

**MODELING OF LANGUAGE-UNIVERSAL SPEECH ATTRIBUTES FOR  
MULTILINGUAL SPEECH RECOGNITION AND PROCESSING**

A Dissertation  
Presented to  
The Academic Faculty

By

Hao Yen

In Partial Fulfillment  
of the Requirements for the Degree  
Doctor of Philosophy in the  
College of Engineering  
School of Electrical and Computer Engineering

Georgia Institute of Technology

December 2025

© Hao Yen 2025

**MODELING OF LANGUAGE-UNIVERSAL SPEECH ATTRIBUTES FOR  
MULTILINGUAL SPEECH RECOGNITION AND PROCESSING**

Thesis committee:

Prof. Chin-Hui Lee  
Electrical and Computer Engineering  
*Georgia Institute of Technology*

Prof. Elliot Moore  
Electrical and Computer Engineering  
*Georgia Institute of Technology*

Prof. David Anderson  
Electrical and Computer Engineering  
*Georgia Institute of Technology*

Prof. Sabato Marco Siniscalchi  
Computer Engineering  
*University of Palermo and Norwegian  
University of Science and Technology*

Prof. Larry Heck  
Electrical and Computer Engineering  
*Georgia Institute of Technology*

Date approved: Aug 5, 2025

## ACKNOWLEDGMENTS

I would never have been able to reach the destination of my PhD journey without the support of my advisors, committee members, mentors, friends, and family members.

First and foremost, I want to express my heartfelt thanks to my advisor, Professor Chin-Hui Lee. His constant support, sharp research insight, and thoughtful guidance have shaped both my academic path and personal growth in meaningful ways. I am especially grateful for his encouragement during times when I struggled to find direction, as his advice always helped me regain clarity and keep moving forward. I also truly appreciate his understanding and support while I pursued internship opportunities during the doctoral journey. Working with Prof. Lee has been one of the greatest privileges of my life. His passion, patience, and dedication continue to inspire me, and he is not just an incredible mentor but also a role model I genuinely look up to.

Second, I am also deeply thankful to my co-advisor, Professor Sabato Marco Siniscalchi, for his consistent support and valuable insight throughout my PhD. His perspective always brought clarity to my work, and his calm, encouraging presence made a lasting impact on how I approach both research and collaboration. Beyond the academic guidance, I've really appreciated the genuine connection we have built over the years. It has been a true pleasure to work with someone who brings both expertise and kindness into every interaction.

I am also grateful to committee members Prof. David Anderson, Prof. Elliot Moore II, and Prof. Larry Heck, for their invaluable contributions as members of my thesis reading committee, and for giving me lots of insightful suggestions and constructive feedback.

I would also like to express my sincere appreciation to the teams and mentors who supported me during my internship experiences. At Apple, I'm especially thankful to Woojay Jeon for his invaluable guidance. At Mitsubishi Electric Research Laboratories, I'm grateful to François G. Germain, Gordon Wichern, and Jonathan Le Roux for their mentorship

and collaboration. I also want to thank Shaoshi Ling and Guoli Ye at Microsoft for their thoughtful supervision and support. Their insights and encouragement made a meaningful impact on both my research and professional growth.

During my time at Georgia Tech, I have been lucky to be surrounded by an amazing group of friends and collaborators. I am especially thankful to Chao-Han Huck Yang, Hu Hu, Jun Qi, Pin-Jui Ku, Hsin-Hung Chen, Chun-Wei Ho, Yvonne Chen, and Sam Li. Their friendship, support, and collaboration have made this PhD journey so much more rewarding. From countless late-night discussions to overcoming challenges together, their presence has played a big role in both my personal and academic growth. A special shout-out to Pin-Jui Ku, who has always been a close friend and collaborator for nearly five years. Thank you for being someone I could always count on, both in research and in life.

Outside of academics, I am also incredibly grateful for the friends I made here in Georgia Tech, including Fan Lok Kong, Wil Ton Chuah, Rachel Ye, Maxx Ma, and many others. The time spent on and off the court, whether during late-night games, tournaments, casual hangouts, or just sharing laughs, has been a perfect balance to the intensity of research life. Their friendship brought so much joy, stress relief, and a sense of community that I will always cherish.

Last but certainly not least, I want to thank my parents for always being there for me. Their unwavering love, support, and belief in me have meant everything throughout this journey. I am incredibly lucky to have their encouragement as the foundation that helped me pursue my PhD dream abroad.

Finally, to my girlfriend, Hsiao-Hsuan Renee Wan. Her love, patience, and constant support have kept me grounded through every high and low. With her by my side, I feel truly ready for whatever comes next.

## TABLE OF CONTENTS

<b>Acknowledgments</b> . . . . .	iii
<b>List of Tables</b> . . . . .	x
<b>List of Figures</b> . . . . .	xiii
<b>Summary</b> . . . . .	xv
<b>Chapter 1: Introduction</b> . . . . .	1
1.1 Overview . . . . .	1
1.2 Motivations and Scientific Goals . . . . .	3
1.3 Main Contributions . . . . .	5
1.4 Thesis Outline . . . . .	7
<b>Chapter 2: Background and Literature Survey</b> . . . . .	9
2.1 Multilingual Modeling for Automatic Speech Recognition (ASR) . . . . .	9
2.2 Language-universal Acoustic Modeling . . . . .	10
2.2.1 Language-universal Acoustic Units . . . . .	10
2.2.2 Language-universal Feature Extraction . . . . .	13
2.3 Challenges and Issues for Language-universal Modeling . . . . .	14
2.3.1 Limitations of Language-universal Units . . . . .	14

2.3.2	Low-resource Languages . . . . .	16
2.3.3	Diversity of Writing systems and Textual Tokenization . . . . .	17
2.4	Articulatory Attribute Modeling for Speech Processing & Speech Recognition	18
2.5	Bottom-up ASR Framework . . . . .	20
<b>Chapter 3: Language-universal Speech Attribute Modeling . . . . .</b>		<b>22</b>
3.1	Universal Attribute Inventory . . . . .	22
3.2	Language-universal Speech Attribute Modeling . . . . .	25
3.3	Datasets and Experimental Setup . . . . .	26
3.4	Analysis and Discussions of Advantages of Speech Attributes . . . . .	28
3.4.1	Compact and Balanced Inventories of Speech Attributes . . . . .	28
3.4.2	Monolingual Speech Attribute Modeling . . . . .	31
3.4.3	Multilingual Speech Attribute Modeling . . . . .	34
3.4.4	Effects of Training Data Size . . . . .	35
3.5	Summary . . . . .	37
<b>Chapter 4: Boosting End-to-End Multilingual Phoneme Recognition through Exploiting knowledge-based Constraints . . . . .</b>		<b>38</b>
4.1	Proposed Language-universal Phoneme Recognition Framework . . . . .	38
4.1.1	Language-universal Speech Attribute Inventory . . . . .	38
4.1.2	Language-universal Attribute and Phoneme Recognizers . . . . .	39
4.1.3	Attribute-to-phoneme Mapping . . . . .	40
4.2	Datasets and Experimental Settings . . . . .	41
4.3	Experimental Results and Discussions . . . . .	43

4.4	Summary . . . . .	45
<b>Chapter 5: Language-Universal Speech Attributes Modeling for Zero-Shot Multilingual Spoken Keyword Recognition . . . . .</b>		<b>46</b>
5.1	Introduction . . . . .	46
5.2	Overview of The Language-universal Attribute Modeling SKR Framework .	47
5.2.1	ASR-based Spoken Keyword Recognition System . . . . .	47
5.2.2	DAT for Universal Spoken Keyword Recognition . . . . .	48
5.3	Datasets and Experimental Settings . . . . .	49
5.4	Experimental Results and Discussions . . . . .	51
5.4.1	DAT Results: Language Identification . . . . .	51
5.4.2	Analysis on Language-universal Latent Features . . . . .	52
5.4.3	Recognition Results . . . . .	53
5.4.4	Analysis of the Zero-shot Transfer Results . . . . .	55
5.4.5	Challenges for Speech Attribute Modeling for Continuous Speech Recognition . . . . .	57
5.5	Summary . . . . .	58
<b>Chapter 6: Keyword Information Extraction through Speech Attribute Modeling: from Isolated Word Recognition to Continuous Speech Recognition . . . . .</b>		<b>59</b>
6.1	Introduction . . . . .	59
6.2	A Two-stage System for KIE: Attribute-based ASR and Utterance Verification	60
6.2.1	Attribute-based ASR . . . . .	60
6.2.2	Utterance Verification . . . . .	61
6.3	Phoneme-based Vocabulary-Independent Utterance Verification (VIUV) . .	62

6.3.1	Subword-based Utterance Verification . . . . .	62
6.3.2	Attribute-specific Cohort Set . . . . .	65
6.3.3	String-Level Utterance Verification . . . . .	66
6.4	DNN-based Utterance Verification: Acoustically Grounded Word Embeddings . . . . .	67
6.5	Datasets and Experimental Settings . . . . .	69
6.5.1	Performance Measurement for Keyword Information Extraction . . . . .	69
6.5.2	Datasets and Keywords . . . . .	72
6.5.3	Models and Parameters . . . . .	73
6.6	Experimental Results and Discussions . . . . .	74
6.7	Summary . . . . .	76
<b>Chapter 7: A Bottom-up Framework for Continuous Speech Recognition in Syllable-based Languages . . . . .</b>		<b>78</b>
7.1	Syllable-based Languages and Speech Recognition . . . . .	78
7.2	A Bottom-up Framework for Syllable-based Speech Recognition . . . . .	79
7.2.1	Bottom-up Attribute Recognition and Knowledge Integration . . . . .	80
7.2.2	Mapping Syllables to Attribute Sequences . . . . .	80
7.3	Dataset and Experimental Settings . . . . .	83
7.3.1	Pronunciation and Syllable Clause Error Rate . . . . .	83
7.3.2	Dataset . . . . .	84
7.3.3	Models . . . . .	85
7.4	Experimental Results and Discussions . . . . .	85
7.4.1	Analysis on Attribute to Syllable Results . . . . .	85

7.4.2	Pronunciation Error Rate on AISHELL-1 . . . . .	87
7.4.3	Bottom-up Syllable-based Recognition Results on AISHELL-1 . . . . .	88
7.4.4	Ablation Study on Low-resource Training . . . . .	90
7.4.5	Cross-lingual Transfer to Japanese Syllable Recognition . . . . .	92
7.5	Summary . . . . .	94
<b>Chapter 8: Conclusions and Future Work . . . . .</b>		<b>96</b>
8.1	Summary . . . . .	96
8.2	Thesis Contributions . . . . .	97
8.3	Future Work . . . . .	99
<b>References . . . . .</b>		<b>100</b>
<b>Vita . . . . .</b>		<b>115</b>

## LIST OF TABLES

3.1	IPA vowel table organized by tongue height (rows) and backness (columns), illustrating the full inventory of vowel attributes used in this work. . . . .	24
3.2	Full inventory of language-universal speech attributes used in this dissertation.	25
3.3	Token error rate (%) for characters in the LibriSpeech test set, where the overall performance is 1.25%. . . . .	32
3.4	Token error rate (%) for phonemes in the LibriSpeech test set, where the overall performance is 1.86%. . . . .	32
3.5	Token error rate (%) for manner and place of articulation in the LibriSpeech test set, where the overall performance is 0.77% and 0.81% respectively. . .	33
3.6	Token error rate (%) for manner-place product features in the LibriSpeech test set, where the overall performance is 0.77%. . . . .	33
3.7	Token Error Rate (%) results eight languages in Multilingual LibriSpeech dataset with 10-hour subset for each language. Avg. stands for the average TER of the eight languages and std. is the standard deviation. . . . .	35
3.8	Token Error Rate (%) results eight languages in Multilingual LibriSpeech dataset with 1-hour subset for each language. Avg. stands for the average TER of the eight languages and std. is the standard deviation. . . . .	36
4.1	Inventory of universal speech attributes used for phoneme recognition. . . .	39
4.2	Training corpora and size in hours for each language. . . . .	41
4.3	Phoneme Error Rate (PER(%)) results for traditional models and proposed methods. Traditional methods include monolingual and multilingual models. Our proposed methods incorporated different categories of attributes, including <i>manner of articulation</i> , and <i>place of articulation</i> . The average PER (Avg) for six languages is also reported. . . . .	43

5.1	Training and testing data size in number of words and samples for each language, including in-domain in-vocabulary ( <b>ID-IV</b> ), in-domain out-of-vocabulary ( <b>ID-OOV</b> ), and unseen languages ( <b>UL</b> ). . . . .	50
5.2	Language identification accuracy (%) with (w/) and without (w/o) DAT for characters, phonemes and attributes. . . . .	51
5.3	Testing WER (%) of the in-domain set on 8 rich-resource languages and the average (Avg.). . . . .	54
5.4	Zero-shot transfer: Testing WER (%) of the 3 in-domain out-of-vocabulary ( <b>ID OOV</b> ) keywords from Russian, Italian, and Polish. . . . .	55
5.5	Zero-shot transfer: Testing WER (%) of the 3 unseen languages ( <b>UL</b> ), namely Turkish, Latvian, and Lithuanian. . . . .	55
5.6	Genetic proximity scores between unseen languages, Turkish (tr), Latvian (lv), and Lithuanian (lt), and the eight training languages. Avg. refers to the average score. . . . .	56
6.1	Examples of phoneme cohort sets based on speech attributes. . . . .	66
6.2	13 keywords (in lowercase) selected in this study. Keywords are divided into three categories according to the number of training instances in the dataset. . . . .	72
6.3	Testing MDR, FAR, and ATWV of the low-resource keywords. . . . .	75
6.4	Testing MDR, FAR, and ATWV of the unseen keywords. . . . .	76
7.1	Mandarin Chinsee initials/consonants classified by manner and place of articulation. The symbols in the square brackets are the closest IPA realizations of the corresponding initials in terms of pronunciation. . . . .	81
7.2	Japanese consonants classified by manner and place of articulation. The symbols in the square brackets are the closes IPA realizations of the corresponding initials in terms of pronunciation. . . . .	82
7.3	Examples of converting syllables to different attribute sequences for Mandarin and Japanese. . . . .	83

7.4	Statistics of mapping from different knowledge sources (Know. Source) to Mandarin (ZH) and Japanese (JP) syllables, including manner of articulation (M), place of articulation (P), voicing (V), height (H) and backness (B) of vowels, and aspiration (A). . . . .	86
7.5	PrER(%) on test set of AISHELL-1 for different categories of speech attributes, where $\prod = M \times P \times H \times B \times A$ is the Cartesian product of all the categories. No language model is applied during decoding. . . . .	88
7.6	Experimental results on the AISHELL-1 test set, reporting SER and SCER. For the proposed bottom-up system, different knowledge sources (attribute categories) are added incrementally. SCER is normalized with the full set of knowledge sources, ensuring maximum disambiguation of syllables. . . .	89
7.7	Two qualitative examples from the AISHELL-1 test set. . . . .	90
7.8	PrER(%) on subset of JSUT dataset ("basic5000") for different categories of speech attributes, from different source language ( $S$ ), including 150 hours of Mandarin Chinese (ZH) and 80 hours of Multilingual LibriSpeech (MLS), to target ( $T$ ) Japanese (JP). No language model is applied during decoding. . . . .	92
7.9	SER(%) on the subset of JSUT dataset ("basic5000") test set. All the WavLM systems are trained on 80 hours MLS dataset and tested on Japanese. * refers to systems use JSUT during training. For the proposed bottom-up system, different knowledge sources (attribute categories) are added incrementally. . . . .	94

## LIST OF FIGURES

2.1	Different token sequences of the sentence "she likes cats". . . . .	12
2.2	Words, phonemes (ARPABET), and IPA (slashes). Figure adapted from [95].	15
2.3	Distributions of writing systems in FLEURS (y-axis is the count). Figure adapted from [103]. . . . .	18
3.1	IPA chart of vowels. . . . .	24
3.2	Illustration of (a) separate recognizers for each category of speech attributes, and (b) joint modeling the product space of all categories of attributes. . . .	26
3.3	Unit inventory size vs. number of languages from the FLEURS dataset. . .	28
3.4	Token distribution of characters (in uppercase) in the LibriSpeech train, validation, and test sets. . . . .	29
3.5	Token distribution of phonemes in the LibriSpeech train, validation, and test sets. . . . .	30
3.6	Token distribution of manner of articulations in the LibriSpeech train, validation, and test sets. . . . .	30
3.7	Token distribution of manner-place product in the LibriSpeech train, validation, and test sets. . . . .	31

4.1	An overview of the traditional and proposed approaches. In (a), a multi-lingual approach predicting all the phoneme tokens for all languages. In (b), our proposed system contains two articulatory attribute recognizers for manner and place of articulation respectively, that are shared across languages. The attribute recognizers will produce attribute logits that will later be transformed to phoneme logits with an attribute-to-phoneme mapping. $S^C$ stands for the deterministic articulatory attribute-to-phoneme mapping matrix for a specific attribute category $C$ . . . . .	40
4.2	A qualitative example for utterance numbered 19557664 from the Swedish dataset. The above two plots are frame level posteriors of each phoneme corresponding to the ground truth, and the below two plots depict the activation curve for <i>nasal</i> and <i>bilabial</i> . . . . .	44
5.1	Overview of the language-universal spoken keyword recognition framework. "W2V2" refers to the Wav2Vec2.0 pre-trained model, and "PM" refers to pronunciation model. GRL refers to gradient reversal layer and Classifier denotes a language classifier. . . . .	48
5.2	PCA visualization of latent features from the final hidden layer of $DAT_{attr}$ , showing three manner-place articulation classes: nasal-bilabial (nas-blb), stop-alveolar (stp-alv), and fricative-labiodental (frc-lbd). Each cluster includes features from five languages. . . . .	52
5.3	Illustration of a language family tree for 8 in-domain training languages (English, German, French, Persian, Spanish, Russian, Italian, Polish) and 3 unseen languages (Turkish, Latvian, Lithuanian). . . . .	57
6.1	Overview of the attribute-based ASR system for KIE . . . . .	60
7.1	Overview of the bottom-up framework for syllable-based ASR. "AM" refers to acoustic model and "LM" stands for language model. . . . .	80
7.2	Experimental results (PrER, SER, and SCER) of using different amount of training data. The results for $WavLM_{syl}$ (syl) and $WavLM_{BU}$ (ours) are shown. . . . .	91

## SUMMARY

The performance of multilingual automatic speech recognition (ASR) systems critically depends on their ability to generalize across diverse languages and linguistic environments. Conventional ASR approaches typically involve training separate models for each language, which can be challenging due to the inherent variability in phonetic and linguistic structures across languages. Alternatively, some methods train a single model on data from multiple languages, but these models often struggle to effectively capture the unique characteristics of each language, particularly when there is a significant imbalance in training data. This disparity can lead to substantial performance degradations, especially in scenarios involving languages with limited or no training data. To address these issues in multilingual speech recognition, it is essential to develop approaches that operate uniformly across multiple languages without being constrained by language-specific characteristics. Current systems often fail to scale effectively due to their reliance on language-dependent tokens such as phonemes and characters, which are not universally applicable. This limitation poses significant challenges in building robust multilingual systems capable of accommodating a wide variety of languages and dialects.

In this dissertation, we aim to establish a language-universal framework for ASR that overcomes language-specific limitations by leveraging universal speech attributes, such as manner and place of articulation. These attributes, which remain consistent across all languages, serve as a foundation for building multilingual models capable of performing effectively across diverse linguistic settings. Our approach seeks to address the lack of knowledge sharing across languages due to linguistic distance, where traditional language-dependent tokens are insufficient. By utilizing a compact set of language-universal speech attributes, we aim to bridge the performance gap for low-resource and unseen languages, enhancing the adaptability and scalability of ASR systems.

As a first step in our proposed research, we build a language-universal framework by

modeling various speech attributes. By optimally utilizing advanced models and training techniques, the proposed framework effectively handles both seen and unseen languages with an improved accuracy. Our investigations demonstrate significant performance improvements in two preliminary multilingual speech recognition tasks: phoneme recognition and isolated spoken keyword recognition. Expanding upon these findings, we further extend our approach to continuous speech recognition and introduce a two-stage keyword information extraction system. This system consists of a first-stage attribute-based ASR model, which detects candidate keywords, followed by a second-stage phoneme-based utterance verification process, which refines detections by verifying keyword presence and boundaries, thereby reducing false alarms. The integration of this two-stage approach achieves superior performance over conventional character- and phoneme-based ASR systems. Finally, we propose a bottom-up ASR system as opposed to conventional top-down approaches to syllable-based continuous ASR, demonstrating the applicability of our proposed attribute modeling framework to general continuous speech recognition tasks. Moreover, we conduct comprehensive evaluations and analyses across all aforementioned recognition tasks to validate the effectiveness of the proposed framework.

# CHAPTER 1

## INTRODUCTION

### 1.1 Overview

End-to-end (E2E) automatic speech recognition (ASR) has made significant advancements in recent years [1]. However, one of the key research challenges yet to be addressed in ASR is building a language-universal engine for all spoken languages. One potential solution is to explore shared acoustic-phonetic structures among different languages to build a large set of acoustic models [2, 3, 4, 5, 6, 7] that characterize all the phone units needed to cover all the spoken languages being considered. Such approaches are called multilingual E2E ASR models, which have gained significant interest due to their capability to recognize multiple languages without prior knowledge of the language involved, e.g., [8, 9, 10].

However, there are several contrasting factors to be considered when designing multilingual E2E models [11]. On the one hand, expanding a multilingual model to accommodate a wide array of languages offers the benefits of more comprehensive contextual understanding and a greater range of recording conditions. On the other hand, this scaling introduces its own set of risks, such as data impurities stemming from variations between source and target languages, which could adversely affect the acoustic model of the target language. In addition, the presence of imbalanced data [12] from different languages can result in mingled acoustic characteristics and context mismatches, such as tonal variations from Mandarin mixing with non-tonal patterns from English. This can lead to significant degradation in models that are sensitive to context and are trained with diverse speech data from various language sources. Therefore, a primary challenge in multilingual acoustic modeling lies in striking a balance between leveraging the rich, diversified acoustic data from multiple languages and mitigating the negative effects of data impurity and language

mismatches. This has led to a growing interest in exploring knowledge sharing among multiple languages so as to define a universal set of acoustic-phonetic units that work for multiple or even for all languages.

The benefits associated with a language-universal ASR system are several: 1) unseen target languages can be decoded; 2) the number of model parameters can be reduced; and 3) training material can be shared. A common approach to achieving language-universal ASR involves defining a set of fundamental speech units that capture similar sounds across different languages. One method for this is to group acoustically and phonetically similar sounds under a single unified speech symbol, as proposed in [11]. A universal phonetic inventory can be constructed based on acoustic-phonetic knowledge, such as the International Phonetic Alphabet (IPA) [13], or through data-driven approaches that leverage large-scale speech corpora. By collecting diverse speech samples that encompass all possible speech units and their contextual variations, a comprehensive acoustic model can be trained to represent these units, thereby enabling ASR systems to function even for languages not explicitly present in the training data. These speech units are typically defined at the subword level, with phonemes being the most commonly chosen fundamental units. However, a major challenge with this approach is that as the number of languages increases, the number of required speech symbols also grows, leading to inconsistent parameter estimation and degradation in ASR performance [14]. Furthermore, since not all phonemes in a universal phone inventory are present in every individual language, it becomes necessary to group multiple languages together to construct a shared phonetic set that ensures broad coverage. Even in cases where a shared phoneme set can be defined for multiple languages, studies have reported significant performance degradation when transitioning from a language-dependent ASR setup to a language-independent model, as observed in [14, 15, 4, 2].

Speech attributes are unique features that explain the production of speech sounds by the mouth's articulators and are consistent across all languages. Unlike traditional graphemes or phonemes, speech attributes provide a lower-level articulatory representa-

tion that remains stable across linguistic variations. Opting for an attribute-based language characterization provides two main benefits over traditional high-level tokens. Firstly, these attributes are universally defined, eliminating the need to expand the token inventory or modify model architectures when new languages are introduced. This is particularly advantageous in multilingual ASR and zero-shot learning scenarios, where conventional systems struggle due to inconsistencies in unit inventories across languages. Additionally, speech attributes offer a dynamic and adaptive modeling approach, as they can be updated and refined as new linguistic data becomes available, regardless of the language. This adaptability is crucial for scaling ASR models to low-resource languages, where training data may be insufficient. Modeling speech attributes has been adopted in conventional statistical models [16, 17, 18, 19] and deep neural network architectures [20, 21] for different speech processing tasks, which demonstrate its effectiveness and robustness to acoustic variability.

