degradation in *multi* is not consistent with other works that have attempted multilingual training: we suspect this is due to the model sharing all the parameters (especially the output layer) between all languages; whereas [108, 109, 110] observed multilingual improvements when each language had its own output layer.

The distribution of error types shifts with training setups and the phonotactic model's strength. E2E *multi* systems are somewhat "balanced" with about 50% substitutions, 30% deletions, and 20% insertions. The three best hybrid systems have a strong phonotactic constraint (i.e., they use a word-level LM) and produce a similar error distribution, but with insertions and deletions equally likely. As the phonotactic constraint weakens, we observe that the share of deletions drastically increases in the hybrid systems, achieving a 70-80% share of all error types when unigram phone LMs are used. The differences between the E2E and hybrid models are very pronounced in the *cross* scenario: in E2E, there are relatively many more substitutions than insertions and deletions, while for the hybrid systems, the share of deletions is dominant at 40-72.5%, depending on the phone LM's order. These findings mean that even if the error rates of the E2E and hybrid models are similar, the transcripts will look rather different, with generally much shorter transcripts for the hybrid model.

Multilingual E2E is significantly more robust than monolingual E2E. We previously found that multilingual E2E provides significant gains with respect to mono in [21]. Here, we observe the same trend in the E2E P system measured with PER in Table 3.



(a) Multilingual system (coloring threshold 0.75).



(b) Zero-shot system (coloring threshold 0.35).

Figure 5: The results of the hierarchical clustering of the phone confusion matrix, presented as dendrograms. The height of each bar corresponds to a distance score (i.e., a lower bar indicates a higher similarity). We selected the threshold for cluster coloring to highlight some of the more meaningful similarities found between phones.

The dendrograms resulting from the hierarchical clustering of the phone token confusions of the multilingual phone E2E system are shown in Figure 5a and for the zero-shot phone E2E system in Figure 5b [**RQ2**]. The hierarchies in the figure can be understood as groups of phone tokens that tend to have similar patterns of confusions – the symbols that are clustered "lower" in the Y axis of the figure have more similarity (smaller JSD).

At a glance we see that a much higher threshold is required to cluster phone tokens in the multi system (Figure 5a) than in the zero-shot system (Figure 5b); this is because there are fewer recognition errors in the multi system, and the clustering algorithm therefore considers its phone tokens to be more distinct. The high error rates obtained in the zero-shot ASR experiments (see the *Zero-shot systems* section of Table 2) could suggest that the systems' outputs are close to random. However, we observe the pat-

terns in the zero-shot system's confusions to be quite meaningful from an articulatory feature point of view (as indicated by different phones within a cluster of the same color) and similar to the multilingual system's patterns. For instance, both the multilingual and the zero-shot systems cluster vowels separately from consonants. The phone tokens in the zero-shot system are further clustered, with a clustering threshold of about 0.55, into approximately eight categories defined by articulatory features, with meaningful sub-categories. Moving from left to right through Figure 5b, we find that the first group contains voiced fricatives /β,ɮ/, a voiced plosive /ɡ/, and the three clicks /ǀ,ǃ,ǁ/. The next group contains coronal laterals, taps, trills, plosives, and non-sibilant fricatives /ɽ,r,l,ɾ,ð,ɻ,ɹ,ɗ,z,c,t,ʈ,d,ɖ/. The third group contains coronal sibilant fricatives /ʒ,ɕ,ʂ,s,ʃ,ɬ/. The fourth group contains the nasal consonant (/m,ŋ,n,ɲ/). The fifth group contains the vowels; these are further subdivided into non-low back rounded vowels /u,ʊ,ɔ,o/, the non-low front vowels and glides /i,y,ɪ,j,ɨ,ə,ɜ,ɛ,e/, the non-low back unrounded vowels /ɤ,ɯ/, the low vowels /œ,ɐ,ɑ,a,æ/, and two glottal consonants /ʔ,h/. The sixth group contains the non-coronal obstruents, further subdivided into the pharyngeal stop /q/, the unvoiced labials /f,p/, the unvoiced velars /x,k/, the voiced velars /ɣ,g,ɠ/, and the voiced labials /v,w/. The seventh group contains glottal and uvular fricatives and trills /ɦ,ʀ,ʁ/. The last group contains a set of palatal and retroflex phones: two approximants /ɥ,ɻ/, a vowel /ɐ̟/, and an afffricate /č/ that are apparently sufficiently different from the other phones to be separately clustered.

The UMAP projection of the crosslingual ASR confusion matrix is shown in Figure 6, with each phone represented as a data point (projected from a vector describing the label categorical probability distribution) and colored/shaped with its corresponding manner of articulation. The 2-D projection allows to further inspect the learned similarities between various phones [**RQ2**], and UMAP – unlike other projection methods such as T-SNE – attempts to preserve the global (in addition to local) structure in the projection. In principle, the phones that are frequently confused together should lie close to each other on the projected surface. Indeed, this display shows the same local clustering structure that is visible in Fig. 5b, but enhanced with global relationships: vowels are on the lower right-hand side of the plot; moving towards left and top, there are tongue body and labial consonants, tongue-blade consonants are on the left-hand side, nasals are in the lower center, and clicks are at the top. Among the vowels, back rounded vowels are at the top, low vowels are in the center, and the high vowels and back

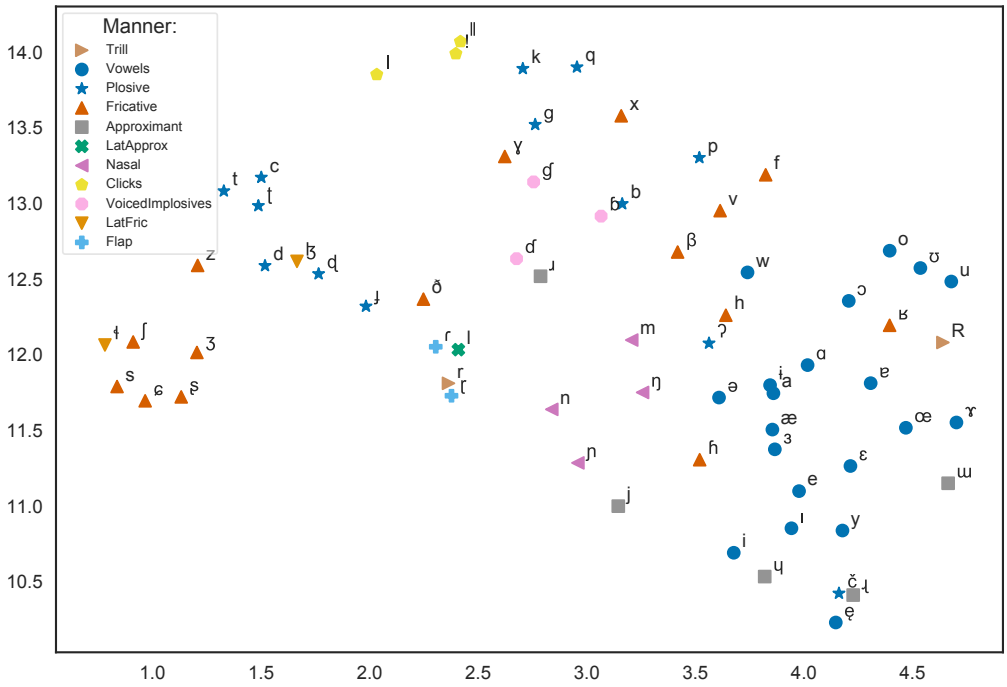


Figure 6: UMAP projection of the phone confusion matrix for the cross-lingual E2E P system. Each manner of articulation class has a unique tick and a unique color to help discriminate them visually.

unrounded vowels are toward the bottom.

The results in Section 5.1 show that in general the phone recognition results of the crosslingual model are rather poor. However, as these analyses of the phone confusions show, many of the phones in the target language are confused with phones that are highly similar to the target phone: phones that share many articulatory features, e.g., the vowels or nasals, are clustered together (see Figure 6). This finding resembles that of a foreign language learner, for whom one of the tasks is to learn new phonemes that are not part of their native language and consequently often not getting the articulation entirely right (yet). It also explains the massive difference between the zero-shot and multilingual systems' performance - while the zero-shot system is able to learn some phonetic representations, the presence of the target language's data closes the gap between the learned articulation styles.