## **1.2 Motivations and Scientific Goals**

Previous research on speech attribute modeling for speech recognition primarily centered around Automatic Speech Attribute Transcription (ASAT) [7, 22], which developed and applied attribute detectors for various speech processing tasks [16, 17, 18, 19]. However, these early approaches were predominantly built on Hidden Markov Model (HMM)-based systems [23, 24, 25], which suffer from several limitations compared to modern deep learning architectures. Due to the modular nature of HMM-based ASR, these systems required separate components for acoustic modeling, feature extraction, and decision-making, increasing complexity and reducing adaptability. Additionally, the accuracy of these HMM-based methods was significantly lower than contemporary neural network-based models, particularly in challenging multilingual or low-resource scenarios [26, 27]. More critically, these methods were not end-to-end (E2E), necessitating additional post-processing steps, such as rescore, and fine-tuning to integrate attribute predictions into downstream ASR tasks, thereby limiting scalability and efficiency.

Recently, phonetic knowledge and speech attributes have been introduced to the task of multilingual end-to-end ASR. However, most of these works focused on the recognition of characters, phonemes, or syllables [28, 29, 30, 31], which should be considered as an initial step to build a multilingual ASR. In addition, some of the systems [32, 33, 34] require a second stage of feature integration and additional fine-tuning, deviating from our primary goal of building an E2E system. The ambiguity of high-level units [31], such as phonemes and characters can also cause the modeling to be rather restrictive and oversimplified, resulting in inferior performance. Overall, the challenge of developing a fully end-to-end, attribute-based multilingual ASR system that overcomes the limitations of high-level linguistic units is yet to be solved. A comprehensive study on speech attribute modeling across different speech recognition tasks remains a critical area of focus.

In this dissertation, we present a comprehensive study on modeling speech attributes for multilingual ASR, beginning with phoneme recognition and progressively extending to continuous speech recognition. We construct a language-universal attribute inventory that maps a diverse range of phonemes into a compact set of articulatory attributes, enabling a more transferable and interpretable representation of speech sounds. Unlike prior approaches that incorporate attributes as auxiliary features or treat them as independent classification targets, we explore two primary modeling strategies: separate attribute recognizers and joint modeling within a product space. In the separate recognizer approach, each attribute is modeled as an independent prediction task, allowing the system to learn fine-grained articulatory patterns without being constrained by phoneme-level supervision. These attribute information can then be integrated with higher-level units, such as phonemes, to serve as knowledge-based constraints that improve recognition performance. The product space approach treats speech attributes as an interdependent, structured representation, capturing joint co-occurrence patterns and enforcing consistency across different articulatory dimensions. By modeling attributes in this integrated space, the system learns richer correlations between features such as manner, place, and voicing, which are often lost

in independently trained classifiers. These complementary modeling strategies allow for a detailed analysis of how various levels of attribute abstraction influence ASR performance and cross-lingual generalization. Moreover, the flexibility of the proposed framework enables the selection of task-specific attribute subsets, thereby optimizing performance across a range of applications, including phoneme recognition, isolated keyword recognition, keyword information extraction, and ultimately, continuous speech recognition.

By systematically exploring the role of speech attributes across various ASR tasks, this work seeks to address fundamental questions regarding their effectiveness in multilingual, low-resource, and zero-shot recognition settings. We evaluate how attribute-based models compare to traditional phoneme- and character-based systems, and demonstrate how they can be effectively integrated into end-to-end ASR architectures. Through this comprehensive investigation, we aim to establish articulatory attributes as a interpretable, scalable, and generalizable foundation for the next generation of multilingual speech recognition systems.

### **1.3 Main Contributions**

The goal of this dissertation is to advance multilingual speech recognition and processing techniques by modeling language-universal speech attributes. In particular, we focus on leveraging fundamental articulatory attribute information to improve speech recognition performance and capabilities in multilingual, low-resource, and zero-shot scenarios. The main contributions of this dissertation are summarized as follows:

1. We construct a language-universal inventory of fundamental speech attributes. This inventory covers articulatory features, such as manner and place of articulation and voicing, that are shared across a wide range of languages, based on linguistic knowledge grounded in the International Phonetic Alphabet (IPA). Attribute recognition models are trained on multilingual speech data to predict these features with high accuracy and generalizability.

2. We propose an initial approach toward multilingual end-to-end ASR by incorporating attribute predictions as auxiliary supervision and knowledge-based constraints. A deterministic attribute-to-phoneme mapping is introduced to project rich articulatory posteriors into phoneme-level outputs, thereby improving recognition accuracy while retaining interpretability.
3. We extend our attribute modeling framework to end-to-end spoken keyword recognition (SKR). The system combines a self-supervised pretrained encoder with a non-trainable pronunciation model and is further enhanced by domain adversarial training (DAT), which promotes language-invariant representation learning and improves zero-shot keyword recognition for unseen languages.
4. We develop a two-stage keyword information extraction (KIE) system that utilizes language-universal speech attributes in the first-stage ASR to provide high-recall keyword candidates. In the second stage, two utterance verification modules, including a statistical hypothesis testing process and a deep neural embeddings method, are employed to validate the keyword presence and boundaries, effectively reducing false alarms.
5. We introduce a bottom-up ASR framework for syllable-based continuous speech recognition. The proposed system first predicts sequences of articulatory attributes and then reconstructs syllables through a structured attribute-to-syllable mapping. We design evaluation metrics that assess both pronunciation quality and syllable-level accuracy, accounting for mapping ambiguities. Experimental results on Mandarin Chinese demonstrate the framework’s effectiveness under both full and low-resource conditions, with potential for extension to other syllable-based languages such as Japanese.

## 1.4 Thesis Outline

The rest of this thesis is organized as follows:

In chapter 2, we provide a review of relevant literature and fundamental background for this research, including key concepts in multilingual speech recognition and language-universal speech attributes. We also describe the datasets and tools used throughout the study to establish the experimental context.

In chapter 3, we construct a comprehensive language-universal speech attribute inventory and analyze its effectiveness compared to conventional modeling units such as characters and phonemes. Through experiments, we demonstrate that articulatory speech attributes offer a more compact, interpretable, and language-universal representation. This modeling choice enables better generalization across languages and plays a critical role in building scalable and robust multilingual ASR systems.

In chapter 4, we introduce our first practical application and contribution, focusing on language-universal speech attribute modeling for phoneme recognition in a multilingual setting. This chapter details how incorporating knowledge-based speech attribute constraints can improve phoneme recognition across multiple languages from diverse regions, laying the groundwork for subsequent chapters.

In chapter 5, we leverage the speech attribute modeling in an integrated product space to develop a spoken keyword recognition framework. The methods and findings in this chapter demonstrate the effectiveness of our attributes-based approach in isolated word recognition tasks and zero-shot transfer scenarios. Furthermore, we highlight potential challenges that arise when moving to continuous speech recognition scenarios, which are addressed in the next chapter.

In chapter 6, we move a step forward to utilize the language-universal speech attribute modeling techniques to tackle continuous speech recognition. This chapter introduces a two-stage system for keyword information extraction in continuous speech, effectively

bridging the gap between isolated word recognition and full continuous speech recognition. The proposed framework is built upon an attribute-based ASR model as the first stage, followed by two distinct utterance verification modules: one based on traditional hypothesis testing approach and the other leveraging deep neural network (DNN-based) techniques.

In chapter 7, we propose a bottom-up framework for continuous speech recognition in syllable-based languages. Building upon the language-universal speech attribute modeling developed in earlier chapters, this system first predicts sequences of articulatory attributes and then reconstructs syllables through a structured, interpretable mapping process. We introduce two evaluation metrics, Pronunciation Error Rate (PrER) and Syllable Clause Error Rate (SCER), to assess the model’s ability to capture underlying phonological structure and syllable-level accuracy, even in the presence of mapping ambiguities. Experimental results on Mandarin Chinese demonstrate the effectiveness of this approach under both full and low-resource conditions. While this chapter focuses on Mandarin, the proposed framework is designed with cross-lingual scalability in mind and is applicable to other syllable-based languages such as Japanese.

Finally, chapter 8 concludes the dissertation and discusses future research directions. In this concluding chapter, we summarize the key contributions of the work and reflect on how each chapter’s developments build toward a unified goal. We also outline potential directions for extending this research, highlighting how the insights from this thesis can inform and drive subsequent studies in multilingual speech recognition and processing.

## CHAPTER 2

### BACKGROUND AND LITERATURE SURVEY

#### 2.1 Multilingual Modeling for Automatic Speech Recognition (ASR)

In recent years, conventional multilingual ASR approaches can be broadly categorized into two main strategies. The first strategy involves merging the token sets (e.g., phonemes, characters, subwords, or even bytes) of all languages into a unified output token set and training an end-to-end (E2E) model on all available multilingual data [8, 35, 9, 10]. This method results in a universal ASR model capable of recognizing speech in any language included in its training data, leveraging shared representations across languages and simplifying the training process. Such E2E models benefit from extensive parameter sharing, which helps in learning generalized features applicable to multiple languages. This approach normally does not require any prior knowledge of the language involved.

The second approach focuses on sharing only certain parameters of the acoustic model across languages [36, 37]. In this method, the input and hidden layers of the model are shared among all languages, forming a universal front-end that captures common acoustic features, while the output layers remain language-specific. Each language has its own dedicated output layer to handle the language-dependent aspects of recognition. This strategy encourages the model to learn both shared and language-specific representations, making it particularly effective for multilingual ASR systems. Many studies have used this approach to optimize ASR models for low-resource languages by leveraging shared learning across languages, thereby enhancing generalization and performance [11, 38, 39, 40].

When designing multilingual E2E ASR models, several key factors must be taken into account. One major challenge is the imbalance in data availability across languages, especially when scaling to include a large number of languages [41]). Such imbalance often

results in the model performing well on high-resource languages while underperforming on low-resource ones. To address this, data sampling techniques are commonly employed to balance the amount of training data across languages [12, 41]. Additionally, the capacity of the model must be expanded to accommodate the complexity of recognizing multiple languages; as demonstrated in [42], this often involves scaling multilingual models to billions of parameters. However, a straightforward E2E architecture may not optimally handle the linguistic diversity present in multilingual data. To better adapt to language-specific variations, some approaches incorporate language-specific adaptation layers to enhance model performance on diverse languages [12]. Similarly, the mixture of experts architecture has been applied in multilingual models to further boost performance by dynamically selecting specialized subnetworks or "experts" based on the input language [43, 44]. These strategies allow for more flexible and efficient handling of multilingual ASR tasks by enhancing model's capability to better capture language-specific characteristics. Despite these advancements, building robust multilingual ASR systems remains a challenging task.

## **2.2 Language-universal Acoustic Modeling**

### 2.2.1 Language-universal Acoustic Units

In automatic speech recognition (ASR) systems, the selection of acoustic units plays a crucial role in determining the system performance, flexibility, and effectiveness across different languages and tasks. The choice of units in an ASR model directly affects its ability to capture language-specific details, manage out-of-vocabulary (OOV) words, and maintain computational efficiency. In modern ASR, three types of acoustic units, namely characters, phonemes, and subwords, are commonly adopted due to their distinct advantages and trade-offs.

Character-based models [45, 46, 47, 48, 49] operate at the level of individual letters or symbols in a written language. This approach is straightforward to implement because it does not require complex linguistic resources, such as phonetic dictionaries, to map sounds

to symbols. Character-based models are effective in handling OOV words, as any word can be represented by a sequence of its constituent characters from a small, fixed vocabulary. This makes them particularly suitable for languages with simple orthographies where there is a direct mapping between sounds and written forms.

Phoneme-based models [50, 51, 52, 25, 53] use phonemes, the smallest distinguishable sound units that differentiate words in a language, as their acoustic units. By representing speech as sequences of phonemes rather than characters or words, these models provide a more compact and efficient representation of speech, reducing the length of sequences needed for processing. Phonemes are relatively consistent across languages, making phoneme-based models well-suited for multilingual ASR and transfer learning across different languages.

Subword-based models [54, 55, 56, 57] provide a middle ground between the granularity of characters and the compactness of phonemes. Subwords are typically generated using data-driven techniques like byte-pair encoding (BPE) [58, 59] or unigram language modeling, which produce units such as frequently occurring prefixes, suffixes, and root morphemes. This approach allows subword-based systems to represent languages more flexibly, efficiently managing OOV words by breaking them down into smaller, known subword components. Subword models also maintain a compact vocabulary size while preserving important linguistic information, making them particularly effective for morphologically rich languages with numerous word variations. Their ability to capture meaningful language patterns without excessively long sequences makes subword units a popular choice for modern end-to-end ASR systems.

Despite the successes of character-, phoneme-, and subword-based ASR systems, these approaches are often considered language-specific and require substantial amounts of training data to perform well, which is a significant limitation for low-resource languages. To address these challenges, recent research has explored multilingual modeling using universal speech units that generalize more effectively across languages. One such approach

Token	Token sequence
Character	['s', 'h', 'e', 'l', 'i', 'k', 'e', 's', 'c', 'a', 't', 's']
Subword	['she', '_lik', 'es', '_c', 'ats']
Phoneme	['SH', 'IY1', 'L', 'AY1', 'K', 'S', 'K', 'AE1', 'T', 'S']
IPA	['ʃ', 'i:', 'l', 'aɪ', 'k', 's', 'k', 'æ', 't', 's']


Language-dependent  
  
 Language-universal

Figure 2.1: Different token sequences of the sentence "she likes cats".

leverages the International Phonetic Alphabet (IPA) [60], a standardized system that represents the sounds of spoken language across the world. Unlike language-specific phonetic alphabets, the IPA provides a universal set of symbols for phonemes, making it a powerful tool for multilingual ASR. By offering a consistent framework for phonetic transcription, IPA-based models improve cross-lingual transfer learning and zero-shot learning [61, 62, 63]. Incorporating IPA into ASR systems [64, 65] can lead to more robust models capable of handling pronunciation variations and dialectal differences, which is especially valuable in low-resource settings. The standardized representation provided by the IPA allows for leveraging phonetic similarities between languages, thereby enhancing the efficiency and accuracy of multilingual ASR systems. Figure 2.1 provides an example of how different selection of tokens represent the same sentence. As we move from characters to IPA, we transition from a more language-dependent unit to a language-universal representation.

Additionally, some studies have proposed using a universal narrow phone set, known as allophones [66], to build phone-based recognition models [67, 68, 69]. Allophones represent different pronunciations of a phoneme that do not alter the word's meaning. By employing a universal set of allophones, ASR systems can recognize these variations across multiple languages, improving performance in diverse linguistic contexts. This approach supports the development of more language-agnostic ASR models that are less dependent on large, language-specific datasets, making them highly suitable for low-resource scenar-

ios. Together, the use of universal speech units like the IPA and allophones pushes ASR technology towards more inclusive and scalable multilingual systems.

### 2.2.2 Language-universal Feature Extraction

Traditional language-universal feature extraction in speech processing relies on fundamental acoustic and prosodic characteristics shared across all spoken languages. These features include spectral features like Mel-Frequency Cepstral Coefficients (MFCCs) [70] and Perceptual Linear Prediction (PLP) [71], which represent the speech signal’s short-term power spectrum, mimicking the human auditory system’s response. Prosodic features [72, 73] such as pitch, energy, and duration capture intonation, stress, and rhythm, which are crucial for tasks like speaker recognition and speech synthesis. Temporal features like formant frequencies and transitions capture phonetic information across languages, while voice quality features such as jitter and shimmer [74] help distinguish speakers and detect emotions [75]. These handcrafted features, being language-universal, are widely used in various speech applications across multiple languages.

While these traditional features provide a solid foundation, recent advances in Deep Neural Networks (DNNs) have revolutionized feature extraction in speech processing by learning data-driven representations directly from raw speech data. Unlike handcrafted features, DNN-based feature extraction methods automatically learn complex patterns in speech data that capture both local and global dependencies [51]. These models produce abstract and robust features that generalize well across languages, enabling more effective multilingual ASR systems. Building upon the success of DNN, domain adversarial training (DAT) [76] is a powerful technique for improving the adaptability of DNN models, especially in minimizing discrepancies between the training and testing data. The idea is to learn features that are consistent across different domains with the help of adversarial training. Such domain-invariant features ensure that models can retain high performance even when deployed in environments different from their original training conditions. Re-

cently, DAT is extensively used in enhancing the language universality of ASR systems [77, 78, 79].

Recently, large pre-trained models have significantly advanced ASR by providing robust and generalizable representations that can be fine-tuned for various tasks and languages. Innovations like Wav2Vec [80] utilize contrastive learning to predict future representations based on past contexts. Its successor, VQ-Wav2Vec [81], builds on this architecture by quantizing representations into discrete tokens. Wav2Vec2.0 [82] expands this technique by using contrastive learning to predict the representations of masked sections, both past and future. HuBERT [83] enhances this approach by using k-means clustering to map speech signals to discrete labels, facilitating a BERT-style pre-training. BEST-RQ [84] learns a model to predict the masked speech signals, in the form of discrete labels generated with a random-projection quantizer. WavLM [85] further enhances speech representation learning by focusing simultaneously on masked speech prediction and speech denoising. In addition, multilingual models such as Whisper [86], Google USM [87], and SeamlessM4T [88] are trained on vast amounts of multilingual and multi-domain speech data, capturing rich acoustic, phonetic, and linguistic patterns. These models are highly effective in multilingual ASR as they learn language-agnostic features that are transferable across different languages, supporting cross-lingual transfer and zero-shot learning [89, 90]. By fine-tuning a single pre-trained ASR model on multiple languages, these models can perform well on both seen and unseen languages [91] without needing extensive additional training data [92, 93, 94]. This capability is particularly advantageous in low-resource settings and contributes to more scalable and efficient multilingual speech recognition systems.

## **2.3 Challenges and Issues for Language-universal Modeling**

### 2.3.1 Limitations of Language-universal Units

It is well established that grapheme-based units such as characters and subwords are inherently language-dependent. Characters are tied to a specific writing system, and sub-



eling. One key issue is the presence of language-specific symbols. Some phonemes that are distinct in one language may not have direct equivalents in another. For instance, English differentiates between [θ] and [ð] in *thin* and *this*, but many languages do not make this distinction, leading to potential inconsistencies in phonetic transcription and model training. Additionally, phonetically similar sounds may be represented using different IPA symbols depending on the language. For instance, the English /ʃ/ (as in *she*) and the Mandarin Chinese /ʃ/ (as in *XI* in Pinyin) are acoustically similar but transcribed differently in IPA. This inconsistency can introduce errors when training multilingual models, as similar sounds may be treated as entirely separate phonemes.

### 2.3.2 Low-resource Languages

A common requirement to build a high-performance ASR system is to prepare a large amount of labeled training data (speech utterances and corresponding transcriptions of the utterances). Such a requirement is not always realizable in real-world scenarios. In fact, according to Ethnologue [96], there are above 7000 languages spoken by people across the world but only a fraction have sufficient linguistic resources for speech recognition development. Furthermore, many languages are unwritten or lack standardized spelling [97], complicating the application of traditional grapheme-based ASR models. In particular, the majority of the languages spoken in Asia and Africa lack the training data that is required to build accurate state-of-the-art ASR systems. These languages are referred to as low-resource languages. This data scarcity issue presents a major bottleneck in building effective ASR systems, as speech corpora are crucial for training acoustic models, language models, and pronunciation lexicons. Such a scenario is often referred to as a low-resource training scenario [62, 98].

Recently, numerous algorithms have been developed to train ASR systems under low-resource scenarios. A promising direction is transfer learning [99, 100, 101], where ASR models trained on high-resource languages are adapted to low-resource languages with

minimal supervision. This approach leverages shared phonetic structures across languages, allowing multilingual models to generalize across different linguistic domains. Another category of approaches is to adopt a large pretrained model [86, 87, 88, 102] to extract representative features to facilitate efficient and effective training for the ASR systems. The pretrained model is generally trained on a large-scale dataset with either a supervised or self-supervised training manner.

However, as shown in [86], linguistic divergence and phonetic variations can still lead to performance degradation, especially for languages with unique phoneme inventories or writing systems that are not well represented in existing multilingual datasets.

### 2.3.3 Diversity of Writing systems and Textual Tokenization

Another major limitation lies in token modeling across diverse writing systems. This variation complicates the design of tokenization and output vocabularies, especially when the goal is to train a single universal ASR model across hundreds of languages.

In current multilingual ASR systems, a common strategy is to adopt a unified output vocabulary composed of characters or subwords across all supported languages. While this simplifies model architecture, it introduces significant challenges when dealing with diverse writing systems. For instance, Chinese text contains no spaces between words, making traditional word or subword tokenization methods ineffective without dedicated word segmentation tools. Similarly, scripts like Japanese and Korean often exhibit inconsistent spacing conventions, depending on the source or transcriber. These inconsistencies make it difficult to standardize token boundaries and hinder reproducibility across datasets.

Furthermore, there is a pronounced skew in the distribution of writing systems used in publicly available multilingual speech datasets. As illustrated in Figure 2.3, scripts such as Latin and Cyrillic dominate, while dozens of others, including Devanagari, Arabic, and various Brahmic or Indic scripts, are underrepresented. This imbalance leads to uneven model performance, as the unified vocabulary is implicitly biased toward the more preva-

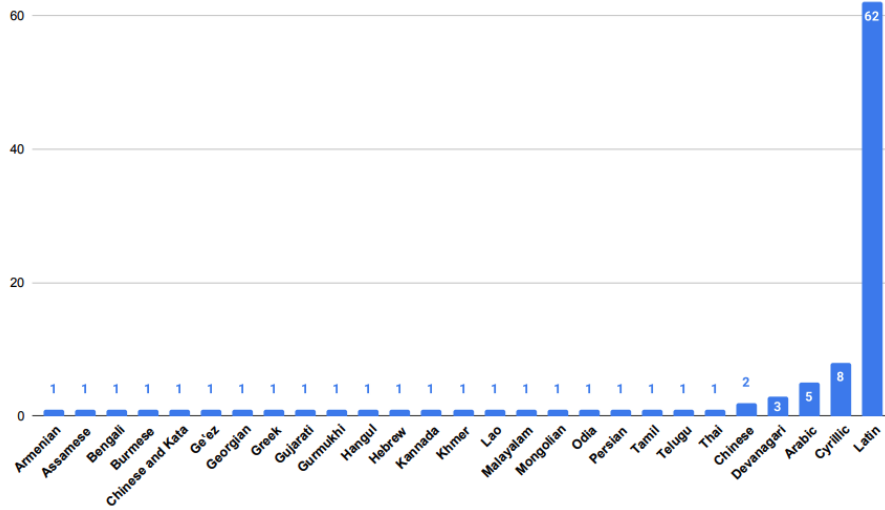


Figure 2.3: Distributions of writing systems in FLEURS (y-axis is the count). Figure adapted from [103].

lent scripts. In multilingual settings, languages using non-Latin scripts often suffer from higher error rates, not necessarily because they are linguistically more complex, but due to issues such as mismatched tokenization, limited script representation in the training data, or inconsistencies in how similar characters are handled across languages.

## 2.4 Articulatory Attribute Modeling for Speech Processing & Speech Recognition

Articulatory attributes [104, 105] are a set of distinct features that describe how speech sounds are produced by the articulators in the mouth, and these features are universally shared across all languages. Characterizing languages based on attributes offers two significant advantages over high-level characterization. Firstly, these attributes are universal across all languages, meaning that fundamentally there is no requirement to expand the number of attributes or their models when extending additional target languages. Furthermore, regardless of the languages, they can be refined once new training data becomes available. In previous studies on attribute modeling, automatic speech attribute transcription (ASAT) [7, 22] has been utilized with success as a key component for several attribute-based speech applications. In ASAT framework, various detectors are trained to generate

a bank of speech attributes. These detectors can be useful in many speech related task such as mispronunciation detection [106, 19, 107], spoken language recognition [17, 108], lattice rescoring [109, 110], and continuous speech recognition [16, 111, 112, 113, 114]. The inherent sharable properties of articulatory attributes across languages make them a promising choice for building a multilingual ASR system, and we thereby expect to extend the technique to an E2E multilingual scenario that can maximize knowledge-sharing and naturally allow universal recognition.

An attribute modeling technique has been introduced to the task of end-to-end (E2E) speech recognition. Müller et al. [33] combined language feature vectors with articulatory attributes in an effort to improve low-resource ASR. Seven articulatory attribute classifiers were defined, and feature combinations for ASR was performed. A slight word error rate reduction is observed by combining these classifiers within the ASR system for final recognition. In [32, 34], multiple articulatory feature extractors are trained with CTC [115] criterion for speech recognition. Li et al. [31] investigated an efficient method of encoding multilingual transcriptions with articulatory representation and build a single attribute model based on E2E framework. They observed that this method underperforms under most of the monolingual settings. However, it can significantly outperform usual E2E models under multilingual training, which implies that articulatory features can be shared across languages. In [28], the authors incorporated articulatory information by embedding constraints in layer initialization. They built articulatory attributes predictors without explicit supervision and relied on the model itself to learn accurate predictions. By representing each token as a fixed-length encoding vector, they constructed an articulatory attribute projection matrix for each target language and expect the deep learning to fine-tune any incomplete, ambiguous or incorrect details.

## 2.5 Bottom-up ASR Framework

Traditional top-down ASR systems [116] follow a model-driven approach, in which the acoustic signal is decoded using statistical acoustic models, such as HMMs or DNNs, a pronunciation lexicon, and a language model to obtain the most likely word sequence. These systems rely on higher-level constraints and predefined word structures, making them highly effective in large-vocabulary read speech recognition tasks but less robust in spontaneous speech or low-resource scenarios. In contrast, the bottom-up ASR approach [7] is motivated by human speech recognition (HSR) [117, 118] theories, which suggest that the linguistic identity of speech sounds is determined through the incremental accumulation of acoustic evidence. Bottom-up ASR frameworks reverse the hierarchy in top-down ASR systems by focusing first on the detection of low-level speech events, such as articulatory attributes, and then composing higher-level linguistic structures from these observations.

Bottom-up ASR architectures offer several compelling advantages compared to traditional approaches. First, they are highly modular and interpretable. Because processing is organized into distinct detectors and integration stages, the system provides rich diagnostic information at each level. It becomes easier to analyze errors, where one can inspect which attribute detector failed or which combination of features led to a wrong prediction. This advantage is largely impossible in an top-down or end-to-end recognizer. This modularity also means one can evaluate and tune individual components, e.g., a manner of articulation detector, in isolation using feature-specific test sets. Second, designing and refining an individual detector is often more straightforward than modifying a large and integrated ASR system. Researchers can focus on one module at a time and incorporate domain knowledge to make it as reliable as possible. Over time, this leads to a library of expert classifiers that collectively cover the acoustic space. Third, bottom-up systems naturally exploit multi-level evidence and can be more robust in challenging conditions [119]. Since attributes

are detected independently across the speech signal, the framework can capture overlapping and redundant cues. For example, manner and place features spread over adjacent phonemes can be picked up without needing perfectly timed phone boundaries. Fourth, bottom-up ASR is advantageous in low-resource settings and for unseen inputs. Since it does not rely strictly on a fixed vocabulary or massive language model training, it can recognize patterns not seen in training by piecing together detected pronunciation units [120]. As a result, standard top-down HMM-based ASR notoriously struggles with out-of-vocabulary (OOV) words, whereas a bottom-up system can hypothesize an OOV word by its attribute sequence.

A number of studies have demonstrated the effectiveness of bottom-up frameworks. For instance, Bromberg et al. [22] trained neural detectors for 14 phonological features and used a CRF-based integration model, achieving competitive results on TIMIT. Metze and Waibel [16] showed that flexible streaming architectures leveraging articulatory features can improve robustness in reverberant conditions. Liu’s landmark-based approach [121] and Juneja’s event-based recognition system [122] further demonstrate the feasibility of using detected speech events to guide recognition. These validate the viability of bottom-up systems as an alternative or complement to traditional top-down ASR.

## CHAPTER 3

### LANGUAGE-UNIVERSAL SPEECH ATTRIBUTE MODELING

In this chapter, we will introduce the proposed language-universal speech attributes as an alternative to conventional modeling units in automatic speech recognition (ASR). Modern ASR systems commonly adopt token units such as characters or phonemes to represent and transcribe speech. These tokenization strategies, while effective under high-resource and monolingual settings, present several challenges when extended to multilingual or low-resource scenarios. To address these limitations, we propose to model speech using articulatory attributes, which are compact and interpretable features that describe how speech sounds are produced. These attributes, such as manner of articulation, place of articulation, and voicing, are shared across languages and grounded in universal phonetic theory. By adopting these language-independent features as modeling units, we aim to improve the generalization and efficiency of ASR systems, particularly in low-resource and multilingual settings.

The remainder of this chapter introduces a unified inventory of speech attributes, derived from phonetic principles and large-scale phoneme inventories. We then explore methods for modeling these attributes both independently and jointly, and evaluate their effectiveness in monolingual and multilingual ASR tasks. Through extensive experiments and analysis, we demonstrate that speech attribute modeling leads to more compact and balanced token sets, better utilization of limited training data, and improved performance under challenging conditions.

#### 3.1 Universal Attribute Inventory

The set of speech attributes used in this dissertation are distinctive features [104, 123, 105, 124]. The attribute inventory normally consists of three major categories: *manner of ar-*

*ticulation* (M), *place of articulation* (P), and *voicing* (V). These attributes are designed to capture articulatory properties that are consistent across languages, providing a compact yet expressive representation of speech sounds. The manner of articulation describes how the airflow is manipulated in the vocal tract during speech production, i.e., whether it is completely blocked, partially obstructed, or allowed to flow freely. It reflects the dynamic interaction between articulators that gives rise to different sound types. The *place of articulation*, on the other hand, indicates where the constriction occurs in the vocal tract, i.e., the specific anatomical location or set of articulators involved in producing a sound. Voicing simply refers to whether the vocal cords vibrate during the production of a sound

The *manner of articulation* includes eleven features (nasal, stop, affricate, fricative, flap, trill, approximant, click, ejective, implosive, and vowel). These classes span the major airflow patterns and articulatory gestures found in human languages. The *place of articulation* set comprises ten features (bilabial, labiodental, dental, alveolar, postalveolar, retroflex, palatal, velar, uvular, glottal). These places cover the full length of the vocal tract from the front of the mouth to the vocal folds, capturing the spatial diversity of consonant production. In addition, *voicing* consists of voiced or voiceless.

As pointed out in [125], a difficulty in using manner and place of articulation for ASR applications is that vowels and consonants cannot be mapped into a common linguistic space, because place of articulation has been differently defined for them. As shown in the official IPA chart of vowels (Figure 3.1), we can see that vowels are characterized by a totally different set of phonological features. Therefore, there are two ways to categorize vowels. The first approach simply views all vowels as the same attribute, resulting in a total of fifteen attributes in place of articulation.

The second approach aims to distinguish vowels using their own set of articulatory features, rather than treating them as a single undifferentiated category. To ensure clear separation avoid representational ambiguity, we define a dedicated set of vowel-specific attributes based on the IPA vowel chart shown in Table 3.1. Specifically, we introduce

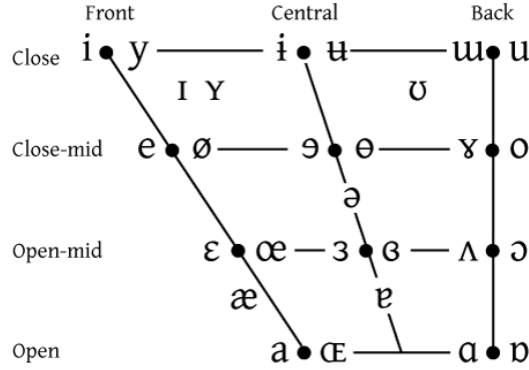


Figure 3.1: IPA chart of vowels.

two articulatory dimensions: Height (H) and Backness (B). Height captures the vertical tongue position during vowel articulation and consists of seven classes: high, semi-high, upper-mid, mid, lower-mid, semi-low, and low. These categories extend the conventional IPA classification by incorporating intermediate levels such as semi-high, mid, and semi-low, allowing for finer-grained distinctions. Backness, on the other hand, represents the horizontal tongue position and is categorized as front, central, or back. Together, the Height and Backness features offer a systematic, interpretable, and language-independent scheme for categorizing vowels, enabling their integration into a unified attribute-based modeling framework for ASR.

The inventory of all universal attribute units adopted in this dissertation is shown in Table 3.2. Phonemes are categorized into each attribute based on the IPA [13] and PHOIBLE [126], a large database of phone inventories for more than 2000 languages. It is impor-

Table 3.1: IPA vowel table organized by tongue height (rows) and backness (columns), illustrating the full inventory of vowel attributes used in this work.

	front		central		back	
high	i	y	ɨ	ʉ	ɯ	u
semi-high	ɪ	ʏ				ʊ
upper-mid	e	ø	ɘ	ɵ	ɤ	o
mid			ə			
lower-mid	ɛ	œ	ɜ	ɞ	ʌ	ɔ
semi-low	æ		ɶ			
low	a	ɶ			ɑ	ɔ̄

Table 3.2: Full inventory of language-universal speech attributes used in this dissertation.

Category	Attributes
Manner (M)	nasal, stop, affricate, fricative, flap, trill, approximant, click, ejective, implosive, vowel
Place (P)	bilabial, labiodental, dental, alveolar, palato-alveolar, retroflex, alveolo-palatal, palatal, velar, uvular, glottal, vowel
Voicing (V)	Voiced, Voiceless
Height (H)	high, semi-high, upper-mid, mid, lower-mid, semi-mid, low
Backness (B)	front, central, back

tant to note that not all attribute categories need to be used in every speech recognition task. In fact, the proposed attribute modeling framework offers the flexibility to select an appropriate and task-specific subset of attributes, in order to achieve improved recognition performance.

### 3.2 Language-universal Speech Attribute Modeling

We can model multiple categories of speech attributes either separately, using individual recognizers for each, or jointly by combining different categories into a Cartesian product space. In order to model speech attributes separately, as shown in Figure 3.2(a), a shared acoustic model is used to process input utterance and produce input features. Then, a bank of speech attribute recognizers is trained to map acoustic features to corresponding articulatory-level posteriors.

When model jointly using product space of the two categories, each attribute token is a combination of manner and place of articulation. For instance, the consonant /r/ consists of articulatory attributes tap and alveolar, so will be assigned the token "tap-alveolar"; whereas the vowel /i/ will be assigned "vowel-high" if vowel height is used. As shown in Figure 3.2(b), a single recognizer is utilized to recognize all tokens in the defined product space. By doing so, we significantly make the number of tokens more compact, resulting

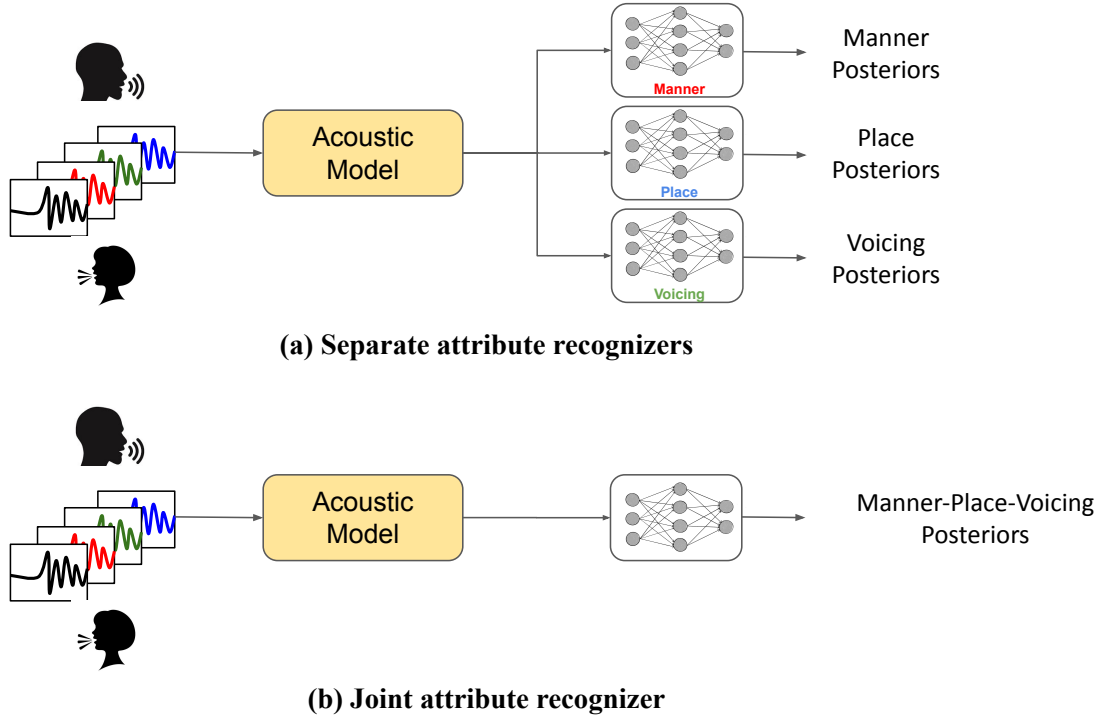


Figure 3.2: Illustration of (a) separate recognizers for each category of speech attributes, and (b) joint modeling the product space of all categories of attributes.

in more training data for each token. Therefore, the attribute-based approach is expected to capture most of the distinct features that describe human sounds. In addition, the proposed universal approach can naturally support zero-shot transfer. This is possible because all phonemes can be categorized in one of the attributes based on Table 3.2. Even if we encountered unseen phonemes from different languages, we can still convert it into attribute tokens and expect the model to infer from those that share the similar acoustic-phonetic features.

### 3.3 Datasets and Experimental Setup

To demonstrate the advantages of modeling speech attributes, we consider two experimental scenarios: a monolingual setting and a more challenging multilingual setting. For the monolingual experiment, we adopt the benchmark English dataset LibriSpeech [127], using its 100-hour clean training set to train the acoustic model illustrated in Figure 3.2. For

the multilingual experiments, we utilize the Multilingual LibriSpeech (MLS) corpus [128], which includes eight languages: English, German, Spanish, French, Italian, Dutch, Portuguese, and Polish, with transcribed speech and standardized train/dev/test splits. We consider two subsets of MLS to simulate different resource conditions: a 10-hour subset for the eight languages, totaling 80 hours, to match the scale of the monolingual setup, and a 1-hour subset to represent low-resource conditions. These settings allow us to evaluate both the scalability and data efficiency of speech attribute modeling.

We use a publicly available pre-trained self-supervised model as our encoder model for all experiments, namely Wav2Vec2.0 [82]. We choose the "base" architecture with 95M parameters leveraging 12 Transformer encoder layers, with embedding size of 768 and 8 attention heads. It is pre-trained on 960 hours of unlabeled audio from LibriSpeech dataset [127]. AdamW optimizer [129] with  $\beta_1 = 0.9$  and  $\beta_2 = 0.98$  is used in training all models, and early stopping criterion is used to terminate the training phase. CTC loss is adopted as main training criterion. During evaluation, we employ Token Error Rate (TER) to measure the performance of the systems. TER is calculated as follows:

$$TER = \frac{S + D + I}{N} \tag{3.1}$$

where  $S$ ,  $D$ ,  $I$  are the numbers of substitution, deletion, and insertion errors, respectively, and  $N$  is the length of the reference text. Three common tokens are considered and compared in the experiments, namely characters, phonemes, and proposed speech attributes. For the speech attributes, we primarily focus on two categories: manner of articulation (M) and place of articulation (P), as these two subsets are generally sufficient to characterize the speech recognition tasks, which will be discussed in the following chapters.

### 3.4 Analysis and Discussions of Advantages of Speech Attributes

#### 3.4.1 Compact and Balanced Inventories of Speech Attributes

We begin by demonstrating several desirable properties of speech attributes. To this end, we use the FLEURS dataset [103], which offers broad linguistic coverage across diverse regions. We randomly select 20 languages spanning four major geographic areas: Europe, Central Asia/Middle East, Africa, and South Asia. The unit counts are plotted against the number of languages in this regional order, as shown in Figure 3.3. As shown in the figure, the number of characters and phonemes steadily increases as more languages are added, whereas sets of speech attributes based on manner of articulation (M), place of articulation (P), and voicing (V), including combined feature product sets M-P and M-P-V, remain relatively compact in size. For the first 12 languages, which mostly originate from European families, the character and phoneme inventory sizes remain nearly constant, reflecting shared alphabets and similar phonemic inventories. However, as additional languages from Central Asia, Africa, and South Asia are introduced, the total character count expands almost exponentially, while the phoneme count grows more slowly. This divergence occurs because new languages often bring distinct writing systems, which dramatically increas-

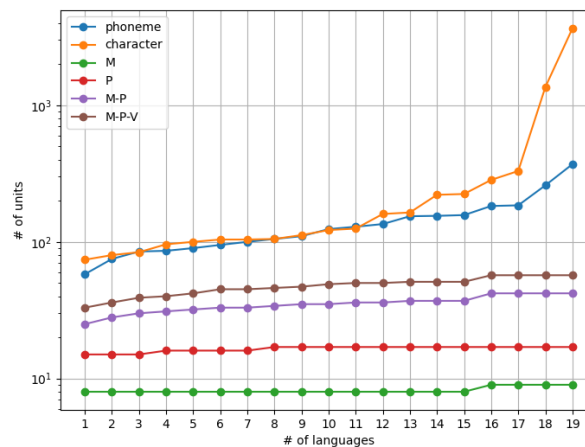


Figure 3.3: Unit inventory size vs. number of languages from the FLEURS dataset.

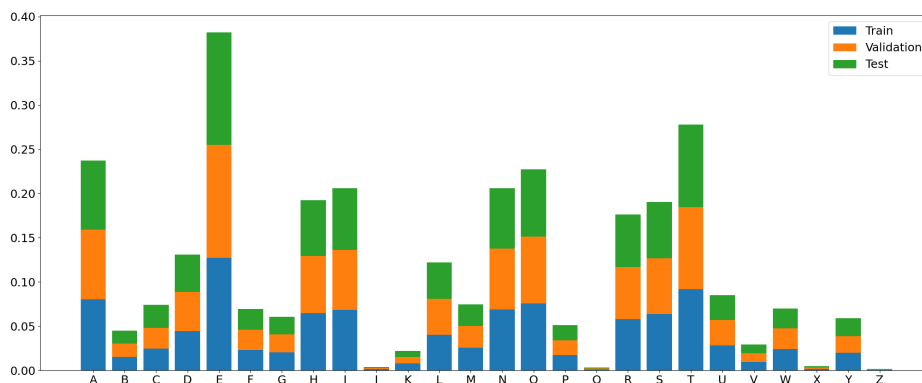


Figure 3.4: Token distribution of characters (in uppercase) in the LibriSpeech train, validation, and test sets.

ing the unique character set and some novel phonemes, whereas several sounds are shared across languages, so the phoneme inventory enlarges more modestly. In contrast, articulatory speech attributes, such as manner and place of articulation are universal categories of speech sounds, so the attribute inventory remains fixed and sharable across languages.

In addition, we analyze the token distributions of characters and phonemes, as shown in Figure 3.4 and Figure 3.5, respectively. Both distributions, taken from the LibriSpeech dataset, exhibit pronounced long-tailed behavior, with a small number of highly frequent tokens dominating the corpus. For example, characters such as /T/ and phonemes such as /n/ occur with high frequency, while others such as /J/ and /r/ appear only sparsely. This skew is consistently observed across the training, validation, and test splits, and is indicative of the inherent frequency imbalance present in natural language. In particular, these figures reflect statistics from a single monolingual corpus, i.e., English. In a multilingual setting, especially one including typologically diverse languages, such an imbalance would likely be significantly worse. Differences in orthographic systems, phoneme inventories, and usage patterns across languages would introduce even greater variability and sparsity into the token distributions.

By contrast, the attribute-based tokens exhibit significantly more balanced and compact

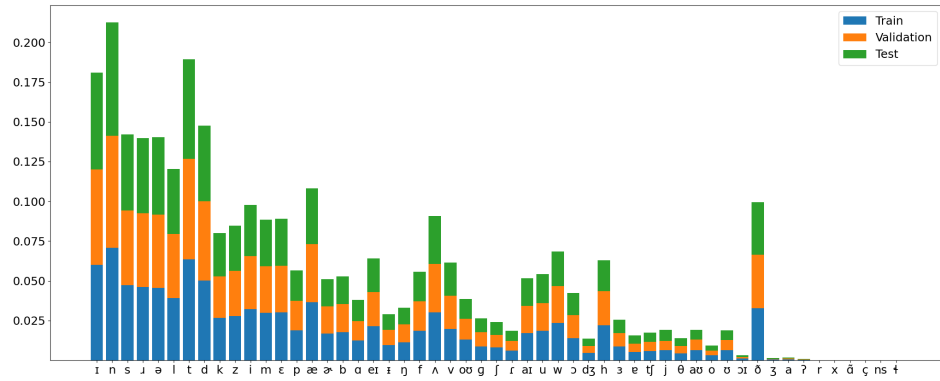


Figure 3.5: Token distribution of phonemes in the LibriSpeech train, validation, and test sets.

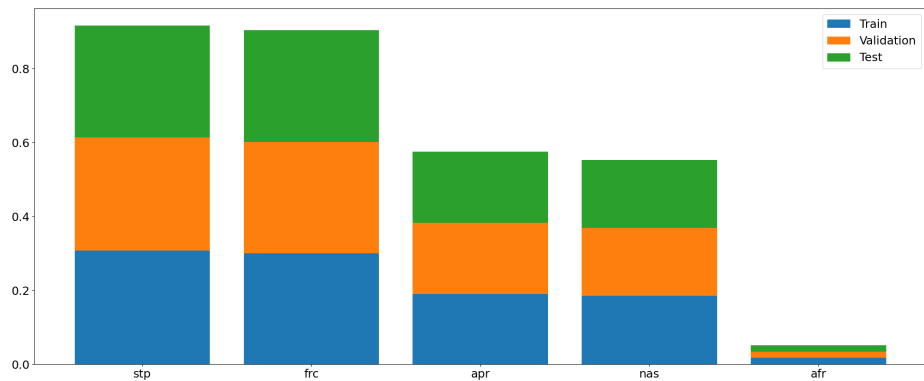


Figure 3.6: Token distribution of manner of articulations in the LibriSpeech train, validation, and test sets.

distributions. As shown in Figure 3.6 and Figure 3.7, pooling phonemes into broader articulatory categories results in relative frequency ratios that are substantially more uniform across the training, validation, and test sets. To focus on consonantal contrasts, we omit all tokens associated with the “vowel” category, which tend to dominate due to their higher frequency in all languages. In Figure 3.6, the coarse manner of articulation comprises five categories for English: stop (stp), fricative (frc), approximant (apr), nasal (nas), and affricate (afr). With the exception of affricates—which are less frequent but still account for approximately 2% of the total tokens—all other categories each represent around 20%

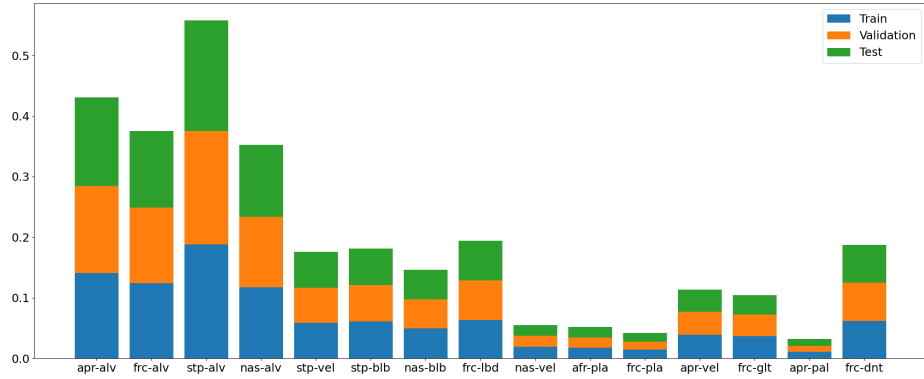


Figure 3.7: Token distribution of manner-place product in the LibriSpeech train, validation, and test sets.

of the dataset. When we incorporate place of articulation as additional information and form the manner–place (M-P) product space, as shown in Figure 3.7, the token usage is further distributed across the combined categories in a relatively even manner. This improved balance is a direct result of aggregating phonemes into shared articulatory classes, which increases coverage and reduces the number of underrepresented units. Such uniformity in the token distribution is expected to enhance model robustness and generalization, particularly in multilingual and low-resource ASR scenarios where data scarcity and class imbalance pose substantial challenges, which we will demonstrate in the next section.