System	Precision	Recall	F1
Zero-shot systems			
Hybrid, cross AM, cross phone 3-gram	65.7	67.6	66.7
E2E, phone tokens, cross AM	70.1	65.1	67.5
E2E, phones, cross AM	69.7	65.6	67.6
Hybrid, cross AM, cross phone 1-gram	69.7	67.4	68.6
Systems with oracle knowledge			
Hybrid, cross AM, mono phone 3-gram	100.0	69.2	81.8
E2E, phone tokens, multi AM	99.5	81.9	89.8
E2E, phones, multi AM	99.7	82.1	90.1
Hybrid, mono AM, mono word 3-gram	100.0	82.8	90.6

Table 4: Phone token inventory discovery performance of various ASR systems. The phone level systems transcripts were tokenized into phone tokens for fair comparison. All systems discover phone tokens at their relative frequency of 0.004.

5.3. Phonetic inventory discovery

The phonetic inventory discovery results, aggregated over all test languages, are presented in Table 4 for **phone token inventory discovery** and in Table 5 for **phone inventory comparison** [**MA**]. To score the systems, we used precision, recall, and F1-score measures: to compute them, we count the number of True Positive (TP): the symbols correctly assigned to the phone inventory; False Positive (FP): the symbols that were incorrectly indicated as being part of the target language's phone inventory; and False Negative (FN): the symbols that were not assigned to the predicted inventory, but should have been as they are present in the true phone inventory. The cumulative size of phone token inventories was 448, and the cumulative size of phone inventories was 1127 (see Table 6 for per-language details).

We start with the analysis of phone token inventory discovery in Table 4. We observe that both E2E crosslingual ASR systems are on par with 67.5% and 67.6% F1-score, suggesting that phone and phone-token distinction is not relevant for discovering the phonetic inventory of a language [**RQ1**]. Interestingly, we see that the hybrid system with a weak phonotactic model performs best (68.6%), and the one with a strong phonotactic model performs worst (66.7%). This finding is consistent with our previous hypothesis that learning non-target languages' phonotactics is harmful for a zero-shot ASR

System	Threshold	Precision	Recall	F1
Zero-shot systems				
E2E, phones, cross AM	min	14.1	32.9	19.7
	0.0001	20.3	31.3	24.7
	0.001	35.6	27.6	31.1
	0.002	43.2	25.8	**32.3**
	0.004	52.4	22.4	31.4
Hybrid, cross AM, cross phone 1-gram	min	12.0	34.1	17.8
	0.0001	17.5	32.4	22.7
	0.001	36.9	27.9	31.7
	0.002	47.9	25.6	**33.4**
	0.004	59.9	22.3	32.5
Systems with oracle knowledge				
Hybrid, cross AM, mono phone 3-gram	**min**	98.0	43.7	**60.4**
	0.0001	98.1	42.1	59.0
	0.001	99.0	37.0	53.9
	0.002	99.0	34.3	51.0
	0.004	98.8	30.3	46.3
E2E, phones, multi AM	min	55.8	85.5	67.5
	0.0001	89.6	80.3	**84.7**
	0.001	99.9	62.7	77.1
	0.002	100.0	53.2	69.5
	0.004	100.0	42.9	60.1
Hybrid, mono AM, mono word 3-gram	**min**	99.3	94.6	**96.9**
	0.0001	99.5	85.9	92.2
	0.001	100.0	64.0	78.0
	0.002	100.0	54.0	70.2
	0.004	100.0	43.5	60.6

Table 5: Phone inventory discovery performance of various ASR systems, depending on the minimum relative frequency threshold needed for discovery. *min* indicates the frequency of the least frequent phone in the transcripts of the target language.

system [22] [**RQ3**]. If the phonotactic model had perfect knowledge of the phonotactics of the target language, the F1-score raises to 81.8%, which is mostly due to a perfect precision score of 100%. The reason for the high precision is that the monolingual phonotactic model has no knowledge of those IPA symbols present in the other training languages but not present in the target language. So even though the AM can predict these phones, they are never selected as a part of the shortest lattice path during ASR decoding.

If we train multilingual E2E ASR models with knowledge of the target language, we again observe very high precision scores, and a substantial increase in recall scores as well. So, despite lack of explicit constraint on the phonotactic information, we see an increase in F1 scores for the E2E ASR multi models. This is consistent with the hypothesis of [22] that the encoder-decoder transformer systems are learning a representation of the target languages' phonotactics [**RQ3**]. Finally, the best result of 90.6% F1-score is obtained by a fully monolingual hybrid system with a word-level trigram LM, likely due to its stronger phonotactic constraint imposed by the presence of a decoding graph. Interestingly, even this system fails to recall about 17% of the IPA symbols in all languages due to ASR errors. This suggests that some phone tokens are not well modelled in the AM and/or are still so rare in the test material that they do not occur [**MA**].

We move onto phone inventory discovery in Table 5 [**MA**]. This task is more difficult than phone token discovery, as indicated by generally lower F1 scores achieved by the zero-shot systems. This is explained by much larger sizes of the inventories, in which we expect there to be even more rare symbols than in the phone token case. The hybrid ASR again outperforms E2E with higher F1 scores, boosted mostly by a higher precision.

Table 5 presents the results for different discovery threshold values, which reveals interesting differences between the behaviour of the E2E and hybrid ASR. Had we fixed the best crosslingual phone discovery threshold at 0.002 for the systems with the oracle knowledge, we'd observe a lower phone inventory discovery performance at 70.6% for the best oracle system (i.e., the hybrid mono AM with mono word 3-gram). However, that threshold is too high and filters out a significant portion of the phones. Indeed, as we gradually lower the threshold towards the frequency of the rarest phone in the true inventory, we observe an F1 score that gradually increases to a maximum of 96.9% F1.

The role of phonotactics [**RQ3**] in phone inventory discovery is even more pronounced than in phone token inventory discovery. The hybrid system with

	TP	FP	FN	Precision	Recall	F1
Cantonese	29	7	4	80.6	87.9	84.1
Bengali	26	14	7	65.0	78.8	71.2
Vietnamese	29	9	13	76.3	69.0	72.5
Lao	27	9	5	75.0	84.4	79.4
Zulu	26	12	19	68.4	57.8	62.7
Amharic	24	14	7	63.2	77.4	69.6
Javanese	25	8	8	75.8	75.8	75.8
Georgian	22	16	6	57.9	78.6	66.7
Czech	24	11	6	68.6	80.0	73.8
French	19	7	23	73.1	45.2	55.9
Mandarin	15	7	21	68.2	41.7	51.7
Spanish	19	12	11	61.3	63.3	62.3
Thai	17	5	16	77.3	51.5	61.8
ALL	302	131	146	69.7	67.4	68.6

Table 6: Per-language breakdown of **phone token inventory discovery** performance for the zero-shot hybrid system with 1-gram LM (threshold = 0.004).

a crosslingual AM and an oracle 3-gram phone LM has an F1-score almost twice that of the same hybrid system using crosslingual LM. We also see that the oracle E2E P system scores worse than the oracle hybrid system due to false positives, which are nearly impossible for the oracle hybrid system, because of its strong phonotactic LM.

To investigate whether the task of automatic phonetic inventory discovery is dependent on the language and language characteristics, we investigated the phonetic inventory discovery performances per language in more detail [**MA**]. Table 6 shows the True Positive (TP), False Positive (FP), False Negative (FN)), precision, recall, and F1 scores on the phone token inventory discovery task for the crosslingual hybrid system with 1-gram phone LM, i.e., the best system from Table 4, for each of the 13 languages. As Table 6 shows, there are large performance differences between languages: the difference in F1 scores between the lowest scoring language Mandarin and the highest scoring language Cantonese is 32.4%.

Comparing these results to Table 1 shows that many of the best performing languages are, surprisingly, the languages with the highest number of unique phones (in decreasing order of number of unique phones: Viet-

	TP	FP	FN	Precision	Recall	F1
Cantonese	19	44	116	30.2	14.1	19.2
Bengali	30	25	22	54.5	57.7	56.1
Vietnamese	13	27	216	32.5	5.7	9.7
Lao	22	30	145	42.3	13.2	20.1
Zulu	29	23	40	55.8	42	47.9
Amharic	29	14	31	67.4	48.3	56.3
Javanese	27	51	7	34.6	79.4	48.2
Georgian	26	28	9	48.1	74.3	58.4
Czech	23	18	12	56.1	65.7	60.5
French	21	7	24	75	46.7	57.5
Mandarin	16	13	77	55.2	17.2	26.2
Spanish	19	20	11	48.7	63.3	55.1
Thai	15	14	128	51.7	10.5	17.4
ALL	289	314	838	47.9	25.6	33.4

Table 7: Per-language breakdown of **phone inventory discovery** performance for the zero-shot hybrid system with 1-gram LM (threshold = 0.002).

namese, Cantonese, Lao), with the exception of Mandarin which has the 4th highest number of unique phones; while those languages with only a few unique phones (in increasing order of number of unique phones: Georgian, Spanish, Czech) typically perform worse, but with the exception of Javanese which has only 1 unique phone but has an F1 score of 75.8%.