### 3.4.2 Monolingual Speech Attribute Modeling

Table 3.3 presents the TER for each English character in the LibriSpeech test set. The overall TER is 1.25%, demonstrating strong recognition performance when using a self-supervised pretrained model combined with a large-scale training corpus. Nevertheless, substantial variation exists across individual characters. For instance, the character /M/ achieves the lowest TER at 0.23%, while /Z/ records the highest at 4.00%. Notably, two of the most error-prone characters, /J/ and /Z/, also appear among the least frequent tokens in the dataset, as shown in Figure 3.4. This observation suggests a strong correlation between

Table 3.3: Token error rate (%) for characters in the LibriSpeech test set, where the overall performance is 1.25%.

Character	TER (%)	Character	TER (%)	Character	TER (%)
A	1.66	B	0.35	C	0.93
D	1.97	E	1.49	F	0.86
G	0.69	H	1.05	I	0.97
J	2.18	K	1.04	L	0.97
M	0.23	N	0.54	O	0.75
P	0.45	Q	1.28	R	0.36
S	0.86	T	0.66	U	1.81
V	0.59	W	0.73	X	1.48
Y	1.13	Z	4.00		

Table 3.4: Token error rate (%) for phonemes in the LibriSpeech test set, where the overall performance is 1.86%.

Phoneme	TER (%)	Phoneme	TER (%)	Phoneme	TER (%)
t	1.59	s	1.22	b	0.63
n	0.62	d	3.08	w	0.60
z	2.25	ɹ	1.68	k	0.55
f	0.51	ð	0.91	h	1.15
l	0.75	r	3.80	v	1.22
ʃ	1.84	ŋ	0.59	tʃ	1.78
θ	2.60	p	0.66	dʒ	2.24
j	2.59	g	1.54	ʔ	4.29
ʒ	1.12	m	0.49	ɪ	2.97
i	1.31	u	1.54	a	4.87
ʌ	1.20	ɔ	2.67	ɛ	2.73
aʊ	0.98	ɔɪ	0.54	ɐ	8.64

token frequency and recognition accuracy. As the ASR system scales to a multilingual setting, where additional writing systems and characters are introduced, the number of infrequent or sparsely represented characters is likely to increase. Such an expansion would exacerbate the imbalance in the training data and further degrade recognition performance, especially for underrepresented tokens.

A similar trend is observed at the phoneme level, as shown in Table 3.4. The overall TER for phoneme recognition is slightly higher, at 1.86%, yet still demonstrates strong performance in a monolingual setting with sufficient training data. Among the consonants, /m/ achieves the best performance with a TER of 0.49%, and the highest counterpart is

Table 3.5: Token error rate (%) for manner and place of articulation in the LibriSpeech test set, where the overall performance is 0.77% and 0.81% respectively.

<b>Manner</b>	<b>TER (%)</b>	<b>Place</b>	<b>TER (%)</b>
nasal	0.35	bilabial	1.14
stop	1.28	labiodental	0.96
affricate	1.19	dental	1.05
fricative	0.83	alveolar	0.53
approximant	1.32	postalveolar	1.21
vowel	0.38	palatal	1.40
		velar	0.43

Table 3.6: Token error rate (%) for manner-place product features in the LibriSpeech test set, where the overall performance is 0.77%.

<b>Manner-Place</b>	<b>TER (%)</b>	<b>Manner-Place</b>	<b>TER (%)</b>
approximant-palatal	2.66	stop-alveolar	1.74
fricative-postalveolar	1.44	affricate-postalveolar	1.34
approximant-alveolar	1.02	fricative-dental	0.97
fricative-alveolar	0.84	nasal-velar	0.74
fricative-labiodental	0.62	nasal-alveolar	0.57
approximant-velar	0.55	stop-velar	0.49
stop-bilabial	0.46	nasal-bilabial	0.30

4.29% for phoneme /ʔ/, which is the least frequent consonant as shown in Figure 3.5. For vowels, the TER also varies widely, i.e., vowel /ɔɪ/ achieves the best TER of 0.54% while /e/ has the worst TER of 8.64%. These results reveal that, despite strong average performance, there remains a substantial performance gap between high-frequency and low-frequency phonemes.

In contrast, Table 3.5 presents the recognition results for both manner and place of articulation categories in the LibriSpeech test set. We model the speech attributes separately as described in section 3.2. The overall TERs are 0.77% and 0.81% for manner and place, respectively which is substantially lower than those observed for character (1.25%) and phoneme (1.86%) recognition. This demonstrates that speech attribute modeling can offer more accurate and reliable recognition, especially when leveraging universal, interpretable features rooted in articulatory phonetics. In particular, manner categories such as nasal and vowel achieve TERs below 0.4%, indicating high discriminability for these broad ar-

tulatory classes. Even for more complex or less frequent classes, such as affricate and approximant, the TER remains consistently low. Similarly, among place categories, common articulatory features like alveolar is recognized with high accuracy, achieving TERs of 0.53%. To further investigate the effectiveness of combining articulatory features and joint modeling, Table 3.6 reports the recognition performance for the manner–place product space. This result confirms that modeling manner and place jointly does not compromise recognition accuracy, and instead offers a richer representation of speech sounds. Several manner–place combinations exhibit particularly strong performance. For example, nasal-bilabial achieves a TER of just 0.30%, while stop-bilabial and stop-velar achieve TERs of 0.46% and 0.49%, respectively. These results suggest that common and acoustically distinct categories are reliably recognized. Even relatively complex or lower-frequency combinations, such as affricate-postalveolar and approximant-palatal, yield reasonable error rates. Overall, the M-P token space provides a finer-grained articulatory structure while preserving robustness, making it a promising alternative to traditional modeling units, particularly in tasks requiring compact and interpretable representations.

### 3.4.3 Multilingual Speech Attribute Modeling

In this section, we evaluate the effectiveness of speech attribute modeling under a multilingual scenario using the MLS dataset. Our goal is to examine how well different modeling units, such as characters, phonemes, and speech attributes, perform when trained jointly across multiple languages with limited supervision. To keep the analysis tractable, we report only the overall TER for each language, as character and phoneme token sets are large and language-specific, making per-token analysis impractical in the multilingual setting. For speech attribute modeling, we report the manner of articulation and manner-place combination modeling results similar to the monolingual setup.

Table 3.7 presents results when each of the eight MLS languages is trained with only 10 hours of labeled data. The multilingual character and phoneme models, denoted as  $\text{Multi}_{\text{char}}$

and  $\text{Multi}_{\text{phn}}$ , achieve comparable average TERs of 7.35% and 8.02%, respectively. Compared to their monolingual counterparts (Table 3.3 and Table 3.4), the performance drop is expected due to the reduced amount of data per language. However, we observe large discrepancies across languages. For instance, English shows the highest TER of 16.17%, while Spanish achieves a low 4.36%. This highlights the difficulty of building balanced multilingual systems when supervision is limited and suggests that language-specific phonetic properties and data variability strongly influence multilingual ASR performance.

On the other hand, speech attribute-based systems demonstrate not only superior TERs, with 5.33% for  $\text{Multi}_{\text{M}}$  and 5.55% for  $\text{Multi}_{\text{M-P}}$ , but also substantially lower variability across languages. The standard deviations reduce to 1.77% and 1.79%, respectively. This consistent performance can be attributed to the universality of articulatory features, which are more evenly distributed and better aligned with cross-linguistic phonetic structures. Moreover, despite operating with a smaller and more compact unit inventory, the attribute-based models maintain robust recognition quality, even with only 10 hours of supervision per language.

#### 3.4.4 Effects of Training Data Size

To further evaluate the robustness of different modeling units under extremely limited supervision, we further reduce the amount of training data to 1 hour per language from the MLS dataset. Table 3.8 presents the TER results across eight languages for each system,

Table 3.7: Token Error Rate (%) results eight languages in Multilingual LibriSpeech dataset with 10-hour subset for each language. Avg. stands for the average TER of the eight languages and std. is the standard deviation.

<b>System</b>	en	de	es	fr	it	nl	pt	pl	Avg. (std.)
$\text{Multi}_{\text{char}}$	15.44	6.75	4.15	11.52	5.80	7.39	8.24	5.31	7.35 ( $\pm$ 3.47)
$\text{Multi}_{\text{phn}}$	16.17	7.77	4.36	7.36	6.27	7.87	8.88	5.44	8.02 ( $\pm$ 3.37)
$\text{Multi}_{\text{M}}$	9.45	4.97	3.50	5.11	3.95	5.25	6.40	4.05	5.33 ( $\pm$ 1.77)
$\text{Multi}_{\text{M-P}}$	9.61	5.18	3.73	5.32	4.04	5.47	6.81	4.27	5.55 ( $\pm$ 1.79)

Table 3.8: Token Error Rate (%) results eight languages in Multilingual LibriSpeech dataset with 1-hour subset for each language. Avg. stands for the average TER of the eight languages and std. is the standard deviation.

<b>System</b>	en	de	es	fr	it	nl	pt	pl	Avg
Multi <sub>char</sub>	20.68	12.68	10.21	21.43	9.66	13.45	16.33	10.79	14.40 ( $\pm$ 4.32)
Multi <sub>phn</sub>	21.42	14.21	10.01	14.56	12.71	14.27	17.94	11.68	14.60 ( $\pm$ 3.37)
Multi <sub>M</sub>	11.71	8.90	8.03	9.63	7.75	9.27	12.86	8.68	9.60 ( $\pm$ 1.67)
Multi <sub>M-P</sub>	12.12	9.06	8.42	9.53	7.67	9.35	12.62	8.67	9.68 ( $\pm$ 1.65)

along with the average TER and standard deviation.

As shown in the table, both Multi<sub>char</sub> and Multi<sub>phn</sub> models suffer from higher TERs and large performance variability across languages, with standard deviations of 4.32 and 3.37 respectively. Notably, English continues to exhibit the highest TER for both systems, likely due to its distinct grapheme-to-phoneme correspondence and phonological structure, which differ significantly from those of the other languages. On the other hand, attribute-based models, Multi<sub>M</sub> and Multi<sub>M-P</sub>, show strong robustness under the 1-hour low-resource condition. Not only do they outperform both character and phoneme systems in average TER by over 34% of relative reduction, but they also exhibit significantly lower variance across languages, with standard deviations around 1.65.

These results highlight two important benefits of articulatory attribute modeling: (1) its compact and interpretable representation allows for more efficient parameter sharing across languages, and (2) it demonstrates better cross-lingual generalization and stability when trained with extremely minimal supervision. It is also worth noting that all eight languages in the MLS dataset belong to the Latin language family, and therefore share relatively similar phonological and orthographic characteristics. However, as we expand to languages from more diverse linguistic families, character- and phoneme-based modeling is likely to encounter greater challenges and performance degradation. This reinforces our core hypothesis that articulatory attributes serve as robust universal units for multilingual ASR, especially in multilingual and low-resource settings.

### 3.5 Summary

In this chapter, we introduced a language-universal speech attribute modeling framework as an alternative to conventional character and phoneme-based units for ASR. We constructed a unified inventory of articulatory attributes grounded in phonetic theory and evaluated both separate and joint modeling strategies using manner, place, and voicing categories. Through extensive monolingual and multilingual experiments on the LibriSpeech and MLS dataset, we demonstrated that attribute-based tokens yield more compact and balanced distributions, with significantly lower token error rates than phoneme and character units. Modeling attributes independently or in structured product spaces enables more robust learning by aggregating training data across similar phonemes, leading to better generalization and robustness. These findings suggest that attribute-based modeling offers a scalable and interpretable solution for multilingual and low-resource ASR.

## CHAPTER 4

### BOOSTING END-TO-END MULTILINGUAL PHONEME RECOGNITION THROUGH EXPLOITING KNOWLEDGE-BASED CONSTRAINTS

In this chapter, we propose a first step toward multilingual end-to-end automatic speech recognition (ASR) by integrating knowledge about speech articulators. Specifically, several deterministic attribute-to-phoneme mapping matrices are constructed based on the pre-defined set of universal attribute inventory, which projects the knowledge-rich articulatory attribute logits, into output phoneme logits. The mapping puts knowledge-based constraints to limit inconsistency with acoustic-phonetic evidence in the integrated prediction. Combined with phoneme recognition, our phone recognizer is able to infer from both attribute and phoneme information. The proposed joint multilingual model is evaluated through phoneme recognition. Experimental results show that our proposed solution outperforms conventional multilingual approaches with a relative improvement of 6.85% on average, and it also demonstrates a much better performance compared to monolingual model. Further analysis conclusively demonstrates that the proposed solution eliminates phoneme predictions that are inconsistent with attributes.

#### 4.1 Proposed Language-universal Phoneme Recognition Framework

##### 4.1.1 Language-universal Speech Attribute Inventory

The set of speech attributes used in this work is a subset of Table 3.2, namely, seven *manner of articulation* (M) classes (approximant, tap, fricative, affricate, nasal, stop, and vowel), and ten *place of articulation* (P) (bilabial, labiodental, dental, alveolar, postalveolar, retroflex, palatal, velar, uvular, glottal). Since we mainly focus on using speech attribute modeling as a knowledge-based constraints, we do not categorize vowel based on

Table 4.1: Inventory of universal speech attributes used for phoneme recognition.

Category	Attributes
Manner (M)	nasal, stop, affricate, fricative, approximant, vowel
Place (P)	bilabial, labiodental, dental, alveolar, postalveolar, retroflex, palatal, velar, uvular, glottal, vowel

their height or backness as shown in Table 3.2. Instead, we map all vowels in the same ‘vowel’ feature in this chapter. This is motivated by the fact that vowel segments typically have more training data and are acoustically more stable, reducing the need for fine-grained subdivision. Grouping vowels under a unified feature helps avoid over-fragmentation of the attribute space while preserving modeling efficiency and robustness. The overall inventory of universal attribute units adopted in this chapter is shown in Table 4.1.

#### 4.1.2 Language-universal Attribute and Phoneme Recognizers

Figure 4.1 (b) illustrates our proposed approach, which comprises a language-universal encoder along with a set of attribute recognizers, a phoneme recognizer, and a set of attribute-to-phoneme mapping matrices. The encoder first produces the input features and feeds them to both the phoneme recognizer and attribute recognizers. The phoneme recognizer aims to predict output phoneme tokens for each language.

A set of two attribute recognizers, for *manner of articulation* and *place of articulation* respectively, are placed to project input features to attribute logits. The attribute recognizers are trained as a predictor for articulatory attributes and optimized with full supervision via CTC criterion. Not only does the attribute recognizers abstract away from the phoneme recognizers, which contributes to the improvement in the multilingual acoustic modeling, the models also provides us the capability to predict universal articulatory speech attributes themselves. When trained with a large amount of available languages, these recognizers are expected to capture most of the distinct features that describe human sounds. Moreover, it has been shown that by joint modeling the articulatory and phonetic features, the

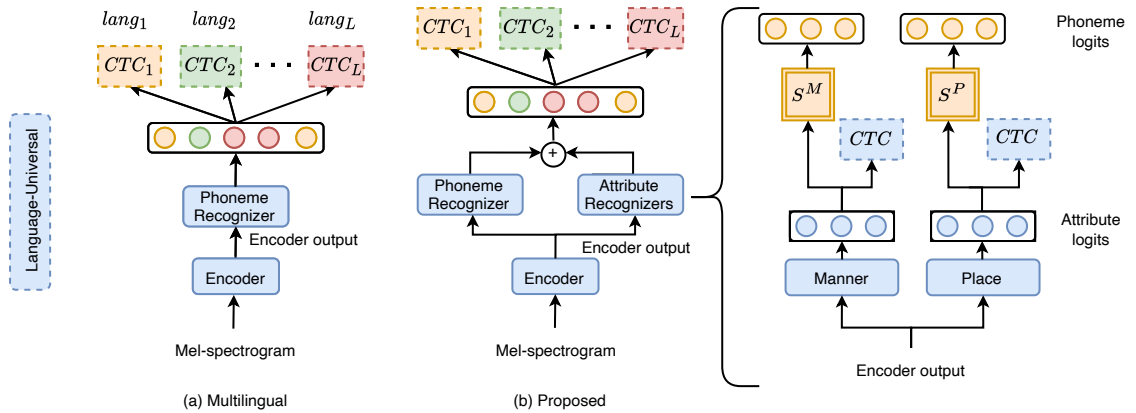


Figure 4.1: An overview of the traditional and proposed approaches. In (a), a multilingual approach predicting all the phoneme tokens for all languages. In (b), our proposed system contains two articulatory attribute recognizers for manner and place of articulation respectively, that are shared across languages. The attribute recognizers will produce attribute logits that will later be transformed to phoneme logits with an attribute-to-phoneme mapping.  $S^C$  stands for the deterministic articulatory attribute-to-phoneme mapping matrix for a specific attribute category  $C$ .

performance of phoneme recognition can be improved accordingly [112, 130].

#### 4.1.3 Attribute-to-phoneme Mapping

Following the language-universal attribute recognizers, two attribute-to-phoneme mapping matrices are used to project the attribute logits to the phoneme logits. Suppose there are  $L$  training languages and each language  $l$  has its own phoneme inventory  $P_l$ . A union phoneme inventory  $P_{\text{uni}}$  can be created by enumerating all the phonemes appearing in the data for all languages:

$$P_{\text{uni}} = \bigcup_{1 \leq l \leq L} P_l \quad (4.1)$$

We can then construct an articulatory attribute-phoneme mapping matrix  $S^C \in \mathbb{R}^{|C| \times |P_{\text{uni}}|}$ , where  $|P_{\text{uni}}|$  represents the number of phoneme tokens for the target language and  $|C|$  represents the number of articulator attributes for a particular category of attribute, such as *manner of articulation* ( $M$ ), and *place of articulation* ( $P$ ), which means  $C$  is one of  $M$  or  $P$ . We can obtain  $S^C = \{0, 1\}^{|C| \times |P_{\text{uni}}|}$ , describing the association of attributes and

Table 4.2: Training corpora and size in hours for each language.

Language	Family	Corpora	Train (hrs)	Test (hrs)
English	West Germanic	LibriSpeech	100.4	5.4
Kyrgyz	Turkic	CommonVoice	2.6	2.0
Dutch	West Germanic	CommonVoice	11.5	7.0
Russian	Slavic	CommonVoice	23.5	13.3
Swedish	North Germanic	CommonVoice	2.1	2.0
Tatar	Turkic	CommonVoice	11.9	4.6

phonemes in each language  $l$ : Suppose the phoneme  $p \in P_{\text{uni}}$  has the column index  $j$  where  $1 \leq j \leq |P_{\text{uni}}|$ , attribute  $a \in C$  has the row index  $i$  where  $1 \leq i \leq |C|$ , if  $p$  has the presence of attribute  $a$ , then  $(i, j)$  entry of the matrix  $S^C$  is set to 1, otherwise, it is assigned to 0. The attribute-to-phoneme mapping matrix is deterministic and fixed to accurately describe the relation between each attribute-phoneme pair. It also serves as a strong knowledge-based constraints to reduce any nonsense phoneme outputs.

As shown in Figure 4.1, the attribute-to-phoneme mapping aims to project attribute logits to phoneme logits. However, the attribute-to-phoneme projection does not yield a one-to-one correspondence, meaning multiple phonemes could potentially share identical logits after mapping. For example, as we mentioned in section 3.1, vowels are categorized into the same attribute *vowel*, and therefore, the phoneme logits for all the vowels will simultaneously receive the same logits after projection. To address this limitation, it becomes necessary to introduce supplementary information for distinguishing among phonemes grouped within the same attributes. To address this, we combine the output logits of the attribute-to-phoneme projection with conventional phoneme logits to make the final predictions.

## 4.2 Datasets and Experimental Settings

In contrast to chapter 3, this chapter addresses a more challenging scenario involving a more diverse set of language families. Specifically, six languages are included in our experiments, English (en) from LibriSpeech [127] and Kyrgyz (ky), Dutch (nl), Russian (ru),

Swedish (sv), and Tatar (tt) from the CommonVoice database [131]<sup>1</sup>. Among these, English, Dutch, and Swedish belong to the Germanic family, Kyrgyz and Tatar are from the Turkic family, and Russian is a Slavic language. These languages exhibit substantial differences in pronunciation and writing systems, which further increase the complexity of the task. Non-aligned phoneme transcription of each audio sample are obtained by running the open-source tool phonemizer<sup>2</sup> on their corresponding text scripts. After processing all the transcriptions and removing some of the utterances that contain ambiguous phoneme tokens, the resulting training and testing dataset, and their corresponding sizes are shown in Table 4.2.

The Conformer architecture [132] is investigated in this work. Our encoder first processes the input mel-spectrogram with a convolution subsampling layer and then with a number of conformer blocks. The Conformer block consists of three modules; a feed-forward module, a multi-head self-attention module, and a convolution module. The feed-forward module consists of a normalization layer and a linear layer with a Swish activation function [133] followed by another linear layer. The multihead self-attention module is composed of a layer-normalization and multi-head self-attention with relative sinusoidal positional embedding employed in Transformer-XL model [134]. The convolution module includes a layer normalization, a point-wise convolution layer with a gated linear unit (GLU) activation function [135], and a 1-D depth-wise convolution layer. The depth-wise convolution layer is succeeded by a batch normalization layer, a Swish activation, and a point-wise convolution layer. We follow the *small* Conformer setting, which incorporates 16 conformer blocks, each with 4 attention heads and the dimension is 144.

Phoneme and attribute recognizers are both one single-LSTM-layer decoder followed by a linear layer. The decoder dimension is 320. The models are trained with the Adam optimizer [136] with  $\beta_1 = 0.9$ ,  $\beta_2 = 0.98$  and  $\epsilon = 10^{-9}$ . A transformer learning rate scheduler [137], with 10k warm-up steps and peak learning rate  $0.05/\sqrt{d}$  where  $d$  is the model

---

<sup>1</sup><https://voice.mozilla.org>

<sup>2</sup><https://github.com/bootphon/phonemizer>

Table 4.3: Phoneme Error Rate (PER(%)) results for traditional models and proposed methods. Traditional methods include monolingual and multilingual models. Our proposed methods incorporated different categories of attributes, including *manner of articulation*, and *place of articulation*. The average PER (Avg) for six languages is also reported.

System	en	ky	nl	ru	sv	tt	Avg
Mono <sub>base</sub>	7.25	46.86	47.37	23.7	64.13	54.31	40.60
Multi <sub>base</sub>	16.45	41.13	<b>44.82</b>	28.94	45.79	56.84	39.00
Multi <sub>attr</sub> (ours)	<b>13.59</b>	<b>36.91</b>	47.34	<b>25.92</b>	<b>40.06</b>	<b>55.30</b>	<b>36.53</b>
Relative improvement (%)	17.39	10.26	-5.62	10.44	12.51	2.71	6.85

dimension in conformer encoder, is used for adjusting learning rate. Each experiment is trained for 50 epochs with early stopping criterion.