Turning to Table 7, we observe that the five languages with the worst phone inventory discovery rates are the tone languages Vietnamese, Thai, Cantonese, Lao, and Mandarin. The low F1 scores of the tone languages are primarily caused by large false negative counts, specifically, by the inability of a cross-lingual ASR to recognize most tone-marked vowels. Our phone notation treats the lexical tone as a diacritic modifier of the vowel. About two thirds of the lexical tone contours occur in only one language each, therefore about two thirds of the vowels in a tone language cannot be discovered using cross-lingual ASR. Non-tone languages are generally characterized by higher recall than precision (Bengali, Javanese, Georgian, Czech, Spanish) due to more false positives. The exceptions are French and Amharic, which tend to have the least number of false positives (along with Mandarin and Thai). In fact, French is the only language which achieves a higher F1 score in phone

inventory discovery than in phone token inventory discovery.

The differences between languages are potentially due to some sounds being easier to recognize cross-linguistically than others. Phoneme inventories of languages are not created in a "random" fashion, rather phoneme inventories are build up in an "ordered" fashion [111]. For instance, all languages have at least two of the three voiceless plosives /p, t, k/; these phonemes are considered "simple" sounds, i.e., sounds that are easy to produce. The simplicity of a phoneme has many different definitions, but one definition that is common in the literature is age of acquisition (expressed in months): It has been suggested that a phoneme's typical age of acquisition is a measure of its articulatory complexity [112]. It has been demonstrated that age of acquisition is a better predictor of articulation errors in dysarthria than is either manner or place of articulation [113], and a similar ranking of the difficulty or simplicity of phones can be observed in the transcription errors of non-native transcribers [114]. Romani et al. [115] compared the phonemes of Italian using three metrics: the rank-ordered age of acquisition, the frequency of articulation errors in a population of 11 patients with apraxia of speech, and the frequency of occurrence of the phoneme cross-linguistically. They found strong correlations, suggesting that, despite language-dependent variation, there are universal tendencies for some phones to be easier to learn and produce than others despite language-dependent variation in the list of which phones are easier to learn and produce than others. We therefore investigated whether or not phones that are easier for humans to learn and produce are also easier for a cross-linguistic automatic speech recognizer to correctly detect in the phone inventory of an unknown language.

In our analyses we focused on those phone tokens (IPA symbols) that are the most and least accurately recognized symbols by the hybrid crosslingual ASR system with a uni-gram phone LM. Table 8 shows the top 20 phone tokens that are successfully discovered (i.e., the true positives). Table 9 shows the top 20 phone tokens that are falsely discovered (i.e., the false positives). Table 10 shows the top 20 phone tokens that are missed (i.e., the false negatives). In these tables, the columns labeled "Count" show the number of our 13 languages that contain this phone symbol in their phoneme inventory. The columns labeled TP, FP, and FN show the number of languages in which the IPA symbol was correctly detected (true positives; TP), incorrectly detected (false positives; FP), or incorrectly not discovered (false negatives; FN) from the inventory by a cross-language speech recognizer, respectively. The columns labeled "Prec," "Rec," and "F1" show the token-level preci-

Symbol	TP	FP	FN	Count	Prec. [%]	Recall [%]	F1 [%]	Phoible [%]	AoS [%]	AoA Rank
m	13	0	0	13	100.0	100.0	100.0	96	3.6	2
i	13	0	0	13	100.0	100.0	100.0	92	-	-
j	13	0	0	13	100.0	100.0	100.0	90	7.0	4
k	13	0	0	13	100.0	100.0	100.0	90	5.6	1
u	13	0	0	13	100.0	100.0	100.0	88	-	-
p	13	0	0	13	100.0	100.0	100.0	86	4.5	2
n	13	0	0	13	100.0	100.0	100.0	78	3.7	2
l	13	0	0	13	100.0	100.0	100.0	68	6.6	1
t	13	0	0	13	100.0	100.0	100.0	68	2.2	1
s	13	0	0	13	100.0	100.0	100.0	67	5.8	2
a	12	1	0	12	92.3	100.0	96.0	86	-	-
o	12	1	0	12	92.3	100.0	96.0	60	-	-
b	11	2	0	11	84.6	100.0	91.7	63	15.8	3
d	11	2	0	11	84.6	100.0	91.7	46	16.4	3
h	8	0	2	10	100.0	80.0	88.9	56	-	-
e	10	2	1	11	83.3	90.9	87.0	61	-	-
w	10	2	1	11	83.3	90.9	87.0	82	5.9	4
z	7	0	3	10	100.0	70.0	82.4	30	8.4	6
r	8	5	0	8	61.5	100.0	76.2	44	9.3	3
ʃ	6	2	2	8	75.0	75.0	75.0	37	13.9	5

Table 8: Top 20 IPA symbols that are successfully discovered with the hybrid crosslingual ASR system. Phoible = % of 3183 phoneme inventories containing this symbol as a phoneme [32]. AoS = articulation errors produced by patients with apraxia of speech, % of all errors [115]. AoA = age of acquisition rank [115].

Symbol	TP	FP	FN	Count	Prec. [%]	Recall [%]	F1 [%]	Phoible [%]
˥	3	8	2	5	27.3	60.0	37.5	3
ɑ	0	8	3	3	0.0	0.0	0.0	7
˧	3	8	2	5	27.3	60.0	37.5	10
˩	2	8	2	4	20.0	50.0	28.6	18
ɪ	3	7	1	4	30.0	75.0	42.9	15
ɨ	0	6	3	3	0.0	0.0	0.0	16
˩	3	6	2	5	33.3	60.0	42.9	2
ˈ	0	5	1	1	0.0	0.0	0.0	-
ʊ	2	5	0	2	28.6	100.0	44.4	14
ː	5	5	0	5	50.0	100.0	66.7	-
ɾ	8	5	0	8	61.5	100.0	76.2	44
ɛ	7	4	2	9	63.6	77.8	70.0	37
ʔ	3	4	2	5	42.9	60.0	50.0	37
ə	5	4	2	7	55.6	71.4	62.5	22
ŋ	6	4	3	9	60.0	66.7	63.2	63
ʰ	6	4	2	8	60.0	75.0	66.7	-
ɾ	0	3	1	1	0.0	0.0	0.0	26
c	1	3	3	4	25.0	25.0	25.0	14
ɔ	7	3	3	10	70.0	70.0	70.0	35
ɤ	0	2	2	2	0.0	0.0	0.0	3

Table 9: Top 20 IPA symbols that are falsely discovered with the hybrid crosslingual ASR system (**false positives**). Phoible = % of 3183 phoneme inventories containing this symbol as a phoneme [32].