### 4.3 Experimental Results and Discussions

Table 4.3 shows the phoneme error rate (PER) for several phoneme architectures in both monolingual and multilingual settings, including our system. The monolingual baseline system, denoted as Mono<sub>base</sub>, is a model trained from scratch using only the target language data; whereas, Multi<sub>base</sub> refers to the baseline of multilingual modeling end-to-end trained on the union of all the phonemes from all target languages. On the one hand, the Multi<sub>base</sub> can improve the performance for some languages owing fewer training samples, for example, Kyrgyz and Swedish; on the other hand, it also creates confusion for resource-rich languages, such as English (7.25% to 16.45%), Russian (23.70% to 28.94%), and Tatar (54.31% to 56.84%), causing a degradation of the overall PER. Comparing the third with the second row in Table 4.3, we can see that our proposed Multi<sub>attr</sub> outperforms the Multi<sub>base</sub> for the majority of the target language with a relative improvement of 17.39%, 10.26%, 10.44%, 12.51%, and 2.71% for English, Kyrgyz, Russian, Swedish, and Tatar, respectively.

Comparing with Mono<sub>base</sub>, we observe significant improvements for Kyrgyz (46.86% to 36.91%), and Swedish (64.13% to 40.06%). For resource-rich languages, such as En-

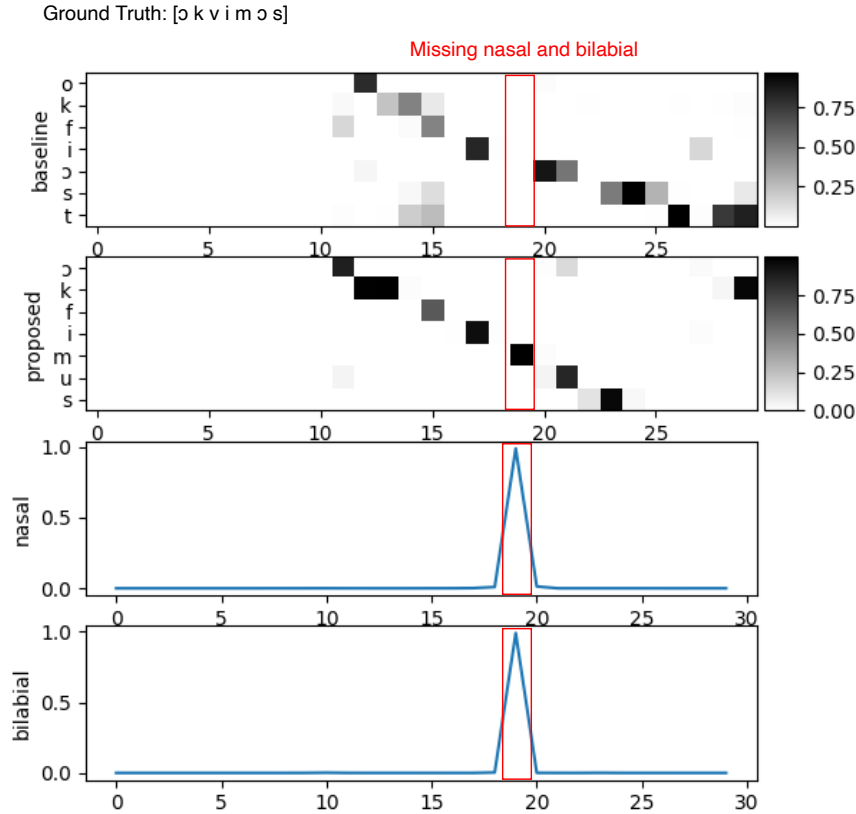


Figure 4.2: A qualitative example for utterance numbered 19557664 from the Swedish dataset. The above two plots are frame level posteriors of each phoneme corresponding to the ground truth, and the below two plots depict the activation curve for *nasal* and *bilabial*.

lish, we see that, while we do not yet compete with the monolingual results in terms of PER, we effectively shrink the performance gap between the multilingual model and the monolingual model. It is also worth pointing out that our proposed  $\text{Multi}_{\text{attr}}$  requires no prior knowledge regarding the target languages involved in the experiments. In contrast,  $\text{Mono}_{\text{base}}$  needs to be trained and tested with specific language information.

In order to gain more insights into the properties of our approach, we investigate into its ability to reduce inconsistent phonetic-acoustic evidence. Figure 4.2 shows one qualitative example from the Swedish dataset. While the baseline system fails to predict the output phoneme [m] which has the presence of *nasal* and *bilabial* attributes from the ground truth, our manner and place recognizers clearly indicate the presence of those two events in the area where the mistakes occurs. The error is corrected by our solution since it has the

ability to predict and integrate attributes information.

#### **4.4 Summary**

In this chapter, we have presented the initial step to build a multilingual ASR system. Our results show that the proposed system can yield better performance than the conventional multilingual model in terms of phoneme error rate. Evaluated on six languages, we observed significant improvements for resource-limited languages, namely, Kyrgyz and Swedish. For languages with abundant resources, we successfully mitigated performance degradation. Further analysis has demonstrated that our proposed method can effectively eliminate inconsistencies with acoustic-phonetic evidence. Although the system does not yet fully achieve the ultimate goal of multilingual speech recognition, our findings support the initial motivation that incorporating speech attribute modeling is a promising direction for improving multilingual ASR performance.

## CHAPTER 5

### LANGUAGE-UNIVERSAL SPEECH ATTRIBUTES MODELING FOR ZERO-SHOT MULTILINGUAL SPOKEN KEYWORD RECOGNITION

In this chapter, our aim is to extend the universal attribute modeling approach from phoneme to word and present a solution toward designing multilingual E2E ASR systems. The design concept of the proposed system consists of a comprehensive inventory of fundamental attribute units, a pre-trained Wav2Vec2.0 encoder, a linear output layer, and a non-trainable pronunciation model. The encoder initially generates robust features. Then, these features are fed into a linear layer modeling articulatory characteristics directly by generating sequences of posteriors for each attribute token. A non-trainable pronunciation model is adopted to map such sequences to the most likely keywords. To further enhance the language-universality, a domain adversarial training (DAT) [138] technique is integrated into our proposed system, aiming to assist the pre-trained encoder in learning language-invariant features.

#### 5.1 Introduction

Spoken Keyword Recognition (SKR) is an important sub-task within ASR. SKR aims to identify a specific target keyword out of a set of predefined candidates [139, 140], also known as speech command recognition. Owing to its wide applicability to various domains, such as smart home devices [141] or crime defections [142], SKR has long been an important research topic in the speech processing field [143, 144]. Moreover, the rising popularity of AI-assisted devices has sparked a heightened interest in SKR for customizable user experiences. However, a common requirement to build a high-performance SKR system is to prepare extensive labeled training data, which is often impractical in real-world scenarios. In fact, it is generally favorable to build a language-universal SKR system that

can handle to any input or language, particularly in zero-shot transfer [89] or low-resource scenarios [62].

Recent studies have treated SKR as a classification task [145, 146, 147], which directly associates audio speech with specific labels. These ASR-free systems have a fixed output dimension that corresponds to the number of the predefined keywords and cannot be altered during testing, limiting their ability to handle out-of-domain scenarios like out-of-vocabulary words or unseen languages. Another category of approaches adopts an ASR-based systems [143, 148], but these typically rely on high-level tokens, such as characters and phonemes. The inherent disadvantage of using those tokens is that they are language-dependent, which means those systems are unable to process tokens from different languages, or keywords that have not been encountered during training. Hence, it is crucial for researchers to explore knowledge sharing among multiple languages, for example, defining a universal set of acoustic-phonetic units that work for multiple or even for all languages, in order to build a language-universal SKR system.

## **5.2 Overview of The Language-universal Attribute Modeling SKR Framework**

### 5.2.1 ASR-based Spoken Keyword Recognition System

Figure 5.1(a) illustrates our proposed ASR-based SKR system, which comprises a pre-trained Wav2Vec2.0 (W2V2) encoder, a Linear output layer, and a non-trainable pronunciation model (PM). The encoder, E, first produces the input features and feeds them to the Linear output layer (O), which aims to predict sequences of token posteriors. The system is trained under full-supervision with corresponding label tokens via the CTC criterion. The non-trainable pronunciation model consists of a lexicon with the collection of all the keywords and their corresponding token sequences in characters, phonemes, or attributes. Given the predicted token posteriors, PM applies a simple beam search to obtain the most likely keyword. The ASR-based system architecture has the capability to generalize to out-of-domain scenarios, such as out-of-vocabulary keywords and unseen languages, due to its

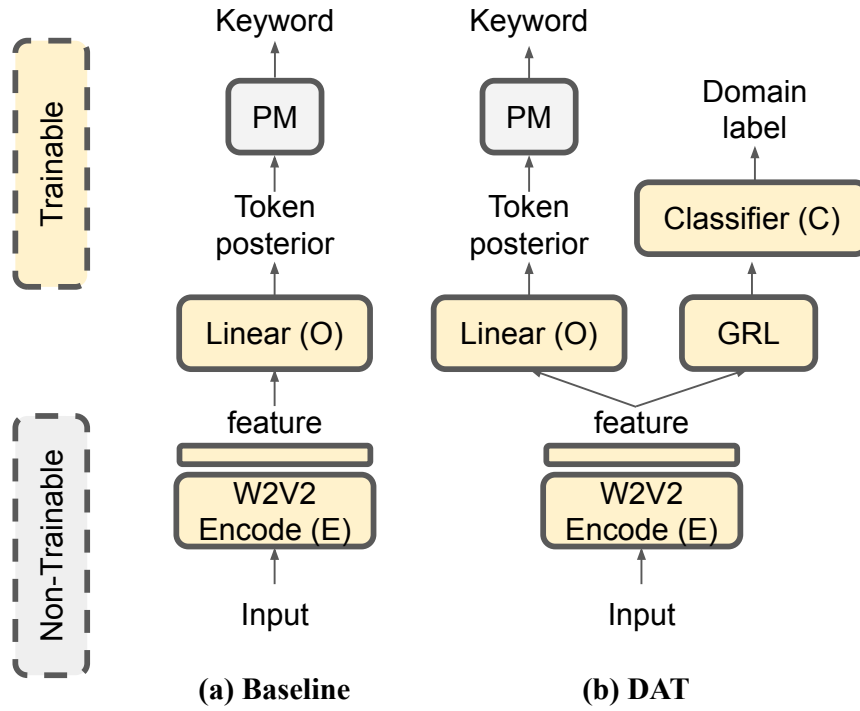


Figure 5.1: Overview of the language-universal spoken keyword recognition framework. "W2V2" refers to the Wav2Vec2.0 pre-trained model, and "PM" refers to pronunciation model. GRL refers to gradient reversal layer and Classifier denotes a language classifier.

ability to generate arbitrary sequences of tokens, and then rely on the provided lexicon to perform recognition.

### 5.2.2 DAT for Universal Spoken Keyword Recognition

As depicted in Figure 5.1(b), The DAT framework we propose consists of a language-invariant encoder  $E$ , linear output layer  $O$ , and a language classifier  $C$ . During training, a gradient reversal layer (GRL) is used to minimize the ability of the encoder to distinguish languages. In other words, the output features from the encoder  $E$  are expected to be language-invariant representations. The losses for  $O$  and  $C$  are denoted as  $\mathcal{L}_O$  and  $\mathcal{L}_C$ , and the respective parameters of  $E$ ,  $O$ ,  $C$  as  $\theta_E$ ,  $\theta_O$ ,  $\theta_C$ . We update each weight following

specified gradient descent rules,

$$\begin{aligned}
 \theta_E &\leftarrow \theta_E - \alpha \left( \frac{\partial \mathcal{L}_O}{\partial \theta_E} - \lambda \frac{\partial \mathcal{L}_C}{\partial \theta_E} \right) \\
 \theta_C &\leftarrow \theta_C - \alpha \frac{\partial \mathcal{L}_C}{\partial \theta_C} \\
 \theta_O &\leftarrow \theta_O - \alpha \frac{\partial \mathcal{L}_O}{\partial \theta_O}
 \end{aligned} \tag{5.1}$$

where  $\alpha$  is the learning rate and  $\lambda$  is the scale of  $L_C$  gradients.

Integrating DAT into our system presents a clear benefit. Given that collecting exhaustive language data is impractical, models often rely on language information, and cues during recognition. This reliance could compromise their capability to handle previously unseen keywords and languages. DAT helps reduce this dependency by minimizing language information in the encoder’s features, promoting better generalization to new scenarios. Since our proposed framework models speech attributes, which are expected to be relatively invariant to languages in nature, DAT can be complimentary to eliminate the influence of any residual language information in the features.

### 5.3 Datasets and Experimental Settings

We evaluate our proposed language-universal approach on the Multilingual Spoken Words Corpus (MSWC) [149] dataset. As shown in Table 5.1, we take into account eight rich-resource languages as our in-domain in-vocabulary (**ID-IV**) set, namely English, German, French, Persian, Spanish, Russian, Italian, and Polish. **ID-IV** set refers to languages and keywords that have been seen and included during training. We select languages from multiple language families (a group of languages that are related through a common ancestral language), and keywords that have more than 500 training samples for each language. For zero-shot transfer, we consider two scenarios: (i) an in-domain out-of-vocabulary (**ID-OOV**) set, (ii) and an unseen language (**UL**) set. **ID-OOV** set denotes keywords that belong to one of the eight seen languages but not included during training. We choose the three

Table 5.1: Training and testing data size in number of words and samples for each language, including in-domain in-vocabulary (**ID-IV**), in-domain out-of-vocabulary (**ID-OOV**), and unseen languages (**UL**).

Set	Language	Family	# words	# train	# test
<b>ID-IV</b>	English (en)	Germanic	1755	1194255	150238
	German (de)	Germanic	602	403101	50799
	French (fr)	Romance	376	253913	31828
	Persian (fa)	Iranian	330	211875	26637
	Spanish (es)	Romance	229	149488	18729
	Russian (ru)	Slavic	97	62324	7857
	Italian (it)	Romance	78	50569	6358
	Polish (pl)	Slavic	62	39513	5047
<b>ID-OOV</b>	Russian (ru)	Slavic	30	-	1698
	Italian (it)	Romance	30	-	4481
	Polish (pl)	Slavic	30	-	2448
<b>UL</b>	Turkish (tr)	Turkic	30	12801	1620
	Latvian (lv)	Baltic	30	2845	389
	Lithuanian (lt)	Baltic	30	4137	330

in-domain languages that have the fewest training data, namely Russian, Italian, and Polish, and select 30 most frequent keywords that are not included in the **ID-IV** set as target keywords. The **UL** set, referring to languages not represented in the training set, includes three low-resource languages, Turkish, Latvian, and Lithuanian, which all belong to different language families than the eight **ID-IV** languages, since we want to consider more extreme scenarios.

While character sequences can be obtained from each keyword itself, non-aligned phoneme sequences are obtained by running the open-source tool `phonemizer`<sup>1</sup>. After processing all keywords and removing some of those containing ambiguous phoneme-based tokens, the resulting training and testing datasets, and the corresponding number of keywords and samples for each language are shown in Table 5.1. After obtaining the phoneme sequences, we then use the mapping relationship described in section 3.1 to transform each keyword into a sequence of attributes. We use the product space of manner and place of articulations as

<sup>1</sup><https://github.com/bootphon/phonemizer>

discussed in section 3.1. In our experiments, there are a total of 114 character tokens, 73 phoneme tokens, and only 33 attribute tokens.

We use a publicly available pre-trained self-supervised model as our encoder model for all experiments, namely Wav2Vec2.0 [82]. We choose the "base" architecture - 95M parameters leveraging 12 Transformer encoder layers, with embedding size of 768 and 8 attention heads. It is pre-trained on 960 hours of unlabeled audio from the LibriSpeech dataset [127]. The language classifier in DAT framework consists of 3 linear layers with a hidden dimension of 1024 and predicts eight in-domain language classes. AdamW optimizer [129] with  $\beta_1 = 0.9$  and  $\beta_2 = 0.98$  is used in training all models, and early stopping criterion is used to terminate the training phase.

## 5.4 Experimental Results and Discussions

### 5.4.1 DAT Results: Language Identification

Table 5.2 presents the language identification accuracy for systems trained with and without Domain Adversarial Training (DAT). In the absence of DAT, our models retain the same architecture depicted in Figure 5.1(b), except for the exclusion of the gradient reversal layer (GRL). The removal of the GRL implies that the feature extractor does not aim to mislead the language classifier. Thus, the feature extractor learns language-specific features to optimize recognition performance.

The results clearly show that DAT leads to a significant decrease in language identification accuracy, confirming the encoders' effectiveness in minimizing language-specific

Table 5.2: Language identification accuracy (%) with (w/) and without (w/o) DAT for characters, phonemes and attributes.

System ( $\downarrow$ )/Units ( $\rightarrow$ )	Accuracy (%)		
	Characters	Phonemes	Attributes (ours)
w/o DAT	91.35	91.10	90.47
w/ DAT	45.24	49.66	<b>34.10</b>

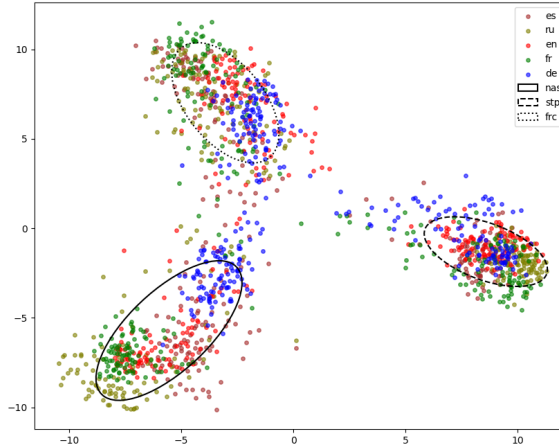


Figure 5.2: PCA visualization of latent features from the final hidden layer of  $\text{DAT}_{attr}$ , showing three manner-place articulation classes: nasal-bilabial (nas-blb), stop-alveolar (stp-alv), and fricative-labiodental (frc-lbd). Each cluster includes features from five languages.

information, and further strengthen our motivation in its use in out-of-domain scenarios. Furthermore, our attribute modeling approach, when combined with DAT, further reduces language identification accuracy. This suggests a stronger capability for feature universality, as characters and phonemes are inherently more language-dependent, making the extraction of language-invariant features more challenging.

#### 5.4.2 Analysis on Language-universal Latent Features

To effectively visualize and analyze the language-universality achieved through Domain Adversarial Training (DAT), we adopt a common technique known as Principal Component Analysis (PCA) [150]. PCA is a linear dimensionality reduction method that projects high-dimensional features into a two-dimensional space while preserving their relative geometric structure. The results are shown in Figure 5.2, which visualizes the latent features extracted from the Wav2Vec2 encoder output. We focus on three of the most common classes in the manner-place attribute product space: nasal-bilabial, stop-alveolar, and fricative-labiodental. The PCA plot reveals that features grouped by attribute class form

three distinct, compact clusters with clear separation between classes. In particular, each cluster contains features from five languages, namely English, German, French, Spanish, and Russian, which overlap significantly when they share the same attribute class. This strongly suggests that DAT successfully removes language-specific cues from the latent features. Overall, the result clearly demonstrates that speech attributes, when used in combination with DAT, organize the representation space in a way that is semantically meaningful and highly language-invariant.

### 5.4.3 Recognition Results

Three different sets of units, namely characters, phonemes and attributes, are modeled and compared using the architectures in Figure Figure 5.1. Although accuracy is commonly used in speech command classification, we report Word Error Rate (WER) to align with standard measurement in ASR for each system. The results of the Baseline training, denoted as  $\text{Base}_{char}$ ,  $\text{Base}_{phone}$ , and  $\text{Base}_{attr}$ , stands for training the systems from scratch on the training set. The results of DAT-based systems are denoted as  $\text{DAT}_{char}$ ,  $\text{DAT}_{phone}$ , and  $\text{DAT}_{attr}$  respectively.

Table 5.3 summarizes the WER for the eight in-domain in-vocabulary (**ID-IV**) languages for both Baseline and DAT settings. For Baseline, all systems, namely  $\text{Base}_{char}$ ,  $\text{Base}_{phone}$ , and  $\text{Base}_{attr}$  have similar results. However, when applying DAT to remove language-specific information and enforce language-invariant features, we observe a notable increase in WER for all systems ( $\text{DAT}_{char}$ ,  $\text{DAT}_{phone}$ , and  $\text{DAT}_{attr}$ ) compared to their Baseline counterparts. This drop in performance is expected because DAT reduces the models' reliance on language-specific cues that can be beneficial for recognizing speech in high-resource languages. By removing these cues, the systems have to rely solely on language-invariant features, which may not be as optimized for capturing language-specific variations. Despite this drop, we argue that since characters and phonemes are relative language-dependent during Baseline training, those system rely more on the language in-

Table 5.3: Testing WER (%) of the in-domain set on 8 rich-resource languages and the average (Avg.).

System	ID-IV								Avg.
	en	de	fr	fa	es	ru	it	pl	
Base <sub>char</sub>	13.14	12.57	13.65	14.77	11.89	15.50	<b>14.39</b>	15.00	13.86
Base <sub>phone</sub>	<b>12.73</b>	<b>11.70</b>	<b>13.29</b>	14.30	<b>11.84</b>	<b>12.96</b>	14.63	15.14	<b>13.32</b>
Base <sub>attr</sub> (ours)	13.28	12.04	13.69	<b>13.53</b>	12.20	13.67	15.13	<b>14.98</b>	13.56
DAT <sub>char</sub>	18.73	16.45	19.55	16.91	14.99	17.98	17.19	18.62	17.55
DAT <sub>phone</sub>	17.76	18.33	21.12	20.00	17.58	14.97	18.91	17.65	18.29
DAT <sub>attr</sub> (ours)	<b>15.51</b>	<b>13.74</b>	<b>16.27</b>	<b>15.47</b>	<b>13.85</b>	<b>13.83</b>	<b>16.73</b>	<b>15.75</b>	<b>15.14</b>

formation to perform recognition. In fact, after removing language information via DAT, our proposed framework, DAT<sub>attr</sub>, consistently outperforms DAT<sub>char</sub> and DAT<sub>phone</sub> for eight in-domain languages, with a relative word error rate (WER) reduction of 13.73%, and 17.22%, respectively. The results provide strong support for our argument that speech attributes are significantly more universal across languages than traditional tokens.

Table 5.4 and Table 5.5 show the WER for the out-of-domain scenarios, namely in-domain out-of-vocabulary (**ID OOV**) and unseen languages (**UL**). For OOV keywords, character-based system degrades significantly, due to the fact that some of the character tokens are not included during training. However, phoneme-based systems, namely Base<sub>phone</sub> and DAT<sub>phone</sub>, perform slightly better than Base<sub>attr</sub> and DAT<sub>attr</sub> but the improvements are not meaningful. We argue that since all the tokens are still covered in the phoneme set, both phoneme and attribute systems have similar results.

For unseen languages, our proposed framework Base<sub>attr</sub> consistency outperforms Base<sub>char</sub>, Base<sub>phone</sub> in all three languages, with relative WER reduction of 26.87% and 17.34% respectively. Comparing DAT and Baseline, we can see that DAT improves the robustness across languages of both phonemes and attributes systems, and slightly degrade the character system on average. However, the attribute-based system clearly benefits more from the language-invariant features generated by the DAT framework as we can see all three unseen languages yield meaningful improvements, while character- and phoneme-based

Table 5.4: Zero-shot transfer: Testing WER (%) of the 3 in-domain out-of-vocabulary (**ID OOV**) keywords from Russian, Italian, and Polish.

System	ID OOV			
	ru	it	pl	Avg.
Base <sub>char</sub>	63.96	44.50	40.77	49.74
Base <sub>phone</sub>	<b>31.15</b>	<b>40.62</b>	33.29	<b>35.02</b>
Base <sub>attr</sub> (ours)	31.57	41.89	<b>32.80</b>	35.42
<hr/>				
DAT <sub>char</sub>	54.89	40.19	37.95	44.34
DAT <sub>phone</sub>	<b>28.15</b>	<b>38.61</b>	34.11	<b>33.62</b>
DAT <sub>attr</sub> (ours)	29.98	40.28	<b>30.23</b>	33.81

Table 5.5: Zero-shot transfer: Testing WER (%) of the 3 unseen languages (**UL**), namely Turkish, Latvian, and Lithuanian.