Symbol	TP	FP	FN	Count	Prec. [%]	Recall [%]	F1 [%]	Phoible [%]	AoS [%]	AoA Rank
f	6	0	7	13	100.0	46.2	63.2	44	8.5	4
v	1	2	6	7	33.3	14.3	20.0	27	18.3	3
x	0	2	5	5	0.0	0.0	0.0	19	-	-
ɲ	4	1	4	8	80.0	50.0	61.5	42	23.8	5
~	0	0	3	3	N/A	0.0	0.0	-	-	-
ɨ	0	6	3	3	0.0	0.0	0.0	16	-	-
ŋ	6	4	3	9	60.0	66.7	63.2	63	-	-
ʒ	4	1	3	7	80.0	57.1	66.7	16	-	-
ɔ	7	3	3	10	70.0	70.0	70.0	35	-	-
c	1	3	3	4	25.0	25.0	25.0	14	-	-
ɕ	0	1	3	3	0.0	0.0	0.0	4	-	-
ɯ	0	0	3	3	N/A	0.0	0.0	6	-	-
y	0	0	3	3	N/A	0.0	0.0	6	-	-
z	7	0	3	10	100.0	70.0	82.4	30	-	-
ʈ	0	0	3	3	N/A	0.0	0.0	16	-	-
ɣ	0	0	3	3	N/A	0.0	0.0	14	-	-
ɑ	0	8	3	3	0.0	0.0	0.0	7	-	-
ɦ	0	1	2	2	0.0	0.0	0.0	4	-	-
ɡ	0	1	2	2	0.0	0.0	0.0	57	29.5	5
ɖ	0	0	2	2	N/A	0.0	0.0	9	-	-

Table 10: Top 20 IPA symbols that are missed with the hybrid crosslingual ASR system (**false negatives**). Phoible = % of 3183 phoneme inventories containing this symbol as a phoneme [32]. AoS = articulation errors produced by patients with apraxia of speech, % of all errors [115]. AoA = age of acquisition rank [115].

sion, recall, and F1 score of each phone, using cross-linguistic ASR. The three remaining columns show extrinsic measures of the universality of each phone. The column labeled "Phoible" shows, as a percentage, the fraction of the 3183 phoneme inventories at [32] that contain each phone symbol as a phoneme. The columns labeled AoS and AoA, both copied from [115], show the frequency of each phone's articulation errors in the speech of people with Apraxia of Speech (AoS), and the rank-ordered Age of Acquisition (AoA) of this phone by Italian children, respectively; these two columns are blank for the phones (including all phones but /r/ in Table 9) that do not appear in [115].

Before we go deeper into the findings, first some general observations [**MA**]. There are 10 symbols that have 100% phone token inventory discovery F1-score; 18 symbols that have $\geq$ 80% discovery F1, and 30 symbols with $\geq$ 60% discovery F1 (see also Table 8). For reference, there are 96 unique IPA symbols, of which 56 symbols have 0% F1. Of those, 36 symbols are present exclusively in a single language, making their discovery impossible with this method; 11 are present in 2-3 languages and never hypothesized

	Phoible	AoS	AoA
Prec	0.76**	-0.34	-0.16
Rec	0.67**	-0.52*	-0.52*
F1	0.83**	-0.47	-0.38

Table 11: Pearson's correlation of cross-language phone token recognition scores (Precision, Recall, and F1) with extrinsic measures of phone token universality. Phoible = % of phoneme inventories containing the symbol [32], AoS = articulation errors produced by patients with apraxia of speech, AoA = age of acquisition rank [115]. Two-sided t-test: *:$p < 0.05$, **:$p < 0.01$, with 39(Phoible) or 15(AoS,AoA) degrees of freedom.

(pure false negatives; see also Table 10); 8 are present in 2-3 languages and are hypothesized, but never correctly (mix of false negatives and false positives). Finally, there is one symbol that has 0% F1 and is present in more than 3 languages, i.e., /x/ in 5 languages, where it's a false negative (see also Table 10) and it is also recognized as a false positive in two languages.

Now looking more closely at the individual phone tokens [**RQ2**]. Table 8 shows the phone tokens with the highest TP rates. Not surprisingly, these include the three "simple" sounds /p, t, k/, which occur in all 13 languages. Also the two nasals /m, n/ and the two approximants /l, j/, which occur in all 13 languages have a precision of 100%. Overall, particularly, plosives, nasals, and alveolar fricatives (/s, z/) have a very high precision. This suggests that the pronunciation of these sounds, or in fact, the pronunciation of these IPA symbols, is similar across the 13 languages in our dataset.

Table 10 shows the phone tokens with the highest FN rates. Interestingly, fricatives are dominant among the phone tokens with high FN rates (/f,z,v,x,ɣ,ʒ/); this point is also visible in Table 12, which shows that non-sibilant fricatives and labiodentals have lower recall than the phones of any other articulatory feature category. Across the rows of Tables 8 and 10, the degree to which ASR performance correlates with extrinsic measures of phoneme universality is striking. Table 11 shows the Pearson's-R correlations of cross-linguistic ASR Precision, Recall, and F1 score, respectively, with the frequency of each phone token in the languages of the world (column "Phoible"), the frequency of phone token error in patients with Apraxia of Speech (column "AoS"), and the phone token's Age of Acquisition among Italian children (column "AoA"). All three ASR measures are correlated with cross-linguistic phone frequency at $p < 0.01$ (two-sided t-test, 39 d.o.f.). Cross-linguistic phone token recall is also correlated with articulation errors

		Languages	Count	Prec. [%]	Recall [%]	F1 [%]
backness	front	1 - 13	55	85.7	76.4	80.8
	back	2 - 13	56	76.3	80.4	78.3
	central	1 - 7	20	40.0	50.0	44.4
	near-back	2 - 2	2	28.6	100.0	44.4
	near-front	4 - 4	4	30.0	75.0	42.9
height	close-mid	1 - 12	26	81.5	84.6	83.0
	close	3 - 13	51	86.7	76.5	81.3
	open	3 - 12	15	57.1	80.0	66.7
	open-mid	1 - 10	22	63.6	63.6	63.6
	mid	1 - 7	15	55.6	66.7	60.6
	near-close	2 - 4	6	29.4	83.3	43.5
	near-open	1 - 1	2	0.0	0.0	0.0
manner	lateral-approximant	1 - 13	14	100.0	92.9	96.3
	approximant	1 - 13	28	92.0	82.1	86.8
	nasal	1 - 13	44	87.8	81.8	84.7
	plosive	1 - 13	87	83.3	80.5	81.9
	sibilant-fricative	2 - 13	43	88.2	69.8	77.9
	trill	8 - 8	8	61.5	100.0	76.2
	non-sibilant-fricative	1 - 13	45	75.0	33.3	46.2
	click	1 - 1	2	0.0	0.0	0.0
	flap	1 - 1	2	0.0	0.0	0.0
	implosive	1 - 2	4	0.0	0.0	0.0
	lateral-click	1 - 1	1	0.0	0.0	0.0
	lateral-fricative	1 - 1	2	0.0	0.0	0.0
place	bilabial	1 - 13	40	90.2	92.5	91.4
	alveolar	1 - 13	88	88.6	88.6	88.6
	labio-velar	11 - 11	11	83.3	90.9	87.0
	palatal	1 - 13	28	81.8	64.3	72.0
	palato-alveolar	7 - 8	15	76.9	66.7	71.4
	glottal	2 - 10	17	68.8	64.7	66.7
	velar	1 - 13	40	72.7	60.0	65.8
	labio-dental	1 - 13	21	77.8	33.3	46.7
	alveolo-palatal	3 - 3	3	0.0	0.0	0.0
	dental	1 - 1	2	0.0	0.0	0.0
	labio-palatal	2 - 2	2	0.0	0.0	0.0
	pharyngeal	1 - 1	1	0.0	0.0	0.0
	retroflex	1 - 3	9	0.0	0.0	0.0
	uvular	1 - 2	3	0.0	0.0	0.0
roundness	rounded	1 - 13	59	83.9	79.7	81.7
	unrounded	1 - 13	78	58.5	70.5	64.0
voicing	voiceless	1 - 13	111	86.4	68.5	76.4
	voiced	1 - 13	169	81.0	70.4	75.3

Table 12: A break-down of phone token inventory discovery performance by different articulatory features. The "Languages" column indicates the minimum and maximum number of languages that share any symbol in a given category – e.g., in the back vowels category, every sound exists in at least 2 languages and at most 13 (all) languages. The "Count" column shows how many symbols there were across all languages.

by patients with Apraxia of Speech, and with Age of Acquisition ($p < 0.05$, two-sided t-test, 15 d.o.f.).

The symbols that are most frequently falsely assigned to inventories (i.e., the highest FP rates) are shown in Table 9. Almost half of the symbols in this table are suprasegmentals (tone symbols, the primary stress symbol /ˈ/, and the gemination symbol /ː/). The frequency of suprasegmentals in this table suggests that a crosslingual ASR system might not be the right tool to automatically assess the prosodic or lexical tone attributes of a language. The remaining symbols in this table are segments that commonly appear as allophones of phonemes written using other IPA symbols. For example, the vowels /a/ and /i/, two of the most common vowels in the languages of the world [32], may be implemented in many languages by their allophones /ɑ,ə/ and /ɨ,ɪ/ without loss of intelligibility; similarly /k/ may be implemented as /c/ or even /cʰ/ (as in the English words "skit" and "kit," respectively).