System	UL			Avg.
	tr	lv	lt	
Base <sub>char</sub>	61.79	61.18	39.39	54.12
Base <sub>phone</sub>	54.50	45.50	43.64	47.88
Base <sub>attr</sub> (ours)	<b>48.33</b>	<b>40.10</b>	<b>30.30</b>	<b>39.58</b>
<hr/>				
DAT <sub>char</sub>	61.23	57.33	45.45	54.67
DAT <sub>phone</sub>	46.67	39.59	49.70	46.33
DAT <sub>attr</sub> (ours)	<b>42.96</b>	<b>36.25</b>	<b>26.36</b>	<b>37.10</b>

systems degrade on Lithuanian. Furthermore, a larger reduction on WER (from 39.58% to 37.10%) can be achieved. Comparing DAT framework on the three tokens, a similar pattern as Baseline can be observed where DAT<sub>attr</sub> consistently outperforms both DAT<sub>char</sub> and DAT<sub>phone</sub>, with relative WER reduction of 32.14% and 19.92%, respectively.

#### 5.4.4 Analysis of the Zero-shot Transfer Results

As shown in Table 5.5, the zero-shot transfer to unseen languages provides a challenging task for our ASR models, requiring them to generalize to entirely new languages without direct training. It is reasonable to assume that the acoustic and linguistic distance between the languages play a crucial role in affecting the performance. Numerous approaches exist for quantifying the distance between languages [151, 152, 153], each grounded in differ-

Table 5.6: Genetic proximity scores between unseen languages, Turkish (tr), Latvian (lv), and Lithuanian (lt), and the eight training languages. Avg. refers to the average score.

	en	de	fr	fa	es	ru	it	pl	Avg.
tr	96.3	87.8	85.4	76.6	82.5	75.3	79.6	77.1	81.5
lv	80.3	73.2	63.1	77.2	64.2	48.4	66.0	47.7	65.0
lt	70.0	60.4	42.4	66.3	44.1	45.7	40.2	41.5	51.3

ent linguistic dimensions. In this study, we adopt the genetic proximity<sup>2</sup> as our measure of linguistic distance. Language relationships are quantified using a probabilistic model that evaluates core vocabulary based on well-defined criteria. This approach enables the automatic classification of languages into families and subfamilies, while also producing objective distance metrics for cross-language comparisons. The distances are represented on a scale from 0 to 100, where a value of 0 indicates the closest linguistic relationship and 100 signifies the most distant. In Figure 5.3, we can draw a subtree of the entire language family tree based on the genetic scores for the eleven languages.

To analyze this relationship, we compare the WER results in Table 5.5 of different ASR systems against the genetic proximity scores presented in Table 5.6. These scores reflect the linguistic relatedness of the three unseen languages (Turkish, Latvian, Lithuanian) to the eight training languages (English, German, French, Persian, Spanish, Russian, Italian, Polish). In general, higher values indicate a greater linguistic distance from the training languages, meaning that languages with higher scores are less related to the ones seen during training. A lower score, in contrast, suggests closer linguistic ties, which may facilitate better generalization.

A key observation is that the degree of performance drop aligns with the genetic proximity of the target language to the training languages. Lithuanian (lt) achieves the best zero-shot performance, exhibiting the best WER among the unseen languages. This is expected as its lower proximity score of 51.3 indicates a closer relationship to the training

<sup>2</sup><http://www.elinguistics.net/>

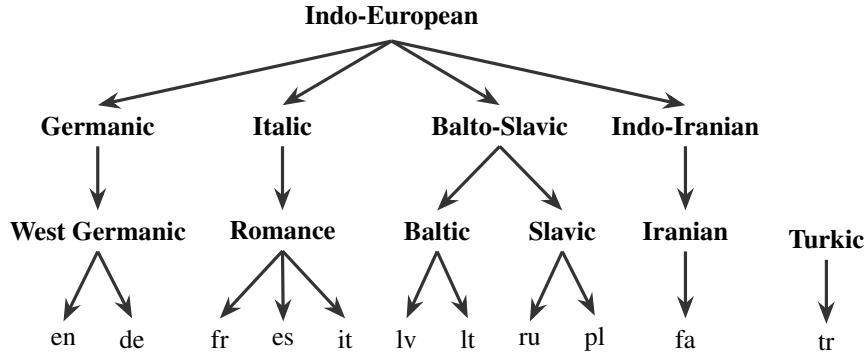


Figure 5.3: Illustration of a language family tree for 8 in-domain training languages (English, German, French, Persian, Spanish, Russian, Italian, Polish) and 3 unseen languages (Turkish, Latvian, Lithuanian).

languages, particularly to Slavic (Russian, Polish) and Germanic (German, English) languages. Latvian (lv) shows moderate recognition results with WER of 47.88% for  $\text{Base}_{attr}$  and 46.33% for  $\text{DAT}_{attr}$ , reflecting its intermediate genetic proximity of 65.0. While Latvian is more distant than Lithuanian, it still retains some phonetic similarities with the training set languages, particularly with Slavic and Germanic languages. Turkish (tr), on the other hand, suffers to generalize, correlating with its high genetic proximity score of 81.5, indicating it is the most distant from the training languages with almost no recognizable relationship. As shown in Figure 5.3, Turkish belongs to the Turkic language family, making it structurally and phonologically distinct from the Indo-European training languages. The lack of shared phonemes and different acoustic structures result in significant mismatch, particularly for character- and phoneme-based models, which struggle to adapt to Turkish’s unique grapheme and phonetic inventory.

#### 5.4.5 Challenges for Speech Attribute Modeling for Continuous Speech Recognition

In the previous sections, we demonstrated that modeling speech with language-universal attributes can achieve high accuracy in isolated word recognition under constrained conditions. A fixed, limited vocabulary and isolated word inputs help mitigate ambiguity, allowing attribute sequences to be matched reliably to the correct word. However, extending this

approach to continuous speech with a large vocabulary introduces significant challenge.

A fundamental limitation of using a compact set of speech attributes for recognizing continuous speech is decoding ambiguity [154]. Because attributes provide a coarse-grained representation, it is possible for different words to yield the same sequence of attributes. In a large vocabulary scenario, many distinct words may share an identical or very similar attribute sequence. For example, an English word and a phonetically similar foreign word could coincidentally map to the same pattern of broad articulatory features. The universality of the attribute inventory means it covers broad categories applicable to many languages, but this also means the distance between representations of different words is reduced, increasing the risk of confusion. In other words, two distinct words can end up nearly overlapping in the attribute space, making it ambiguous which word was actually spoken. This decoding ambiguity becomes more pronounced as the lexicon grows. In a multilingual context, the attribute sequence for a given utterance might correspond to multiple candidate words across different languages. The recognizer must then determine which word (or language) is correct, even though the low-level attribute evidence does not strongly favor one over the others.

## **5.5 Summary**

In this chapter, a novel system to build a language-universal multilingual SKR system leveraging pre-trained self-supervised models and speech attribute modeling has been presented. Our findings strongly confirm that speech attributes are a viable solution toward building multilingual speech recognition models. In fact, the proposed solution has the capability of generalizing toward out-of-domain scenarios, such as out-of-vocabulary keywords and unseen languages. We also observed significant word error rate reduction under zero-shot transfer to three low-resource languages, namely, Turkish, Latvian, and Lithuanian. Further analysis and visualization also demonstrate the language-universality and interpretability of the proposed attribute modeling approach.

## CHAPTER 6

# KEYWORD INFORMATION EXTRACTION THROUGH SPEECH ATTRIBUTE MODELING: FROM ISOLATED WORD RECOGNITION TO CONTINUOUS SPEECH RECOGNITION

### 6.1 Introduction

Keyword Information Extraction (KIE) [155, 18] is the task of detecting and localizing predefined keywords in continuous speech. It is often considered a specialized application of automatic speech recognition (ASR), with a primary focus on identifying specific keywords within an utterance rather than fully transcribing speech. The detection component determines whether the keyword appears in the utterance, while the localization component precisely identifies the start and end time boundaries of the detected keyword. Given a speech utterance  $O$  and a text-based keyword query  $q$ , a KIE system aims to determine whether  $q$  appears in  $O$  and, if so, estimate its temporal boundaries. Existing approaches for KIE can be broadly categorized into three paradigms: keyword-filler-based systems [156, 157], ASR-based systems [158, 159, 160], and end-to-end systems [161].

Among these approaches, ASR-based keyword information extraction is one of the most intuitive and widely adopted methods. It generally follows a two-stage framework: the first stage consists of an ASR model that converts speech into text transcriptions, while the second stage performs keyword search over the transcriptions. The advantage of this approach is its flexibility, as it allows keyword search to be conducted on any arbitrary keyword rather than being limited to a predefined set. This makes ASR-based KIE systems significantly more scalable compared to conventional keyword-filler approaches. Furthermore, because ASR-based KIE shares the same underlying framework as ASR, most modern speech recognition technique can be directly applied to enhance keyword extraction.

When trained on sufficiently large datasets, ASR-based systems tend to yield superior performance compared to earlier keyword-filler based methods [162].

## 6.2 A Two-stage System for KIE: Attribute-based ASR and Utterance Verification

Figure 6.1 shows the overall framework of our proposed two-stage ASR-based system for KIE. As we can see from the framework, the first-stage mainly consists of a conventional ASR system, aiming to predict the most likely sequence of tokens  $W^*$  from a given utterance  $O$ . The keyword search occurs after the first-stage system produces the best possible sequence of tokens  $W^*$ . By passing the desired keyword  $q$  and its corresponding token sequence, the system then output the first-pass detection results with several putative hits. A second-stage utterance verification is adopted to further refine the detection results and remove unwanted false alarms.

### 6.2.1 Attribute-based ASR

For KIE, characterizing speech recognition systems based on articulatory attributes provides significant advantages over high-level representations such as phonemes and characters. As pointed out in chapter 3, attributes form a compact and universal set of speech units, ensuring more balanced data distribution across tokens. This compactness allows for higher training efficiency, as each attribute token benefits from a larger pool of training data, leading to improved recognition accuracy at the token level. In the context of KIE,

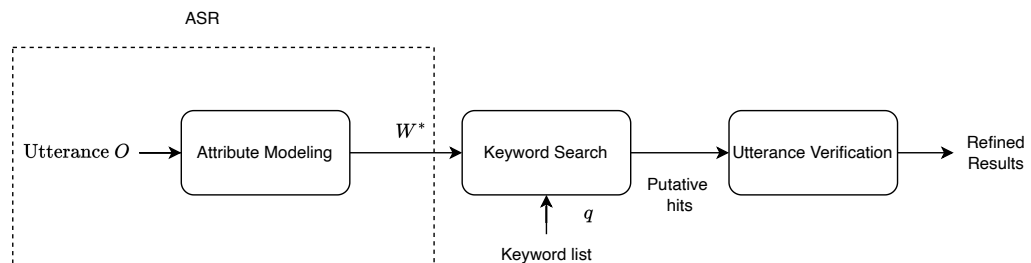


Figure 6.1: Overview of the attribute-based ASR system for KIE

where the goal is to detect a predefined set of keywords, using a well-defined, compact set of attributes is sufficient for distinguishing each keyword without relying on complex phoneme- or character-based representations. Given that first-pass keyword detection relies on the accuracy of attribute recognition, a more robust attribute-based system should directly contribute to a higher detection rate in the first-stage ASR.

However, defining the optimal set of attributes presents a trade-off. A highly compact attribute set can improve recognition accuracy, boosting first-pass detection rates. Yet, if the attribute set is overly compact, it may increase the likelihood of false alarms, extracting more unwanted putative hits and overloading the second-stage verification system. Therefore, balancing detection rate and false alarm rate is critical when selecting the appropriate number of attributes to model. An overly fine-grained attribute set may introduce unnecessary distinctions, while an overly coarse-grained set may fail to adequately differentiate keywords. Striking the right balance is essential to achieving efficient and accurate keyword extraction in the two-stage KIE framework. Empirically, we find that the attribute set defined in Table 4.1 achieves an ideal balance between maximizing detection rate and minimizing false alarms.

In addition to the attributes, we introduced a special token '|' between words to indicate word boundary as suggested in [163]. This is because we observe that false alarms occur a lot when some shorter are substrings in some longer non-keywords. For example, the keyword 'anders' is actually contained in the non-keyword 'commanders'. Therefore, by introducing the word boundary token, we ensure that only when a putative hit is within two word boundaries can it be extracted for second-stage verification. Empirically, we observe a significantly reduction in false alarms in the first stage.

### 6.2.2 Utterance Verification

Utterance verification is a second-stage validation step in two-stage ASR-based KIE systems, performed after the initial detection stage to confirm candidate keywords and filter

out false alarms. By rescoreing each candidate occurrence with additional models or confidence metrics, this verification stage helps eliminate incorrect detections (false alarms) and thus improves the precision of keyword extraction. Common rescoreing and verification methods are HMM [164, 165, 156], logistic regression [166, 167], SVM [168, 169], and DNN [170, 155].

In our proposed two-stage system as illustrated in Figure 6.1, two methods will be introduced in the following sections, namely, a phoneme-based vocabulary-independent utterance verification system and a DNN-based utterance verification system.

### **6.3 Phoneme-based Vocabulary-Independent Utterance Verification (VIUV)**

#### 6.3.1 Subword-based Utterance Verification

Utterance verification (UV) [165] can be formulated as a post-recognition statistical hypothesis test to decide whether the input speech indeed contains the hypothesized keyword or not. In a subword-based recognition framework, achieving vocabulary independence normally requires a two-stage process. First, a subword-level verification is performed for each subword in the hypothesized keyword’s transcription. This yields a series of independent hypothesis tests that examine whether each detected subword truly matches the intended sound. Crucially, no retraining of the verifier is needed when the vocabulary is changed or updated, since the verification operates at the subword level and is thus vocabulary-independent. If the first-stage attribute-based ASR system subsection 6.2.1 were to produce perfectly accurate attribute transcriptions, then any false alarm would occur only in cases where a non-keyword utterance shares the exact same sequence of attribute tokens as the target keyword. Therefore, an additional verification step at units with finer granularity is required to effectively eliminate such false alarms. In this chapter, we perform the vocabulary-independent utterance verification (VIUV) in the *phoneme*-level and thus would refer to the system as phoneme-based vocabulary-independent utterance verification in the following sections.

Consider the lexical phoneme sequence  $p_1^{(k)} p_2^{(k)} \dots p_N^{(k)}$  corresponding to the keyword  $W_k$ . For each phoneme  $p_j^{(k)}$  in  $W_k$ , we perform a hypothesis test to verify if the acoustic segment  $\mathbf{O}_j$ , which is the portion of the input speech aligned to  $p_j^{(k)}$  by the first-stage recognizer truly corresponds to that phoneme. The null hypothesis  $H_0$  represented by the input  $\mathbf{O}_j$  containing a given sound  $p_j^{(k)}$  is tested against the alternative hypothesis  $H_1$  which is represented by the input segment containing a different sound. To simplify the notation and without loss of generality, we will drop the superscript  $(k)$  from  $p_j^{(k)}$ . The likelihood ratio for this phoneme-level test is then formulated as:

$$\Lambda(\mathbf{O}_j; p_j) = \frac{L(\mathbf{O}_j | H_0)}{L(\mathbf{O}_j | H_1)} \quad (6.1)$$

Taking the logarithm of the ratio in Equation 6.1 yields the log-likelihood difference, which can be written as:

$$G(\mathbf{O}_j; p_j) = \log[L(\mathbf{O}_j | H_0)] - \log[L(\mathbf{O}_j | H_1)], \quad 1 \leq j \leq N \quad (6.2)$$

This quantity  $G(\mathbf{O}_j; p_j)$  represents the degree to which the segment  $\mathbf{O}_j$  is more likely under the target phoneme model than under all other possible phoneme models. In practice, the exact distributions for  $L(\mathbf{O}_j | H_0)$  and  $L(\mathbf{O}_j | H_1)$  are unknown, so we seek an approximate verification function to evaluate Equation 6.2. We can interpret the two terms in  $G(\mathbf{O}_j; p_j)$  as follows: the first term  $\log[L(\mathbf{O}_j | H_0)]$  is the log-likelihood of the correct phoneme model  $p_j$ , and the second term  $\log[L(\mathbf{O}_j | H_1)]$  represents the log-likelihood of an anti-phoneme model, i.e. a model representing “ $\mathbf{O}_j$  is not  $p_j$ .” We estimate the anti-phoneme likelihood by combining the contributions from all  $K - 1$  alternative phoneme models. In particular, a phoneme-based verification function  $V_j(\mathbf{O}_j; p_j)$  is defined to ap-

proximate the log-likelihood difference in Equation 6.2 as:

$$V_j(\mathbf{O}_j; p_j) = \log [L(\mathbf{O}_j | p_j)] - \log \left[ \frac{1}{K-1} \sum_{\substack{k=1 \\ k \neq j}}^k \exp(\gamma \log [L(\mathbf{O}_j | p_k)]) \right]^{1/\gamma} \quad (6.3)$$

Here,  $L(\mathbf{O}_j | p)$  represents the HMM likelihood of segment  $\mathbf{O}_j$  given the phoneme model  $p$ , and  $\gamma$  is a positive scaling factor. Consequently, the first term in Equation 6.3 serves as an estimate for  $L(\mathbf{O}_j | H_0)$ , while the second term approximates  $L(\mathbf{O}_j | H_1)$  by computing the geometric mean of the likelihoods from all alternative phoneme models. In practice, when  $\mathbf{O}_j$  truly corresponds to phoneme  $p_j$ , the likelihood  $L(\mathbf{O}_j | p_j)$  will typically be among the largest, while most other  $L(\mathbf{O}_j | p_k), k \neq j$  tend to be much smaller, resulting in a relatively small second term and making  $V_j(\mathbf{O}_j; p_j)$  positive. In contrast, if  $\mathbf{O}_j$  does not match the sound of  $p_j$ , one of the competing phoneme models  $p_k$  will likely produce a higher likelihood, increasing the second term and causing  $V_j(\mathbf{O}_j; p_j)$  to be negative overall. The parameter  $\gamma$  adjusts how strongly competing models influence the score. That is, setting  $\gamma$  to a large value causes the sum to be dominated by the highest competing likelihood, effectively focusing on the strongest anti-phoneme, whereas a smaller  $\gamma$  allows more weight to the contributions of multiple competing models. In practice,  $\gamma$  is tuned to achieve the best overall verification performance for the entire set of phonemes.

The verification function  $V_j(\mathbf{O}_j; p_j)$  in Equation 6.3 provides a discriminative score for each phoneme occurrence, but to compute it requires evaluating  $L(\mathbf{O}_j | p_k)$  for every phoneme model  $p_k$  in the entire set of  $K$  models. This process must be applied to every phoneme segment in the recognized sequence following the primary recognition step. However, calculating these  $K$  likelihoods for each segment can be computationally demanding and may cause substantial delays in generating the verification outcome.

### 6.3.2 Attribute-specific Cohort Set

To reduce the computational complexity of Equation 6.3 while maintaining verification effectiveness, we introduce a cohort set to approximate the anti-phoneme term. The use of cohort models in this context is similar to the speaker-specific cohort technique in speaker verification systems [171, 172, 173] where a small cohort of impostor speaker models is used to normalize scores in a likelihood-ratio test. The key idea is to replace the full sum over all  $K - 1$  models with a sum over a much smaller subset of competing phoneme models that are most likely to be confused with the target phoneme  $p_j$ . For each phoneme  $p_j$ , we pre-define a cohort  $\mathcal{C}(p_j)$  consisting of the  $M$  phoneme models that are most confusable with  $p_j$ . In other words,  $\mathcal{C}(p_j)$  is a fixed list of  $M \ll K$  phonemes that have the highest likelihoods of generating segments intended as  $p_j$ . Unlike [165] where the cohort set is empirically determined from training data by analyzing the confusion matrix of the phoneme recognizer, we use the inherent knowledge of speech attributes to construct the cohort set. As shown in Table 6.1, we simply group all phonemes in addition to  $p_j$  under the same attribute feature. For example, when considering manner of articulation, the phoneme /n/ is classified as a nasal. Therefore, its cohort set consists of other nasal phonemes present in the dataset, such as /m/ and /ŋ/, which share the same articulatory feature. Similarly, when considering the place of articulation, the phoneme /ð/ is categorized as dental, so its cohort set includes only the competing phoneme /θ/ with the same place feature. In practice, cohort sets can be defined based on either manner or place of articulation, or the combination of both, to perform utterance verification.

By considering only a fixed number  $M$  of cohorts for each phoneme, we can redefine the phoneme verification function in Equation 6.3 as:

$$V_j(\mathbf{O}_j; p_j) = \log [L(\mathbf{O}_j | p_j)] - \log \left[ \frac{1}{M} \sum_{m=1}^M \exp(\gamma \log [L(\mathbf{O}_j | p_{c_j(m)})]) \right]^{1/\gamma} \quad (6.4)$$

where  $p_{c_j(m)} \in \mathcal{C}(p_j)$  denotes the  $m$ -th phoneme in the cohort set for  $p_j$ , and  $M = |\mathcal{C}(p_j)|$  is

Table 6.1: Examples of phoneme cohort sets based on speech attributes.

Category	Attribute	Phoneme	Cohort Set
Manner (M)	nasal	n	m ŋ
	stop	p	b t d k ɣ
	fricative	s	z f h ʃ ð θ ʒ
Place (P)	dental	ð	θ
	alveolar	t	d s z l r ɹ
	palato-alveolar	ʒ	ʃ tʃ dʒ

the number of cohort models. In this cohort-based verification function  $V_j$ , the second term now uses the geometric mean of the likelihoods of only the  $M$  most confusable phonemes, rather than all  $K - 1$  anti-phoneme models. This drastically lowers the computation, since  $M$  is small and typically on the order of a few units compared to the full phoneme inventory. The cohort-based method preserves the essence of the likelihood ratio test by focusing on the most likely alternatives to the target phoneme. It effectively normalizes the target phoneme’s score by the scores of its closest “competitors.”

### 6.3.3 String-Level Utterance Verification

The string-level utterance verification stage makes the final decision to accept or reject a hypothesized keyword based on the cumulative evidence from each of its phoneme-level segments. This stage follows the first-pass recognition and phoneme-level verification, integrating the per-phoneme verification scores into a global decision score.

Assuming phoneme-level independence, the overall likelihood ratio for a keyword  $W_k$  is computed as the product of phoneme-level likelihood ratios:

$$\Lambda(\mathbf{O}; W_k) = \prod_{j=1}^N \Lambda(\mathbf{O}_j; p_j) \quad (6.5)$$

where  $N$  is the number of phonemes in the lexical representation of  $W_k$ , and  $\mathbf{O}_j$  is the acoustic segment aligned with the phoneme  $p_j$ . Taking the logarithm of Equation 6.5 yields

the overall log-likelihood difference:

$$G(\mathbf{O}; W_k) = \sum_{j=1}^N [\log L(\mathbf{O}_j | H_0) - \log L(\mathbf{O}_j | H_1)] \quad (6.6)$$

Here,  $L(\mathbf{O}_j | H_0)$  and  $L(\mathbf{O}_j | H_1)$  denote the likelihoods under the target phoneme model and the anti-phoneme model, respectively. These two likelihoods are approximated by the verification function  $V_j(\mathbf{O}_j; p_j)$  defined earlier in Equation 6.4. Substituting these into Equation 6.6, the string-level score can be rewritten in terms of phoneme-level discriminative scores as:

$$G(\mathbf{O}; W_k) = \sum_{j=1}^N V_j(\mathbf{O}_j; p_j) \quad (6.7)$$

This form reflects how each phoneme-level verification score contributes to the overall keyword decision. In practice, this average score serves as the basis for a binary decision. A fixed decision threshold  $\tau$  is applied to determine whether the putative hit is accepted as a valid keyword occurrence:

$$\hat{y} = \begin{cases} 1 & \text{if } G(\mathbf{O}; W_k) \geq \tau \\ 0 & \text{otherwise} \end{cases} \quad (6.8)$$

where  $\hat{y} = 1$  denotes acceptance and  $\hat{y} = 0$  denotes rejection of the keyword hypothesis. The threshold  $\tau$  is tuned empirically on a development set to balance missed detections and false alarms.

#### 6.4 DNN-based Utterance Verification: Acoustically Grounded Word Embeddings

For keywords with plenty of training samples, more complex and discriminative models can be used in addition to the HMM-GMM models for further performance improvement [155]. Usually, keyword-specific DNN verifiers are trained for each keyword, which is not practical once we have more keywords. In addition, for keywords that do not have enough

training samples, such verifiers can not be properly trained. Therefore, it is more favorable to have a DNN-based verification system that can distinguish keywords from non-keywords and simultaneously handle low-resource and unseen keywords during inference time.

Acoustically grounded word embeddings (AGWE) [174, 175, 176] is a type of word embedding that incorporates the acoustic properties of spoken language into the representation of words. Traditional word embeddings, such as Word2Vec [177] or GloVe [178], rely on large text corpora to learn the semantic relationships between words based purely on their co-occurrence in text. However, they do not consider how words sound, which can be a significant limitation in applications involving spoken language [179, 180]. AGWE addresses this gap by integrating acoustic signals, such as phonetic features or audio recordings of speech, into the word representations. This grounding in actual speech sounds provides a richer and more nuanced understanding of words, encompassing both their semantic and phonetic properties.

AGWEs are learned using models that encode both acoustic and phonetic information. Specifically, we use a dual-encoder framework: one encoder processes acoustic sequences (spoken word segments) and produces acoustic embeddings  $f(x)$ , while the other encoder maps phoneme sequences  $p$  to corresponding phonetic embeddings  $g(p)$ . The goal is to train these encoders such that acoustic and phonetic representations of the same word are mapped close together in the shared embedding space, while representations of different words remain far apart.

To measure the similarity between embeddings, we use the cosine distance:

$$d(x, y) = 1 - \frac{x \cdot y}{|x||y|} \quad (6.9)$$

Let  $(x_i, p_i)$  denote the  $i$ -th training pair, where  $x_i$  is the acoustic segment and  $p_i$  is its corresponding phoneme sequence. We adopt a multi-view contrastive loss (adapted from [181]) to jointly train the acoustic and phonetic encoders. The loss consists of two symmetric

terms:

$$\begin{aligned} \min_{f,g} \sum_{i=1}^N & \left[ m + d(f(x_i), g(p_i)) - \min_{p \neq p_i} d(f(x_i), g(p)) \right]_+ \\ & + \sum_{i=1}^N \left[ m + d(g(p_i), f(x_i)) - \min_{\text{phn}(x) \neq p_i} d(g(p_i), f(x)) \right]_+ \end{aligned} \quad (6.10)$$

The first term encourages the acoustic embedding  $f(x_i)$  to be close to its matching phonetic embedding  $g(p_i)$ , and far from the embeddings of mismatched phoneme sequences. The second term does the reverse: it pushes  $g(p_i)$  close to  $f(x_i)$ , and far from non-matching acoustic segments. The margin  $m$  defines how far apart non-matching pairs should be, and  $\text{phn}(x)$  is the phoneme sequence corresponding to the word label of the acoustic sequence  $x$ .

In practice, computing the minimum distance over all mismatched pairs  $p \neq p_i$  and  $\text{phn}(x) \neq p_i$  is computationally expensive. Therefore, we follow the strategy from [182], selecting the  $k$  most offending negative examples, i.e., those closest in distance, from within each mini-batch. We then compute the mean distance over these hard negatives to approximate the minimum.

## 6.5 Datasets and Experimental Settings

### 6.5.1 Performance Measurement for Keyword Information Extraction

To evaluate the effectiveness of a KIE system, we assess its ability to detect and temporally localize keywords within a test corpus. For each target keyword  $q$ , the system outputs a list of hypothesized keyword occurrences, where each instance is represented as a 3-tuple:

$$O_{q,i}^p = (u_i^p, s_i^p, e_i^p)$$

where  $u_i^p$  denotes the utterance identifier, while  $s_i^p$  and  $e_i^p$  represent the predicted start and end times, respectively, of the  $i$ -th occurrence of keyword  $q$ . The full list of predicted

occurrences for keyword  $q$  is denoted as:

$$\mathbf{O}_q^p = \{O_{q,j}^p\}$$

which is evaluated against a reference list of true keyword instances:

$$\mathbf{O}_q^r = \{O_{q,j}^r\}$$

A putative occurrence  $O_{q,i}^p$  is considered a correct match, or a *hit*, if the midpoint of its temporal span lies within a predefined tolerance window around a corresponding ground-truth occurrence. Specifically, we consider a hit valid if its midpoint falls within a time window that begins  $\Delta T$  seconds before the reference start time and ends  $\Delta T$  seconds after the reference end time. In our experiments, we set  $\Delta T = 0.1$  seconds.

To quantify system performance, we employ the Missed Detection Rate (MDR) and the False Alarm Rate (FAR). The MDR measures the proportion of ground-truth occurrences that are not correctly detected, while the FAR indicates the proportion of predicted detections that do not correspond to any true occurrence.

For a keyword,  $q$ , the MDR is defined as:

$$MDR(q) = N_{Miss}(q)/N_{True}(q) \quad (6.11)$$

where  $N_{Miss}(q)$  is the number of missed putative occurrences and  $N_{True}(q)$  is the total number of reference occurrences of the keyword  $q$ . The overall MDR of a KIE system is thus defined as the average MDR across all keywords:

$$MDR = \frac{1}{K} \sum_{q=1}^K P_{Miss}(q) = \frac{1}{K} \sum_{q=1}^K \frac{N_{Miss}(q)}{N_{True}(q)} \quad (6.12)$$

Similarly, the FAR is defined as:

$$FAR(q) = N_{FA}(q)/N_{NT}(q) \quad (6.13)$$

where  $N_{FA}(q)$  is the number of false alarms and  $N_{NT}(q)$  is the number of non-target trials for the keyword  $q$ . In our baseline setup, we assume one trial per second of speech, so the number of non-target trials is computed as:

$$N_{NT}(q) = T - N_{True}(q) \quad (6.14)$$

where  $T$  total duration of the test set in seconds and also equals to the total number of trials of the test dataset. We can then define the overall false alarm rate as the average false alarm rate of all keywords:

$$FAR = \frac{1}{K} \sum_{q=1}^K P_{FA}(q) = \frac{1}{K} \sum_{q=1}^K \frac{N_{FA}(q)}{N_{NT}(q)} = \frac{1}{K} \sum_{q=1}^K \frac{N_{FA}(q)}{T - N_{True}(q)} \quad (6.15)$$

In addition to MDR and FAR, we adopt the Actual Term Weighted Value (ATWV) as a comprehensive metric to jointly assess detection quality and reliability. Introduced by NIST for the spoken term detection task [183], ATWV is defined as:

$$\begin{aligned} ATWV &= 1 - P_{Miss} - \beta \cdot P_{FA} \\ &= 1 - \frac{1}{K} \sum_{q=1}^K \left( \frac{N_{Miss}(q)}{N_{True}(q)} + \beta \frac{N_{FA}(q)}{T - N_{True}(q)} \right) \end{aligned} \quad (6.16)$$

where  $\beta$  is a scaling parameter that adjusts the weight of false alarms in relation to missed detections. In our evaluation, we follow the standard convention and set  $\beta = 999.9$ . A perfect KIE system achieves an ATWV of 1, indicating zero missed detections and no false alarms. Higher ATWV values therefore correspond to better system performance.

Table 6.2: 13 keywords (in lowercase) selected in this study. Keywords are divided into three categories according to the number of training instances in the dataset.

Category	Keywords	# train
Low-resource	missouri	1
	baltimore	3
	jesus christ	8
	excellency	11
	alexander	15
	surrender	18
Unseen	hawkeye	0
	anders	
	edison	
	rodolfo	
	galatians	
	battleaxe	
	ferdinando	

### 6.5.2 Datasets and Keywords

We use the subset of LibriSpeech [127] corpus, involving 100 hours of clean speech audio as our training dataset. The test data are selected from the *'test-clean'* subset. The total number of testing data is 2620. To obtain the reference time boundary for each keyword, we need to align each utterance with corresponding transcription. Since the original LibriSpeech corpus does not provide word-level alignment information, the Montreal Forced Aligner (MFA) [184] is used to produce the ground-truth word-level alignments for each audio file in LibriSpeech. MFA is built on top of Kaldi [25], which mitigates run time when using larger corpora and more computationally intensive training. The ASR pipeline that MFA implements uses a standard HMM-GMM architecture, adapted from existing Kaldi recipes.

A total of thirteen keywords in the LibriSpeech vocabulary are selected in the experiments. The selected keywords were further divided into two categories based on 100 hours of training data: (i) words having less than 10 training data in the training set (6 words), (ii) words having no speech data in the training corpus (7 words). We refer to the first group as

low-resource keywords, and the second unseen keywords. Table 6.2 lists all the keywords in the two groups.

### 6.5.3 Models and Parameters

For the first-stage ASR, we use a publicly available pre-trained self-supervised model as our encoder model for all experiments, namely Wav2Vec2.0 [82]. We choose the "base" architecture - 95M parameters leveraging 12 Transformer encoder layers, with embedding size of 768 and 8 attention heads. It is pre-trained on 960 hours of unlabeled audio from LibriSpeech dataset [127]. AdamW optimizer [129] with  $\beta_1 = 0.9$  and  $\beta_2 = 0.98$  is used in training all models, and early stopping criterion is used to terminate the training phase. During training, the first-stage ASR system in the KIE framework is optimized using the CTC criterion. To obtain the temporal boundaries of each keyword, we apply greedy decoding to align each frame with its corresponding output token. The special blank token in CTC is used to identify frame regions that do not correspond to any token, allowing us to extract contiguous segments aligned to the keyword's token sequence. These segments are then treated as putative keyword occurrences for downstream utterance verification.

For the second-stage utterance verification, the phoneme-based Vocabulary-Independent Utterance Verification (VIUV) system is trained following the standard Kaldi [25] recipe. Each phoneme is represented using a three-state left-to-right continuous-density HMM, where each state is modeled by a GMM with diagonal covariance matrices to estimate the acoustic feature distributions. During inference, Viterbi decoding [185, 186] is employed to compute the most likely phoneme sequence and determine whether the putative keyword is present in the input utterance. We set  $\gamma = 0.5$  in Equation 6.3.

For the AGWE model, we follow the design described in [175]. The acoustic encoder is a 6-layer bidirectional LSTM (BLSTM), with each layer containing 512 hidden units per direction. It takes full utterances as input, and for each word segment within the utterance, the hidden states corresponding to the segment frames are averaged to produce

a fixed 1024-dimensional acoustic embedding. The phonetic encoder comprises an initial embedding layer that maps each of the 47 phonemes to a 64-dimensional vector, followed by a 1-layer BLSTM with 512 hidden units per direction. Both the acoustic and phoneme embeddings are then projected into a shared 256-dimensional embedding space via a linear transformation. During training, we apply random perturbation to the start and end times of sampled word segments within a range of  $\pm 0.1$  second to simulate misaligned boundaries. This improves generalization by making the model less sensitive to segmentation errors. At inference time, cosine distance is used to measure the similarity between a putative acoustic embedding and the corresponding phoneme-based embedding, and a fixed threshold is selected empirically to determine whether to accept or reject a match. This approach enables robust verification of putative keyword hit while maintaining tolerance to minor temporal mis-alignments.

## 6.6 Experimental Results and Discussions

Table 6.3 and Table 6.4 highlight the effectiveness of the proposed two-stage attribute-based KIE system, particularly under low-resource and unseen keyword conditions. For second-stage verification, three types of cohort-based VIUV as mentioned in subsection 6.3.2 are evaluated:  $VIUV_M$ ,  $VIUV_P$ , and  $VIUV_{M+P}$ . These variants differ in how they construct cohort sets for anti-phoneme modeling, where  $VIUV_M$  groups competing phonemes based on shared manner of articulation,  $VIUV_P$  uses place of articulation, and  $VIUV_{M+P}$  combines both to form a more restrictive and confusable cohort set. Additionally, we evaluate a DNN-based utterance verification using AGWE, which learns a joint embedding space between acoustic segments and phoneme sequences. It is worth noting that no verification was applied to the character or phoneme systems, as their false alarm rates were inherently negligible.

In the low-resource keyword scenario, the attribute-based system demonstrates a clear advantage in recall over both character- and phoneme-based baselines. As shown in Ta-

Table 6.3: Testing MDR, FAR, and ATWV of the low-resource keywords.

<b>System</b>	<b>Verification</b>	<b>MDR (%) ↓</b>	<b>FAR (<math>\times 10^{-3}\%</math>) ↓</b>	<b>ATWV ↑</b>
Character	-	41.54	0.0	0.5846
Phoneme	-	20.00	0.0	0.8000
	-	0.0	17.50	0.8250
	VIUV <sub>M</sub>	0.0	12.47	0.8753
Attribute	VIUV <sub>P</sub>	0.0	10.13	0.8987
	VIUV <sub>M+P</sub>	0.0	7.32	0.9268
	AGWE	0.0	<b>3.09</b>	<b>0.9691</b>

ble 6.3, while character- and phoneme-based systems yield Missed Detection Rates (MDR) of 41.54% and 20.00%, respectively, the attribute-based system achieves a perfect MDR of 0%. This suggests that modeling speech at the articulatory attribute level enables better accuracy in recognition, even when training data for individual keywords is scarce. However, the gain in recall comes with a trade-off in precision which is reflected by a higher false alarm rate of  $17.5 \times 10^{-3}\%$ . By introducing utterance verification, we observe consistent reductions in false alarms with phoneme-based VIUV. Specifically, VIUV<sub>M</sub> reduces FAR to  $12.47 \times 10^{-3}\%$ , VIUV<sub>P</sub> to  $10.13 \times 10^{-3}\%$ , and VIUV<sub>M+P</sub> to  $7.32 \times 10^{-3}\%$ , with corresponding increases in ATWV. The AGWE achieves an FAR of just  $3.09 \times 10^{-3}\%$  and the best ATWV of 0.9691, outperforming all systems. These results confirm that the attribute-based system, when equipped with verification, is highly effective in extracting low-resource keywords.

In the more challenging unseen keyword scenario, where none of the target keywords appear in the training set, the benefit of attribute-based modeling becomes even more obvious. As shown in Table 6.4, the character- and phoneme-based systems suffer from high missed detection rates of 70.35% and 69.14%, respectively, and yield very low ATWV scores of 0.2892 and 0.3010. In contrast, the attribute-based system achieves a significantly lower MDR of 27.07% without verification, demonstrating its ability to generalize to entirely novel keywords. However, similar patterns of higher FAR can be observed as in low-resource keywords results, which suppresses the overall ATWV to 0.4721. Integrat-

Table 6.4: Testing MDR, FAR, and ATWV of the unseen keywords.

<b>System</b>	<b>Verification</b>	MDR (%) ↓	FAR ( $\times 10^{-3}\%$ ) ↓	ATWV ↑
Character	-	70.35	0.73	0.2892
Phoneme	-	69.14	0.73	0.3013
	-	27.07	25.72	0.4721
	VIUV <sub>M</sub>	27.63	12.06	0.6031
Attribute	VIUV <sub>P</sub>	27.99	9.63	0.6238
	VIUV <sub>M+P</sub>	<b>28.12</b>	<b>3.36</b>	<b>0.6852</b>
	AGWE	28.66	4.41	0.6693

ing utterance verification into the attribute-based pipeline proves essential in this setting. Notably, VIUV<sub>M+P</sub>, which leverages both manner and place of articulation to form cohort sets, achieves the best performance by reducing FAR to  $3.36 \times 10^{-3}\%$  and achieving the highest ATWV of 0.6852. Unlike low-resource keywords, this surpasses the AGWE-based verification, which attains an ATWV of 0.6693. We contend that this results may stem from the fact that VIUV is entirely vocabulary-independent, while AGWE relies on phoneme-embedding pairs derived from the training set. As a result, AGWE may generalize less effectively to words whose phonemic patterns are unseen during training. These findings highlight the robustness of our VIUV approach for zero-shot and unseen scenarios, especially when attribute-based modeling is paired with carefully designed cohort set for verification.

## 6.7 Summary

In this chapter, we proposed a two-stage keyword information extraction (KIE) framework centered on attribute-based automatic speech recognition (ASR) followed by phoneme-level utterance verification. The first stage leverages articulatory attributes to model speech with a compact and universal token set, enabling accurate detection and localization of keywords, especially in low-resource and zero-shot settings. To address the higher false alarm rates associated with attribute-based decoding, we introduced a second-stage utterance veri-

fication module that operates on finer phoneme-level units. Two verification strategies were presented: a phoneme-based vocabulary-independent utterance verification (VIUV) system that utilizes a novel attribute-specific cohort definition to approximate anti-phoneme likelihoods, and a DNN-based verifier using acoustically grounded word embeddings (AGWE) to evaluate acoustic-phonetic consistency. Both methods significantly reduce false alarms and enhance detection precision, with VIUV demonstrating superior generalization for unseen keywords and AGWE excelling in low-resource conditions. This robust and scalable framework lays a strong foundation for extending attribute-based modeling to broader continuous speech recognition tasks.

## CHAPTER 7

### A BOTTOM-UP FRAMEWORK FOR CONTINUOUS SPEECH RECOGNITION IN SYLLABLE-BASED LANGUAGES

In this chapter, we propose a bottom-up framework for continuous speech recognition in syllable-based languages by unifying articulatory attribute modeling with syllable-level prediction. The system first recognizes sequences or lattices of articulatory attributes that serve as a language-universal, interpretable representation of pronunciation, and then transforms them into syllables through a structured mapping procedure. We introduce two evaluation metrics, namely Pronunciation Error Rate (PrER) and Syllable Clause Error Rate (SCER), to evaluate the model’s ability to capture pronunciation and handle syllable ambiguities. Experimental results on the AISHELL-1 Mandarin corpus demonstrate that the proposed bottom-up framework achieves competitive performance and exhibits better robustness under low-resource conditions compared to direct syllable prediction. Furthermore, we investigate the cross-lingual transferability and demonstrate promising results to extend the proposed framework to other syllabic languages, such as Japanese.

#### 7.1 Syllable-based Languages and Speech Recognition

In languages, such as English, character or subword units are commonly adopted as modeling targets for speech recognition, as these units generally align well with phonological structure. However, in many other languages, particularly those with logographic or syllabic writing systems, characters function primarily as semantic or orthographic symbols and lack a direct correspondence to pronunciation. Consequently, character-based modeling in these languages fails to incorporate phonological information into the acoustic model. Phoneme-based approaches, while theoretically grounded in phonology, also encounter limitations, particularly in languages with strong supra-segmental features, where

they struggle to capture tonal and prosodic variations critical for accurate recognition. Given these limitations, syllables offer a linguistically coherent and acoustically meaningful alternative for speech recognition in such languages, often referred to as syllable-based languages. Both Mandarin Chinese and Japanese belong to this category. In Mandarin, each syllable corresponds to a stable phonological unit composed of an initial, a final, and a tone, while Japanese syllables have a relatively small set of allowed syllable patterns, which primarily follow consonant-vowel (CV) structures. This strong alignment between orthography and pronunciation makes syllables a compelling choice for acoustic modeling in these languages.

Syllable-based acoustic modeling has attracted increasing interest in recent years. Ganapathiraju et al. [187] first demonstrated that syllable-based acoustic models could outperform traditional phone-based models in large-vocabulary tasks using HMM/GMM system. In the context of Mandarin Chinese, many prior studies [188, 189, 190, 191, 192] demonstrate that syllable-level models can outperform context-independent phoneme models and even approach or surpass state-of-the-art context-dependent baselines. In Japanese ASR, traditional systems historically used mora units [193, 194, 195, 196] for acoustic modeling.

## **7.2 A Bottom-up Framework for Syllable-based Speech Recognition**

Building upon the above-mentioned premises, our proposed framework aims to unify the strengths of attribute modeling and syllable-level recognition. By first modeling speech attributes, language universal interpretable features, we introduce a robust front-end that is resilient to phonetic variations across speakers and languages. The attribute sequences are then transformed into syllables, serving as the bridge to higher-level linguistic units. This two-stage approach supports modular training, interpretable error analysis, and extensibility to multilingual and low-resource ASR scenarios.

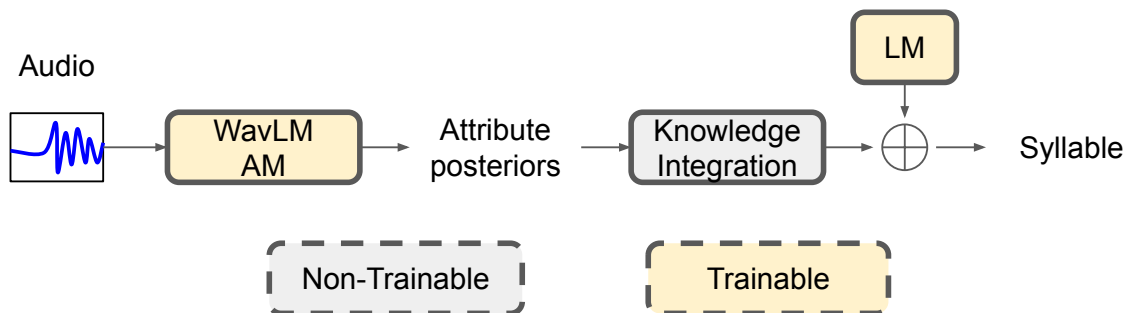


Figure 7.1: Overview of the bottom-up framework for syllable-based ASR. "AM" refers to acoustic model and "LM" stands for language model.

### 7.2.1 Bottom-up Attribute Recognition and Knowledge Integration

The proposed bottom-up framework is illustrated in Figure 7.1. The system begins with an acoustic model built on WavLM [85] and fine-tuned with CTC loss to predict speech attribute posteriors directly from audio. These posteriors represent articulatory features and serve as an intermediate, language-universal representation. Following this, a knowledge integration step is applied to select and combine the appropriate knowledge sources as well as level of attribute granularity. This step enables flexible modeling by empirically choosing which subset of attributes to include based on the desired trade-off between representational compactness and discriminative power. The final recognition output is produced via decoding with a language model trained on syllable sequences.

### 7.2.2 Mapping Syllables to Attribute Sequences

#### *Mandarin Chinese*

Each Mandarin syllable can be decomposed into two phonological components: an initial and a final. The initial is a consonant sound at the start of the syllable, or a null initial if no consonant is present, while the final consists of the vowel or vowel combination that follows, including any semivowel glide and ending nasal if applicable. In our approach, we convert each syllable into a sequence of articulatory attributes by mapping its initial and final components separately, then concatenating their attribute representations.

Table 7.1: Mandarin Chinese initials/consonants classified by manner and place of articulation. The symbols in the square brackets are the closest IPA realizations of the corresponding initials in terms of pronunciation.

		bilabial	alveolar	retroflex	alveolo-palatal	velar
stop	unaspirated	b [p]	d [t]			g [k]
	aspirated	p [p <sup>h</sup> ]	t [t <sup>h</sup> ]			k [k <sup>h</sup> ]
nasal		m [m]	n [n]			
affricate	unaspirated		z [ts]	zh [tʂ]	j [tɕ]	
	aspirated		c [ts <sup>h</sup> ]	ch [tʂ <sup>h</sup> ]	q [tɕ]	
fricative		f [f]	s [s]	sh [ʂ]	x [ɕ]	h [x]
approximant		l [l]		r [ʐ,ʒ]		

As illustrated in Table 7.1, every Mandarin initial can be characterized by its manner of articulation, e.g. stop, nasal, affricate, fricative, and approximant, and its place of articulation, e.g. bilabial, alveolar, retroflex, alveolo-palatal, and velar. In addition, an explicit aspiration feature is included to distinguish aspirated from unaspirated stops and affricates. For example, without the aspiration, the syllable initial /b/ with IPA symbol [p] will be mapped to the same feature as initial /p/. Mandarin relies on aspiration (rather than voicing) as the primary contrast for such consonants, so the aspiration attribute is crucial for correctly distinguishing pairs like /b/ vs. /p/, /d/ vs. /t/, /zh/ vs. /ch/, etc. in the attribute sequence. As for the vowels, we employ the same features as described in Table 3.1. Notably, some exceptions such as semivowel glides like y [j] and w [w] are treated analogously to their corresponding high vowels, sharing the high and front/back attributes of /i/ and /u/, respectively, since they function as vowel-like elements in finals.