Table 12 analyzes performance in the discovery of base phones, as a function of the articulatory features of the base phone. Articulatory features are assigned to each symbol according to the IPA standard [7]: vowel symbols are characterized by backness, height, and rounding, consonant symbols by manner, place, and voicing. For each category, the Count column is the sum, across the inventories of all thirteen languages, of the number of symbols of that category. The Languages column specifies the minimum and maximum number of languages possessing any of the symbols in that category, e.g., there are three distinct symbols in the [implosive] manner category, one of which occurs in two languages, and two of which occur in just one language each, for a total Count of 4. Recall is the number of true positive detections of symbols in that category, summed across all languages, divided by the reference number of symbols in that category, also summed across languages. Precision is the number of true positives divided by the number of detected symbols; precision is set to zero if there are no detections. F1 is the harmonic mean of recall and precision.

Five manner categories, six place categories, and one vowel height category are never detected. All of these twelve categories are infrequent, both in the number of distinct symbols, and in the number of languages in which they occur: clicks, lateral clicks, and lateral fricatives occur only in Zulu. Other than these twelve never-detected categories, frequency of a category is not a certain pre-requisite for high accuracy. For example, rounded vowels are better-detected than unrounded vowels, despite being less frequent. There is some tendency for categories with low articulatory complexity to

be better-detected than those with high articulatory complexity, e.g., the best-detected manner of articulation (lateral-approximant) is exemplified in Table 8 by the phone /l/, which has a very early age of acquisition, while the poorly-detected non-sibilant fricative category is exemplified in Table 10 by the phones /f,v/ that have higher ages of acquisition. Many of the smaller-scale performance disparities are surprising, and perhaps worthy of further study. For example, close and close-mid vowels are detected with significantly better F1 scores than are open and open-mid vowels; we know of no theory of phonological organization that would explain this finding.

6. Conclusion

To the best of our knowledge, the presented study is the first to address the problem of automatic phonetic inventory discovery using IPA symbols with an application of ASR systems. It concludes a series of works that began with the exploratory analysis of the universality of phonetic representations in E2E ASR [21], and then moved onto studying the effect of phonotactic models strength on the performance of crosslingual (zero-shot) hybrid ASR [22]. These works have laid the ground for addressing our ultimate goal – i.e., phonetic inventory discovery.

We compared the performance of two E2E ASR systems which used phones and phone tokens as the output labels [**RQ1**] in Section 5.1. We found that the phone system performs better in multi- and monolingual conditions, whereas the phone token system is preferable for crosslingual recognition. We did not find a strong effect of either label type on the F1 score performance of phone token inventory discovery. However, we found that it is significantly easier to discover phone tokens than to discover phones.

We wanted to find out empirically which phones would be easier or more difficult to automatically discover in a new language [**RQ2**]. To that end, we first analysed the confusions of the crosslingual phone E2E ASR systems using Levenshtein alignments, confusion matrices, hierarchical clustering and a 2-D UMAP projection in Section 5.2. We presented clusters of IPA symbols that are very likely to be confused with each other by a crosslingual system, and found that these clusters have a clear phonological interpretation. Furthermore, we looked closely into which phones tend to be frequently discovered, falsely hypothesized, or undiscovered in Section 5.3. Our findings correlate well with extrinsic indicators such as phoneme frequency in Phoible [32] or the age of acquisition [115].

We previously claimed that the inability to model target language phonotactics inhibits crosslingual ASR performance [22], and here we extended this claim to also involve phonetic inventory discovery with crosslingual ASR in Section 5.3 [**RQ3**]. In particular, we found that a hybrid ASR system with a crosslingual AM (i.e., one that has never seen the target language) could exhibit much stronger phonetic inventory discovery performance if an oracle phonotactic LM is available. Ideally, an acoustic model alone would be able to recognize the sounds with no phonotactic information, although that seems unlikely given the strong reliance of ASR systems (whether hybrid or E2E) on the language model.

Is it possible to discover phonetic inventories with crosslingual ASR [**MA**]? We show that to some extent, the answer is "yes." Our best zero-shot system has an F1 score of 68.6% for phone token inventory discovery, but only 33.4% for phone inventory discovery, which presents clear continuing challenges. Most of the true positive detections are sounds that are common among many languages. However, a widespread presence is not sufficient, as common fricatives such as /f/, /v/, or /z/ tend to be harder to discover. Some sounds tend to be confused with others that differ by just a single articulatory feature, consistently with our past findings [21]. We also observe that the crosslingual ASR struggles with determining the lexical tonality and the prosody of phones. Rare sounds, which are arguably the most interesting to discover, and languages with large inventories (especially tone languages) remain as the largest challenge for the task of automatic phonetic inventory discovery, at least for our, arguably simple, proposed method (if they can be automatically discovered at all). We believe that these challenges are adequate future work candidates.

An interesting future direction is the assessment of discovered inventory quality by training and testing an ASR system using the discovered inventory. This would help assess better how different error types (substitutions, insertions, and deletions) affect the downstream task performance. We would also like to see what patterns are observed with a larger diversity of languages – to what extent could it reduce the number of unique phones, making their discovery possible?

We publicly share the code to reproduce our Kaldi[8] and ESPnet[9] ASR

[8] `https://github.com/pzelasko/kaldi/tree/discophone/egs/discophone`
[9] `https://github.com/pzelasko/espnet/tree/discophone/egs/discophone/asr1`

experiments.

References

[1] Q. Zhang, H. Lu, H. Sak, A. Tripathi, E. McDermott, S. Koo, S. Kumar, Transformer transducer: A streamable speech recognition model with transformer encoders and RNN-T loss, in: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 7829–7833.

[2] P. K. Austin, J. Sallabank, The Cambridge handbook of endangered languages, Cambridge University Press, 2011.

[3] O. Scharenborg, L. Ondel, S. Palaskar, P. Arthur, F. Ciannella, M. Du, E. Larsen, D. Merkx, R. Riad, L. Wang, E. Dupoux, L. Besacier, A. W. Black, M. Hasegawa-Johnson, F. Metze, G. Neubig, S. Stüker, P. Godard, M. Müller, Speech technology for unwritten languages, IEEE/ACM Transactions on Audio, Speech and Language Processing 28 (2020) 964–975.

[4] E. Morrell, J. Rowsell (Eds.), Stories from Inequity to Justice in Literacy Education: Confronting Digital Divides, Routledge, New York, 2020.

[5] L. F. Lamel, R. H. Kassel, S. Seneff, Speech database development: Design and analysis of the acoustic-phonetic corpus, in: Speech Input/Output Assessment and Speech Databases, ISCA, Noordwijkerhout, the Netherlands, 1989, pp. 2161–2170.

[6] H. C. Leung, V. W. Zue, A procedure for automatic alignment of phonetic transcriptions with continuous speech, in: ICASSP, IEEE, 1984, pp. 2.7.1–2.7.4.

[7] I. P. Association, I. P. A. Staff, et al., Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet, Cambridge University Press, 1999.

[8] T. Schultz, A. Waibel, Multilingual and crosslingual speech recognition, in: Proc. DARPA Workshop on Broadcast News Transcription and Understanding, 1998, pp. 259–262.

[9] E. Barnard, M. Davel, C. van Heerden, ASR corpus design for resource-scarce languages, in: Proc. Interspeech, 2009, pp. 2847–2850.

[10] J. Lööf, C. Gollan, H. Ney, Cross-language bootstrapping for unsupervised acoustic model training: Rapid development of a polish speech recognition system, in: Tenth Annual Conference of the International Speech Communication Association, 2009, pp. 88–91.

[11] P. Swietojanski, A. Ghoshal, S. Renals, Unsupervised cross-lingual knowledge transfer in dnn-based lvcsr, in: Proc. IEEE Spoken Language Technology Workshop (SLT), 2012, pp. 246–251.

[12] J.-T. Huang, J. Li, D. Yu, L. Deng, Y. Gong, Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers, in: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013, pp. 7304–7308.

[13] X. Li, S. Dalmia, J. Li, M. Lee, P. Littell, J. Yao, A. Anastasopoulos, D. R. Mortensen, G. Neubig, A. W. Black, F. Metze, Universal phone recognition with a multilingual allophone system, in: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 8249–8253.

[14] K. M. Knill, M. J. Gales, S. P. Rath, P. C. Woodland, C. Zhang, S.-X. Zhang, Investigation of multilingual deep neural networks for spoken term detection, in: Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), IEEE, 2013, pp. 138–143.

[15] P. Swietojanski, A. Ghoshal, S. Renals, Unsupervised cross-lingual knowledge transfer in DNN-based LVCSR, in: Proc. IEEE Spoken Language Technology Workshop (SLT), 2012, pp. 246–251.

[16] S. Dalmia, R. Sanabria, F. Metze, A. W. Black, Sequence-based multilingual low resource speech recognition, in: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2018, pp. 4909–4913.

[17] M. Swadesh, The phomemic principle, Language 10 (2) (2934) 117–129.

[18] K.-F. Lee, H.-W. Hon, Speaker-independent phone recognition using hidden Markov models, Transactions on Acoustics, Speech, and Signal Processing 37 (11) (1989) 1641–1648.

[19] C.-H. Lin, L.-S. Lee, P.-Y. Ting, A new framework for recognition of mandarin syllables with tones using sub-syllabic units, in: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vol. II, 1993, pp. 227–230.

[20] T. Nagamine, M. L. Seltzer, N. Mesgarani, Exploring how deep neural networks form phonemic categories, in: Proc. Interspeech, 2015, pp. 1912–1916.

[21] P. Żelasko, L. Moro-Velázquez, M. Hasegawa-Johnson, O. Scharenborg, N. Dehak, That sounds familiar: an analysis of phonetic representations transfer across languages, Proc. Interspeech (2020).

[22] S. Feng, P. Żelasko, L. Moro-Velázquez, A. Abavisani, M. Hasegawa-Johnson, O. Scharenborg, N. Dehak, How phonotactics affect multilingual and zero-shot ASR performance, in: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2021, pp. 7238–7242.

[23] N. S. Trubetzkoy, Zur allgemeinen theorie der phonologischen vokalsysteme, Travaux du cercle linguistique de Prague 1 (1929) 39–67.

[24] E. Sapir, Sound patterns in language, Language 1 (2) (1925) 37–51.

[25] G. L. Trager, B. Bloch, The syllabic phonemes of English, Language 17 (3) (1941) 223–246.

[26] G. L. Trager, The phonemes of Russian, Language 10 (4) (1934) 334–344.

[27] H. Osborn, Amahuaca phonemes, International Journal of American Linguistics 14 (1948) 188–190.

[28] O. A. Shell, Cashibo I: Phonemes, International Journal of American Linguistics 16 (4) (1950) 198–202.

[29] R. S. Larsen, E. V. Pike, Huasteco intonations and phonemes, Language 25 (1949) 268–277.

[30] H. P. Aschmann, Totonaco phonemes, International Journal of American Linguistics 12 (1) (1946) 34–43.

[31] W. L. Wonderly, Zoque II: Phonemes and morphophonemes, International Journal of American Linguistics 17 (2) (1951) 105–123.

[32] Phoible 2.0, (Available online at http://phoible.org, Accessed on 2020-05-04.) (2019).

[33] M. Versteegh, R. Thiollière, T. Schatz, X.-N. Cao, X. Anguera, A. Jansen, E. Dupoux, The zero resource speech challenge 2015., in: Proc. Interspeech, 2015, pp. 3169–3173.

[34] A. Tjandra, B. Sisman, M. Zhang, S. Sakti, H. Li, S. Nakamura, VQ-VAE unsupervised unit discovery and multi-scale code2spec inverter for ZeroSpeech Challenge 2019, in: Proc. Interspeech, 2019, pp. 1118–1122.

[35] T. Pellegrini, C. Manenti, J. Pinquier, Technical report the IRIT-UPS system@ ZeroSpeech 2017 track1: unsupervised subword modeling, Tech. rep., IRIT, Université de Toulouse (2017).

[36] M. Heck, S. Sakti, S. Nakamura, Feature optimized DPGMM clustering for unsupervised subword modeling: A contribution to ZeroSpeech 2017, in: Proc. IEEE workshop on automatic speech recognition and understanding (ASRU), 2017, pp. 740–746.

[37] H. Shibata, T. Kato, T. Shinozaki, S. Watanabe, Composite embedding systems for ZeroSpeech2017 track 1, in: Proc. IEEE workshop on automatic speech recognition and understanding (ASRU), 2017, pp. 747–753.

[38] T. K. Ansari, R. Kumar, S. Singh, S. Ganapathy, Deep learning methods for unsupervised acoustic modeling - LEAP submission to ZeroSpeech challenge 2017, in: Proc. IEEE workshop on automatic speech recognition and understanding (ASRU), 2017, pp. 754–761.

[39] T. K. Ansari, R. Kumar, S. Singh, S. Ganapathy, S. Devi, Unsupervised HMM posteriograms for language independent acoustic modeling in zero resource conditions, in: Proc. IEEE workshop on automatic speech recognition and understanding (ASRU), 2017, pp. 762–768.

[40] E. Hermann, S. Goldwater, Multilingual bottleneck features for subword modeling in zero-resource languages, in: Proc. Interspeech, 2018, pp. 2668–2672.

[41] E. Hermann, H. Kamper, S. Goldwater, Multilingual and unsupervised subword modeling for zero-resource languages, Computer Speech & Language 65 (2021) 101098.

[42] B. Wu, S. Sakti, J. Zhang, S. Nakamura, Optimizing DPGMM clustering in zero-resource setting based on functional load, in: Workshop on Spoken Language Technology for Under-resourced Languages (SLTU), 2018, pp. 1–5.

[43] D. Renshaw, H. Kamper, A. Jansen, S. Goldwater, A comparison of neural network methods for unsupervised representation learning on the zero resource speech challenge, in: Proc. Interspeech, 2015, pp. 3199–3203.

[44] M. Heck, S. Sakti, S. Nakamura, Supervised learning of acoustic models in a zero resource setting to improve DPGMM clustering, in: Proc. Interspeech, 2016, pp. 1310–1314.

[45] M. Heck, S. Sakti, S. Nakamura, Unsupervised linear discriminant analysis for supporting DPGMM clustering in the zero resource scenario, in: Workshop on Spoken Language Technology for Under-resourced Languages (SLTU), 2016, pp. 73–79.

[46] M. Heck, S. Sakti, S. Nakamura, Iterative training of a DPGMM-HMM acoustic unit recognizer in a zero resource scenario, in: Proc. IEEE Spoken Language Technology Workshop (SLT), 2016, pp. 57–63.

[47] T. Tsuchiya, N. Tawara, T. Ogawa, T. Kobayashi, Speaker invariant feature extraction for zero-resource languages with adversarial learning, in: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 2381–2385.

[48] S. Bhati, S. Nayak, K. S. R. Murty, Unsupervised speech signal to symbol transformation for zero resource speech applications, in: Proc. INTERSPEECH, 2017, pp. 2133–2137.

[49] R. Menon, H. Kamper, E. van der Westhuizen, J. Quinn, T. Niesler, Feature exploration for almost zero-resource ASR-free keyword spotting using a multilingual bottleneck extractor and correspondence autoencoders, in: Proc. Interspeech, 2019, pp. 3475–3479.

[50] K. Levin, A. Jansen, B. V. Durme, Segmental acoustic indexing for zero resource keyword search, in: Proc. IEEE workshop on automatic speech recognition and understanding (ASRU), 2015, pp. 5828–5832.

[51] M. Wiesner, O. Adams, D. Yarowsky, J. Trmal, S. Khudanpur, Zero-shot pronunciation lexicons for cross-language acoustic model transfer, in: Proc. IEEE workshop on automatic speech recognition and understanding (ASRU), 2019, pp. 1048–1054.

[52] M.-H. Siu, H. Gish, A. Chan, W. Belfield, Improved topic classification and keyword discovery using an HMM-based speech recognizer trained without supervision, in: Proc. Interspeech, 2010, pp. 2838–2841.

[53] H. Wang, A. Ragni, M. J. Gales, K. M. Knill, P. C. Woodland, C. Zhang, Joint decoding of tandem and hybrid systems for improved keyword spotting on low resource languages, in: Proc. Interspeech, 2015, pp. 3660–3664.

[54] M.-h. Siu, H. Gish, A. Chan, W. Belfield, S. Lowe, Unsupervised training of an HMM-based self-organizing unit recognizer with applications to topic classification and keyword discovery, Computer Speech & Language 28 (1) (2014) 210–223.

[55] C. Ni, L. Wang, C.-C. Leung, F. Rao, L. Lu, B. Ma, H. Li, Rapid update of multilingual deep neural network for low-resource keyword search, in: Proc. Interspeech, 2016, pp. 3698–3702.

[56] P. K. Muthukumar, A. W. Black, Automatic discovery of a phonetic inventory for unwritten languages for statistical speech synthesis, in: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2014, pp. 2594–2598.

[57] P. Baljekar, S. Sitaram, P. K. Muthukumar, A. W. Black, Using articulatory features and inferred phonological segments in zero resource speech processing, in: Proc. Interspeech, 2015, pp. 3194–3198.

[58] M. Müller, J. Franke, A. Waibel, S. Stüker, Towards phoneme inventory discovery for documentation of unwritten languages, in: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2017, pp. 5200–5204.

[59] E. Dunbar, R. Algayres, J. Karadayi, M. Bernard, J. Benjumea, X.-N. Cao, L. Miskic, C. Dugrain, L. Ondel, A. W. Black, L. Besacier, S. Sakti, E. Dupoux, The zero resource speech challenge 2019: TTS without T, in: Proc. Interspeech, 2019, pp. 1088–1092.

[60] Q. Yu, P. Liu, Z. Wu, S. K. Ang, H. Meng, L. Cai, Learning cross-lingual information with multilingual BLSTM for speech synthesis of low-resource languages, in: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 5545–5549.

[61] C.-y. Lee, T. J. O'donnell, J. Glass, Unsupervised lexicon discovery from acoustic input, Transactions of the Association for Computational Linguistics 3 (2015) 389–403.

[62] M. Bacchiani, M. Ostendorf, Joint lexicon, acoustic unit inventory and model design, Speech Communication 29 (2-4) (1999) 99–114.

[63] H. Kamper, A. Jansen, S. Goldwater, Unsupervised word segmentation and lexicon discovery using acoustic word embeddings, IEEE/ACM TASLP 24 (4) (2016) 669–679.

[64] A. Baevski, W.-N. Hsu, A. Conneau, M. Auli, Unsupervised speech recognition, downloaded 5/22/2021 from https://ai.facebook.com/research/publications/unsupervised-speech-recognition (2021).

[65] T. Kempton, R. K. Moore, Discovering the phoneme inventory of an unwritten language: A machine-assisted approach, Speech Communication 56 (2014) 152 – 166.

[66] C.-Y. Lee, J. Glass, A nonparametric Bayesian approach to acoustic model discovery, in: Proc. Annual Meeting of the Association for Computational Linguistics, 2012, pp. 40–49.

[67] L. Ondel, L. Burget, J. Černockỳ, Variational inference for acoustic unit discovery, Workshop on Spoken Language Technology for Under-resourced Languages (SLTU) 81 (2016) 80–86.

[68] L. Ondel, H. K. Vydana, L. Burget, J. Cernocký, Bayesian subspace hidden Markov model for acoustic unit discovery, in: Proc. Interspeech, 2019, pp. 261–265.

[69] J. Ebbers, J. Heymann, L. Drude, T. Glarner, R. Haeb-Umbach, B. Raj, Hidden markov model variational autoencoder for acoustic unit discovery, in: Proc. Interspeech, 2017, pp. 488–492.

[70] P. A. Estévez, M. Tesmer, C. A. Perez, J. M. Zurada, Normalized mutual information feature selection, IEEE Transactions on Neural Networks 20 (2) (2009) 189–201.

[71] B. Yusuf, L. Ondel, L. Burget, J. Černocký, M. Sara clar, A hierarchical subspace model for language-attuned acoustic unit discovery, in: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 3710–3714. doi:10.1109/ICASSP39728.2021.9414899.

[72] S. Feng, P. Żelasko, L. Moro-Velázquez, O. Scharenborg, Unsupervised acoustic unit discovery by leveraging a language-independent subword discriminative feature representation, in: Proc. Interspeech, 2021, pp. 1534–1538.

[73] T. Schultz, A. Waibel, Language-independent and language-adaptive acoustic modeling for speech recognition, Speech Communication 35 (1) (2001) 31–51, mIST. doi:https://doi.org/10.1016/S0167-6393(00)00094-7.
URL https://www.sciencedirect.com/science/article/pii/S0167639300000947

[74] T. Niesler, Language-dependent state clustering for multilingual acoustic modelling, Speech Communication 49 (6) (2007) 453–463. doi:https://doi.org/10.1016/j.specom.2007.04.001.
URL https://www.sciencedirect.com/science/article/pii/S0167639307000611

[75] A. Ghoshal, P. Swietojanski, S. Renals, Multilingual training of deep neural networks, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2013, pp. 7319–7323.

[76] S. Toshniwal, T. N. Sainath, R. J. Weiss, B. Li, P. Moreno, E. Weinstein, K. Rao, Multilingual speech recognition with a single end-to-end model, in: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 4904–4908. doi:10.1109/ICASSP.2018.8461972.

[77] A. Kannan, A. Datta, T. N. Sainath, E. Weinstein, B. Ramabhadran, Y. Wu, A. Bapna, Z. Chen, S. Lee, Large-scale multilingual speech recognition with a streaming end-to-end model, in: Proc. Interspeech, 2019, pp. 2130–2134.

[78] W. Byrne, P. Beyerlein, J. M. Huerta, S. Khudanpur, B. Marthi, J. Morgan, N. Peterek, J. Picone, D. Vergyri, T. Wang, Towards language independent acoustic modeling, in: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Vol. 2, IEEE, 2000, pp. II1029–II1032.

[79] A. Bruguier, D. Gnanapragasam, L. Johnson, K. Rao, F. Beaufays, Pronunciation learning with RNN-transducers., in: Proc. Interspeech, 2017, pp. 2556–2560.

[80] A. Patel, D. Li, E. Cho, P. Aleksic, Cross-lingual phoneme mapping for language robust contextual speech recognition, in: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2018, pp. 5924–5928.

[81] D. Yu, L. Deng, P. Liu, J. Wu, Y. Gong, A. Acero, Cross-lingual speech recognition under runtime resource constraints, in: 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, 2009, pp. 4193–4196.

[82] S. Tong, P. N. Garner, H. Bourlard, Multilingual training and cross-lingual adaptation on CTC-based acoustic model, arXiv preprint arXiv:1711.10025 (2017).

[83] S. Tong, P. N. Garner, H. Bourlard, Cross-lingual adaptation of a CTC-based multilingual acoustic model, Speech Communication 104 (2018) 39–46.

[84] H. Inaguma, J. Cho, M. K. Baskar, T. Kawahara, S. Watanabe, Transfer learning of language-independent end-to-end ASR with language model fusion, in: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 6096–6100.

[85] Y. Hu, S. Settle, K. Livescu, Multilingual jointly trained acoustic and written word embeddings, in: Interspeech 2020, 2020, pp. 1052–1056. doi:10.21437/Interspeech.2020-2828.

[86] Y. Hu, S. Settle, K. Livescu, Acoustic span embeddings for multilingual query-by-example search, in: 2021 IEEE Spoken Language Technology Workshop (SLT), 2021, pp. 935–942. doi:10.1109/SLT48900.2021.9383545.

[87] C. Jacobs, Y. Matusevych, H. Kamper, Acoustic word embeddings for zero-resource languages using self-supervised contrastive learning and multilingual adaptation, in: 2021 IEEE Spoken Language Technology Workshop (SLT), 2021, pp. 919–926. doi:10.1109/SLT48900.2021.9383594.

[88] M. Mohri, F. Pereira, M. Riley, Weighted finite-state transducers in speech recognition, Computer Speech & Language 16 (1) (2002) 69–88.

[89] A. Graves, S. Fernández, F. Gomez, J. Schmidhuber, Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks, in: Proceedings of the 23rd international conference on Machine learning, 2006, pp. 369–376.

[90] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, Y. Bengio, Attention-based models for speech recognition, in: Advances in neural information processing systems, 2015, pp. 577–585.

[91] S. Watanabe, T. Hori, J. R. Hershey, Language independent end-to-end architecture for joint language identification and speech recognition, in: Proc. IEEE workshop on automatic speech recognition and understanding (ASRU), IEEE, 2017, pp. 265–271.

[92] T. Hori, S. Watanabe, J. R. Hershey, Joint CTC/attention decoding for end-to-end speech recognition, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017, pp. 518–529.

[93] M. Behrisch, B. Bach, N. Henry Riche, T. Schreck, J.-D. Fekete, Matrix reordering methods for table and network visualization, Computer Graphics Forum 35 (3) (2016) 693–716.

[94] L. McInnes, J. Healy, N. Saul, L. Großberger, UMAP: Uniform manifold approximation and projection, Journal of Open Source Software 3 (29) (2018).

[95] M. Hasegawa-Johnson, L. Rolston, C. Goudeseune, G.-A. Levow, K. Kirchhoff, Grapheme-to-phoneme transduction for cross-language ASR, in: International Conference on Statistical Language and Speech Processing, Springer, 2020, pp. 3–19.

[96] T. Schultz, GlobalPhone: a multilingual speech and text database developed at Karlsruhe University, in: Seventh International Conference on Spoken Language Processing, 2002, pp. 345–348.

[97] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, S. Khudanpur, A pitch extraction algorithm tuned for automatic speech recognition, in: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2014, pp. 2494–2498.

[98] S. Kim, T. Hori, S. Watanabe, Joint CTC-attention based end-to-end speech recognition using multi-task learning, in: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2017, pp. 4835–4839.

[99] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N.-E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen, et al., Espnet: End-to-end speech processing toolkit, Proc. Interspeech (2018) 2207–2211.

[100] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin, R. Yamamoto, X. Wang, et al., A comparative study on Transformer vs RNN in speech applications, in: Proc.

IEEE workshop on automatic speech recognition and understanding (ASRU), IEEE, 2019, pp. 449–456.

[101] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in neural information processing systems, 2017, pp. 5998–6008.

[102] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186.

[103] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, et al., The Kaldi speech recognition toolkit, tutorial offered in conjunction with IEEE ASRU 2011 (2011).
URL `https://www.danielpovey.com/files/2011_asru_kaldi.pdf`

[104] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, S. Khudanpur, Semi-orthogonal low-rank matrix factorization for deep neural networks, in: Proc. INTERSPEECH, 2018, pp. 3743–3747.

[105] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, S. Khudanpur, Purely sequence-trained neural networks for ASR based on lattice-free MMI, in: Proc. Interspeech, 2016, pp. 2751–2755.

[106] G. Saon, H. Soltau, D. Nahamoo, M. Picheny, Speaker adaptation of neural network acoustic models using i-vectors, in: Proc. IEEE workshop on automatic speech recognition and understanding (ASRU), 2013, pp. 55–59.

[107] A. Stolcke, SRILM – an extensible language modeling toolkit, in: Proc. ICSLP, 2002, pp. 901–904.

[108] J.-T. Huang, J. Li, D. Yu, L. Deng, Y. Gong, Cross language knowledge transfer using multilingual deep neural network with shared hidden layers, in: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013, pp. 7304–7308.

[109] K. Veselý, M. Karafiát, F. Grézl, M. Janda, E. Egorova, The language-independent bottleneck features, in: Proc. IEEE Spoken Language Technology Workshop (SLT), 2012, pp. 336–341.

[110] K. Kilgour, M. Heck, M. Müller, M. Sperber, S. Stüker, A. Waibel, The 2014 KIT IWSLT speech-to-text systems for English, German and Italian, in: Internat. Worksh. Spoken Language Translation (IWSLT), Lake Tahoe, 2014, pp. 73–79.

[111] C. Gussenhoven, H. Jacobs, Understanding Phonology, 3rd edition, Hodder Education, London, Great Britain, 2011.

[112] R. D. Kent, The biology of phonological development, in: Phonological development: Models, research, implications, York Press, Timonium, MD, 1992, pp. 65–90.

[113] H. Kim, K. Martin, M. Hasegawa-Johnson, A. Perlman, Frequency of consonant articulation errors in dysarthric speech, Clinical Linguistics & Phonetics 24 (10) (2010) 759–770.

[114] P. Jyothi, M. Hasegawa-Johnson, Acquiring speech transcriptions using mismatched crowdsourcing, in: Proc. Conference on Artificial Intelligence (AAAI), Vol. 29, 2015, pp. 1263–1269.

[115] C. Romani, C. Galuzzi, C. Guaraglia, J. Goslin, Comparing phoneme frequency, age of acquisition, and loss in aphasia: Implications for phonological universals, Cognitive Neuropsychology 34 (7-8) (2017) 449–471.

Acronyms

AM Acoustic Model. 9, 18, 19, 22, 24, 30, 31, 40

AoA Age of Acquisition. 35, 36

AoS Apraxia of Speech. 35, 36

ASR Automatic Speech Recognition. 1–10, 13–26, 28–30, 32–36, 38–40

AUC Area Under the Curve. 7

CTC Connectionist Temporal Classification. 10, 11, 17

d.o.f. Degree of Freedom. 36, 38

DNN Deep Neural Network. 5, 8

E2E End-to-End. 2, 5, 6, 10, 11, 17, 19–25, 27, 28, 30, 31, 39, 40

FN False Negative. 28, 31, 33, 36

FP False Positive. 28, 31, 33, 38

G2P Grapheme-to-Phone. 13–15

GMM Gaussian Mixture Model. 18

HMM Hidden Markov Model. 5, 9, 18

IPA International Phonetic Alphabet. 2, 4–6, 8, 11, 14, 15, 17, 18, 21, 23, 30, 33–36, 38, 39

JSD Jensen-Shannon divergence. 13, 25

LF-MMI Lattice-Free Maximum Mutual Information. 18

LM Language Model. 6, 9, 11, 18, 19, 21, 22, 24, 30–33, 40

MA Main Aim. 3, 6, 13, 16, 19, 28, 30, 31, 35, 40

MFCC Mel-Frequency Cepstral Coefficient. 18

NLP Natural Language Processing. 17

NMI Normalized Mutual Information. 7

P Phone. 17, 19, 20, 22, 24, 27, 31

PER Phone Error Rate. 3, 17, 20–24

PL Pronunciation Lexicon. 9

PT Phone Token. 17, 19–22

PTER Phone Token Error Rate. 17, 20, 22, 23

RQ1 Research Question 1. 4, 6, 9, 10, 17, 20, 22, 23, 28, 39

RQ2 Research Question 2. 5, 6, 11, 25, 26, 36, 39

RQ3 Research Question 3. 5, 6, 9, 11, 18, 20, 22, 30, 40

T-SNE T-distributed Stochastic Neighbor Embedding. 26

TDNNF Factorized Time-Delay Neural Network. 18

TP True Positive. 28, 31, 33, 36

UMAP Uniform Manifold Approximation and Projection. 11, 26, 27, 39

WER Word Error Rate. 3

Highlights

Discovering Phonetic Inventories with Crosslingual Automatic Speech Recognition

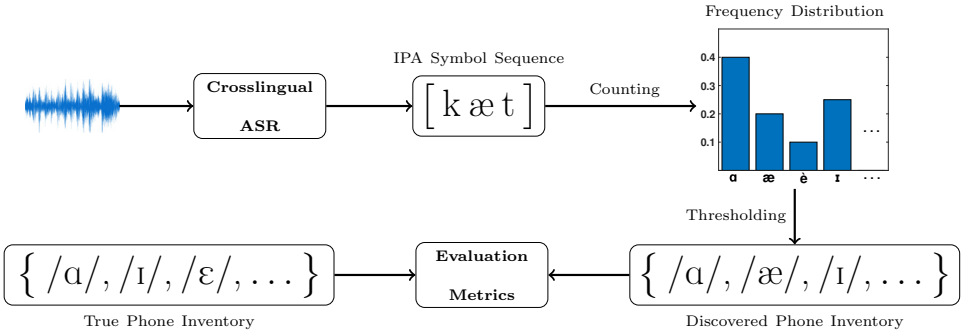
Piotr Żelasko, Siyuan Feng, Laureano Moro Velázquez, Ali Abavisani, Saurabhchand Bhati, Odette Scharenborg, Mark Hasegawa-Johnson, Najim Dehak

- We perform phonetic inventory discovery and present a way of approaching it with ASR.
- We train mono-, multi-, and crosslingual hybrid and E2E ASR on 13 diverse languages.
- Not knowing the target language phonotactics a priori is a limitation.
- We found universal phone tokens: IPA symbols that are well-recognized cross-linguistically.
- Unique sounds, similar sounds, and tone languages remain a challenge.

Graphical Abstract

Discovering Phonetic Inventories with Crosslingual Automatic Speech Recognition

Piotr Żelasko, Siyuan Feng, Laureano Moro Velázquez, Ali Abavisani, Saurabhchand Bhati, Odette Scharenborg, Mark Hasegawa-Johnson, Najim Dehak



Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☒The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

No conflict of interest.