### *Japanese*

Standard Japanese syllables predominantly follow the same consonant-vowel (CV) structure as Mandarin Chinese. Table 7.2 shows the common consonants in Japanese classified by their manner and place of articulation. The consonant symbols are given in Hepburn romanization [197] with their IPA equivalents in brackets. In addition, Japanese contrasts

Table 7.2: Japanese consonants classified by manner and place of articulation. The symbols in the square brackets are the closes IPA realizations of the corresponding initials in terms of pronunciation.

		bilabial	alveolar	alveolo-palatal	velar	glottal
stop	voiceless	p [p]	t [t]		k [k]	
	voiced	b [b]	d [d]		g [g]	
nasal		m [m]	n [n]	ny [ɲ]		
affricate	voiceless		ts [ts]	ch [tɕ]		
	voiced			j [dz]		
fricative	voiceless	f [ɸ]	s [s]	sh [ɕ]		h [h]
	voiced		z [z]			
approximant			r [r]			

consonants primarily by voicing (voiceless vs. voiced) [198], unlike Mandarin, which uses aspiration as the primary contrast for stops and affricates. The inclusion of the voicing feature ensures that Japanese consonant pairs like /k/-/g/, /t/-/d/, /s/-/z/ are properly differentiated when converted into attributes [199]. The vowel portion of Japanese syllables is handled with the same attributes for height and backness as in the Mandarin system. Finally, Japanese also has a moraic nasal sound that can appear as a syllable-final nasal consonant. Unlike previous studies [193, 200], we do not treat these syllable-final nasal sounds as separate and individual units but instead merge them with the preceding syllable. If present at the end of a syllable, it is often realized as [m], [n], or [ɲ], contributing a nasal manner attribute with an appropriate place attribute determined by context.

Table 7.3 shows examples of how we convert original texts in Mandarin Chinese and Japanese to syllables and different attribute sequences. By mapping the initial and final in this way, each syllable is converted into a sequence of attribute: first the attributes corresponding to the initial consonant, followed by the attributes corresponding to the final. This attribute sequence effectively encodes the phonetic content and pronunciation of the syllable in terms of fundamental articulatory categories, providing an intermediate and interpretable representation for the speech recognition system.

Table 7.3: Examples of converting syllables to different attribute sequences for Mandarin and Japanese.

	<b>Mandarin</b>	<b>Japanese</b>
Text	你好	きみ
Syllable	/ni, hao/	/ki, mi/
Manner	[nasal vowel, fricative vowel]	[stop vowel, nasal vowel]
Place	[alveolar vowel, velar vowel]	[velar vowel, bilabial vowel]

### 7.3 Dataset and Experimental Settings

#### 7.3.1 Pronunciation and Syllable Clause Error Rate

To evaluate how accurately the model captures the underlying pronunciation, independent of orthographic symbols, we introduce a metric called Pronunciation Error Rate (PrER). This metric quantifies the recognition quality at the level of articulatory attributes, offering a more direct evaluation of the acoustic modeling capability. Since not all model outputs are in the form of attribute sequences, with higher-level tokens such as syllables, we need to first convert all predicted outputs into attribute sequences using the mapping procedure described in subsection 7.2.2. Likewise, ground-truth syllables are mapped to their corresponding attribute representations to form the reference sequence. PrER enables a uniform, articulatory-level comparison of pronunciation quality across different modeling units, and a lower value indicates that the model has more accurately captured the phonetic structure of the spoken input. In practice, PrER can be calculated on different acoustic resolution, including a single category of attribute or any combination of attributes.

At the syllable level, alongside the commonly used Syllable Error Rate (SER), we introduce an additional metric called Syllable Clause Error Rate (SCER). This metric assesses recognition performance while accounting for the inherent ambiguities in the attribute-to-syllable mapping process, which will be discussed in more detail in the following sections. In practice, certain attribute sequences may correspond to multiple syllables due to over-

lapping articulatory representations, especially when the attribute inventory is not fully discriminative. To reflect this, SCER treats all syllables that share the same underlying attribute sequence as equivalent "clause". During evaluation, both the reference and predicted syllables are first mapped to their corresponding attribute sequences. Then, each sequence grouped into a single clause representing its equivalence class. As a result, SCER represents a relaxed version of the conventional syllable error rate and serves as a lower bound on the syllable error rate of any attribute-based recognition system, under an ideal rescoring or disambiguation scenario. This metric provides an important reference point for evaluating the potential performance of syllable recognition in a bottom-up ASR system.

### 7.3.2 Dataset

The AISHELL-1 corpus [201] is a Mandarin Chinese speech dataset containing approximately 178 hours of transcribed speech recorded from 400 speakers using high-fidelity microphones. For our experiments, we use the official 150-hour training set and an 18-hour development set for validation and early stopping. Evaluation is performed on the 10-hour test set, which includes 7,176 sentences. In the experiments, we will focus on toneless syllable recognition which has a total of 408 units. To simulate low-resource training conditions, we create three training subsets by randomly selecting 25%, 10%, and 5% of the original full training set.

For Japanese, we use the JSUT corpus [202] for evaluation. This corpus provides Japanese text transcriptions paired with read-style speech recordings, sampled at 48 kHz and captured in an anechoic room. All recordings were performed by a native Japanese female speaker. The full corpus contains approximately 10 hours of speech covering various speaking scenarios. In our experiments, we use the "basic5000" subset, which includes 5,000 sentences designed to cover all daily-use Japanese characters specified in the Jouyou Kanji list. We follow the official ESPnet [203] JSUT recipe<sup>1</sup> and use the first 500 sentences

---

<sup>1</sup><https://github.com/espnet/espnet/tree/master/egs2/jsut/asr1>

in the provided list as the test set. The rest 4500 texts are used to train a syllable language model. To generate the corresponding attribute sequences, we first use the Julius [204, 205] recipe to obtain phoneme transcriptions from Japanese text, and then convert these phonemes to articulatory attributes.

### 7.3.3 Models

We adopt a publicly available pre-trained self-supervised model as our acoustic model across all experiments, namely WavLM [85]. We use the "large" version, which comprises approximately 311 million parameters, featuring 24 Transformer encoder layers with an embedding dimension of 1024 and 16 attention heads. The model is pre-trained on 94k hours of unannotated speech data from multiple sources. WavLM is specifically designed to capture both local and global contextual information in speech through its improved masked prediction training objective. All models are trained using the AdamW optimizer [129] with  $\beta_1 = 0.9$  and  $\beta_2 = 0.98$  and optimized with the CTC loss function.

To support decoding with CTC model, we trained separate KenLM<sup>2</sup> language models tailored to each modeling unit. For the syllable-based models, Chinese characters were first converted into their corresponding pinyin syllables using the Python library pypinyin<sup>3</sup>. Each syllable is treated as a standalone token, and no explicit word boundaries are included in the training corpus. During decoding, a placeholder lexicon is employed, in which each token maps directly to its own pronunciation.

## **7.4 Experimental Results and Discussions**

### 7.4.1 Analysis on Attribute to Syllable Results

The output of the attribute recognition posteriors must be transformed into valid Chinese or Japanese syllables. This task can be viewed as a sequence-to-sequence mapping, akin

---

<sup>2</sup><https://github.com/kpu/kenlm>

<sup>3</sup><https://pypi.org/project/pypinyin/>

Table 7.4: Statistics of mapping from different knowledge sources (Know. Source) to Mandarin (ZH) and Japanese (JP) syllables, including manner of articulation (M), place of articulation (P), voicing (V), height (H) and backness (B) of vowels, and aspiration (A).

Lang.	Know. Source	Avg.	Max.	1-to-1 (%)	Oracle SER (%)
ZH	M+P	3.28	8	14.63	43.36
	M+P+H	1.61	4	52.40	7.73
	M+P+H+B	1.33	4	67.88	3.75
	M+P+H+B+A	1.03	2	97.46	0.33
JP	M+P	4.26	10	20.51	28.14
	M+P+H	1.78	4	47.31	9.62
	M+P+V+H+B	1.00	1	100.00	0.00

to translating from sequences of articulatory attributes to syllables. In this work, we define *knowledge sources* as specific combinations of articulatory features, such as manner (M), place (P), voicing (V), vowel height (H), backness (B), and aspiration (A), that together describe the phonetic structure of each speech sound. For example, selecting M+P as the knowledge source implies that each unit is represented by the Cartesian product of manner and place categories, i.e., each  $(m, p) \in M \times P$ . Increasing the number of features in the knowledge source, such as adding H or V, refines the phonetic resolution and reduces ambiguity in the attribute-to-syllable mapping. To quantify the complexity and potential uncertainty of this transformation, as discussed in subsection 5.4.5, we analyze the mapping between various knowledge sources and syllables for both Mandarin Chinese and Japanese. We compute statistics including the average number of syllables per attribute sequence, the maximum overlap, the percentage of one-to-one mappings, and the oracle SER assuming perfect attribute recognition. The results are summarized in Table 7.4.

From the table, we observe that using only the coarse manner and place of articulation (M+P), as defined in chapter 4, results in a highly ambiguous mapping for both Mandarin and Japanese. For Mandarin, a single attribute sequence maps to an average of 3.28 possible syllables, with up to 8 syllables sharing the same sequence in the worst case; only 14.63% of sequences map uniquely, leading to an oracle SER of 43.36%. Similarly, for

Japanese, the average ambiguity with M+P is even higher at 4.26 syllables per sequence, the maximum reaches 10, and only 20.51% of sequences map uniquely, yielding an oracle SER of 28.14%. Adding the height of vowels (H) significantly reduces ambiguity in both languages: for Mandarin, the average drops to 1.61, the unique mapping rate rises to 52.40%, and the oracle SER falls to 7.73%; for Japanese, the average becomes 1.78, with 47.31% unique mappings and an oracle SER of 9.62%. However, the maximum ambiguity remains at 4 and the oracle SERs are still 7.73% and 9.62% respectively, suggesting that further refinements are necessary to fully resolve the ambiguity.

Incorporating additional articulatory features such as voicing (V) and vowel backness (B) further improves disambiguation for both languages. For Mandarin, the M+P+H+B combination reduces the average ambiguity to 1.33 syllables, increases the one-to-one mapping percentage to 67.88%, and lowers the oracle SER to 3.75%, though the maximum remains at 4. For Japanese, this feature set completely resolves the ambiguity: the average ambiguity drops to exactly 1.00, with every sequence mapping uniquely, resulting in a perfect oracle SER. This indicates that once we have a sufficient and accurate pronunciation information, we can obtain a perfect transcription of syllables for Japanese. Finally, by introducing an auxiliary articulatory feature related to aspiration (A) into the full inventory for Mandarin (M+P+H+B+A), the mapping becomes nearly deterministic, with an average ambiguity of 1.03, a maximum of 2, and 97.46% unique mappings, achieving an oracle SER as low as 0.33%. These results demonstrate that a sufficiently rich and well-structured attribute inventory can serve as a reliable intermediate representation for accurate syllable prediction in both Mandarin Chinese and Japanese.

#### 7.4.2 Pronunciation Error Rate on AISHELL-1

Table 7.5 presents the Pronunciation Error Rate (PrER) results on the AISHELL-1 test set using the full 150-hour training data. All systems use the WavLM-large encoder as the acoustic backbone. Specifically, WavLM<sub>syl</sub> denotes the baseline model trained to predict

Table 7.5: PrER(%) on test set of AISHELL-1 for different categories of speech attributes, where  $\prod = M \times P \times H \times B \times A$  is the Cartesian product of all the categories. No language model is applied during decoding.

System	PrER (%)					
	M	P	H	B	A	$\prod$
WavLM <sub>syl</sub>	0.78	0.78	0.79	0.76	0.51	1.30
WavLM <sub>attr</sub> (ours)	<b>0.72</b>	<b>0.75</b>	<b>0.77</b>	<b>0.65</b>	<b>0.53</b>	<b>1.23</b>

syllables directly, while WavLM<sub>BU</sub> represents our proposed bottom-up system where the acoustic model predicts sequences of articulatory attributes at different levels of granularity. For evaluating PrER, no language model is applied during decoding to isolate the pure acoustic modeling performance. We report PrER across all categories of speech attributes, as well as their combined Cartesian product  $\prod = M \times P \times H \times B \times A$ , which reflects the most fine-grained acoustic resolution.

As shown in the results, the bottom-up system consistently outperforms the direct syllable model in terms of PrER across all categories of speech attributes. The performance gap remains clear even at the highest acoustic resolution. Specifically, WavLM<sub>BU</sub> achieves a PrER of 1.23%, improving upon the 1.30% of the syllable-based system. These results highlight the advantage of the bottom-up attribute modeling approach in capturing fine-grained phonetic details that conventional syllable-level models may miss, underscoring its potential for more accurate and language-universal pronunciation modeling.

### 7.4.3 Bottom-up Syllable-based Recognition Results on AISHELL-1

Table 7.6 shows the syllable recognition results on the AISHELL-1 test set. The baseline system, WavLM<sub>syl</sub>, directly predicts syllables in the conventional manner, while the proposed bottom-up system, WavLM<sub>BU</sub>, can incrementally integrate different levels of articulatory knowledge. From the table, we can make several observations. First, the intermediate SER results of WavLM<sub>BU</sub> illustrate the benefit of progressively adding the articulatory knowledge source. The SER drops sharply from 47.62% when using only manner and

Table 7.6: Experimental results on the AISHELL-1 test set, reporting SER and SCER. For the proposed bottom-up system, different knowledge sources (attribute categories) are added incrementally. SCER is normalized with the full set of knowledge sources, ensuring maximum disambiguation of syllables.

<b>System</b>	<b>Know. Source</b>	<b>SER(%)</b>	<b>SCER(%)</b>
WavLM <sub>syl</sub>	–	<b>2.36</b>	2.33
WavLM <sub>BU</sub> (ours)	M+P	47.62	–
	M+P+H	9.83	–
	M+P+H+B+A	2.47	<b>2.12</b>

place, to 9.83% when adding vowel height, and further down to 2.47% with the complete attribute set. This trend confirms that finer-grained articulatory details effectively resolve pronunciation ambiguities, boosting recognition accuracy.

The best bottom-up model using the full knowledge source (M+P+H+B+A) achieves a SER of 2.47%, which is slightly higher than the direct syllable model’s 2.36%. This small gap can be explained by remaining ambiguities in mapping detailed attribute sequences to syllables, as discussed earlier. To better isolate the true acoustic modeling capability, we also compute SCER, where each predicted syllable is normalized according to the complete knowledge source (M+P+H+B+A), meaning two syllables are considered equivalent if their underlying attribute sequences are identical across all categories. Under this relaxed evaluation, WavLM<sub>BU</sub> achieves a lower SCER of 2.12%, outperforming WavLM<sub>syl</sub> with 2.33%. In addition, the difference between the SER and SCER for WavLM<sub>BU</sub> is approximately equal to the oracle SER of 0.33% shown in Table 7.4, confirming that the bottom-up approach captures pronunciation more accurately and that the SER gap is mainly due to final mapping ambiguities.

These results suggest that most of the errors in the bottom-up system occur not from misrecognition of the phonetic content, but from ambiguity in the final syllable mapping process. In contrast, for WavLM<sub>syl</sub>, the close alignment between its SER and SCER indicates that its errors stem directly from acoustic misrecognition, which are inherently more

Table 7.7: Two qualitative examples from the AISHELL-1 test set.

Model	Syllables	
Ground Truth	/... lu ren ji bu shang <b>qian</b> .../	/... wu <b>xian yi jin deng</b> wei jie kou/
WavLM <sub>syl</sub>	/... lu ren ji bu shang <b>xian</b> .../	/... wu <b>qian jie de</b> wei jie kou/
WavLM <sub>BU</sub>	/... lu ren ji bu shang <b>quan</b> .../	/... wu <b>xuan yi jin deng</b> wei jie kou/

difficult to correct during post-processing or rescore. Table 7.7 presents two qualitative examples from the AISHELL-1 test set. In the first example, both WavLM<sub>syl</sub> and WavLM<sub>BU</sub> incorrectly recognize the syllable /qian/. However, WavLM<sub>syl</sub> misclassifies the initial consonant /q/ as /x/, which differ in their manner of articulation. In contrast, although WavLM<sub>BU</sub> outputs /quan/ instead of /qian/, the two syllables share the same attribute sequences in all categories of M, P, H, B, and A. In the second example, WavLM<sub>syl</sub> misrecognizes both /xian/ and /jin/ as /qian/ and /jin/, respectively, and fails to capture the syllable /yi/ and the nasal ending of /deng/. Meanwhile, WavLM<sub>BU</sub> misrecognizes /xian/ as /xuan/, which shares the same pronunciation as well. This contrast underscores that with improved attribute-to-syllable rescoring and additional knowledge sources, our proposed bottom-up approach could close the remaining performance gap.

#### 7.4.4 Ablation Study on Low-resource Training

To further evaluate the robustness of the proposed bottom-up framework under limited supervision, we conduct experiments on the AISHELL-1 dataset using only 5%, 10%, and 25% of the available training data. Figure 7.2 shows the comparative results across three evaluation metrics: PrER, SER, and SCER. As shown in the plot, the bottom-up system consistently achieves lower PrER than the syllable-based model across all training sizes. This highlights the model’s superior ability to learn pronunciation-relevant acoustic features, especially under low-resource conditions. Since PrER is measured without any language model, this advantage can be attributed solely to the stronger acoustic modeling of articulatory attributes with the bottom-up system, which enables more data-efficient learn-

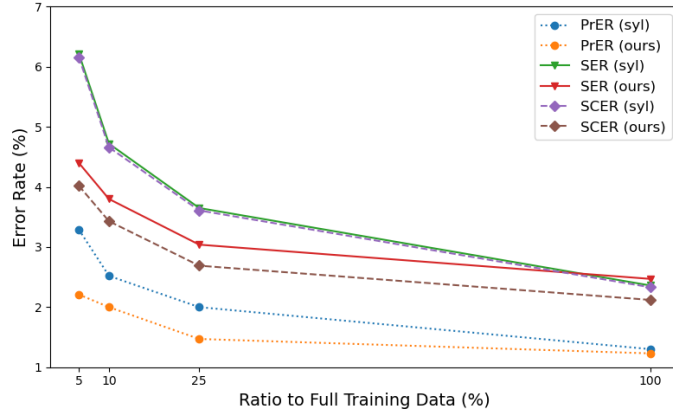


Figure 7.2: Experimental results (PrER, SER, and SCER) of using different amount of training data. The results for WavLM<sub>syl</sub> (syl) and WavLM<sub>BU</sub> (ours) are shown.

ing of pronunciation structure.

While the syllable-based model outperforms the bottom-up system in terms of SER when the full training data is used, the bottom-up system actually surpasses the syllable-based baseline under low-resource conditions where less than 25% of the training data is presented. This advantage is likely due to the significantly higher number of syllable units, which makes it difficult to adequately train all units when data is limited, whereas the bottom-up system benefits from a more compact and generalizable unit inventory. In addition, we observe a similar pattern to that in Table 7.6, where the SCERs of WavLM<sub>syl</sub> closely matches its SERs across all training conditions. In contrast, the SCERs of WavLM<sub>BU</sub> remain consistently lower than their corresponding SERs in all scenarios. This strengthens our assumption that the majority of errors in the bottom-up system arise from ambiguities in the attribute-to-syllable mapping stage, rather than from inaccuracies in the acoustic predictions themselves. These results underscore the potential of bottom-up system as a strong foundation for continuous speech recognition in syllable-based languages, particularly in low-resource settings where direct end-to-end training is less effective.

### 7.4.5 Cross-lingual Transfer to Japanese Syllable Recognition

Table 7.8 shows the PrER results for zero-shot Japanese syllable recognition using our proposed bottom-up framework trained on different source languages. Besides the WavLM<sub>BU</sub> model trained on AISHELL-1 as discussed in previous sections, we also train an additional model on the Multilingual LibriSpeech (MLS) dataset as described in chapter 3, following the same bottom-up manner. We can make several observations from these results. First, the model trained solely on Mandarin Chinese demonstrates a moderate level of cross-lingual transfer capability. Despite never seeing Japanese data during training, it is able to predict Japanese speech attributes to a certain degree. This confirms that a language-universal attribute inventory can generalize beyond the source language. However, its performance remains suboptimal for Japanese because Mandarin alone does not cover all the phonetic nuances needed, particularly distinctions like voicing contrasts that are absent in Mandarin but common in Japanese. This limitation reflects an inherent phonetic gap which degrades the overall performance.

In comparison, the model trained on MLS achieves substantially lower PrER on Japanese despite using less total training data since it is exposed to eight languages, giving it a broader, more universal phonetic foundation. The result is a stronger zero-shot performance: even with only 80 hours of training data, the multilingual model’s Japanese PrER is lower, underscoring the benefit of phonetic diversity in the training corpus. By learning articulatory features from multiple languages, the multilingual acoustic model generalizes

Table 7.8: PrER(%) on subset of JSUT dataset (“basic5000”) for different categories of speech attributes, from different source language ( $S$ ), including 150 hours of Mandarin Chinese (ZH) and 80 hours of Multilingual LibriSpeech (MLS), to target ( $T$ ) Japanese (JP). No language model is applied during decoding.

$S \rightarrow T$	# Train (hrs.)	PrER (%)				
		M	P	V	H	B
ZH $\rightarrow$ JP	150	33.76	32.85	22.55	34.71	31.22
MLS $\rightarrow$ JP	80	18.16	19.87	12.34	23.39	16.50

better to an unseen language, aligning with our findings in chapter 3 that such universal speech attributes yield more consistent cross-lingual performance. In essence, the multi-lingual training allows the model with a wider coverage of speech sounds, making it more robust for zero-shot transfer to an unseen language like Japanese. These findings demonstrate that the bottom-up framework naturally promotes language universality and enables robust cross-lingual adaptability, even in challenging low-resource or zero-shot scenarios.

To integrate all the necessary knowledge sources for effective syllable recognition, a pronunciation dictionary and language model are required. These components serve as additional knowledge sources from the target language and are all that is needed to perform zero-shot speech recognition once a language-universal acoustic model for speech attribute recognition is in place. Table 7.9 shows the results of using our bottom-up framework to perform zero-shot Japanese syllable recognition. Here all our systems with WavLM are trained on the 80 hours MLS dataset as it shows better PrER in Table 7.8. From the results, we observe that although characters and phonemes can be viewed as relatively language-universal units, the systems WavLMchar and WavLMphn struggle with cross-lingual transfer, resulting in high syllable error rates (SERs) of 67.79% and 64.92%, respectively. In contrast, our proposed bottom-up framework demonstrates similar patterns as shown in Mandarin Chinese with consistent improvement as more articulatory knowledge sources are incorporated. Specifically, starting with basic M+P, the SER is 79.43%. Adding vowel height (H) reduces the SER to 63.36%, and with the full categories of attributes (M+P+V+H+B), the SER further drops to 40.08%.

While there remains a performance gap compared to benchmark models such as Whisper [59] and XLS-R [206], it is important to emphasize that these systems are trained on significantly larger amount of data and include the target language during training. For instance, Whisper is trained on 680,000 hours of multilingual speech that includes Japanese, and XLS-R is fine-tuned specifically on Japanese data, including the JSUT dataset, which reasonably yields the best results. In contrast, our framework tackles a more challenging

Table 7.9: SER(%) on the subset of JSUT dataset (“basic5000”) test set. All the WavLM systems are trained on 80 hours MLS dataset and tested on Japanese. \* refers to systems use JSUT during training. For the proposed bottom-up system, different knowledge sources (attribute categories) are added incrementally.

System	Seen $T$ Lang.	Know. Source	SER(%)
Whisper	✓	–	24.01
XLS-R* [206]			13.97
WavLM <sub>char</sub>	✗	character	67.79
WavLM <sub>phn</sub>		phoneme	64.92
WavLM <sub>BU</sub> (ours)	✗	M+P	79.43
		M+P+H	63.36
		M+P+V+H+B	40.08

scenario, which is the actual zero-shot speech recognition with no target language speech data, using only 80 hours of training data from eight Indo-European languages that differ substantially from Japanese. Nevertheless, the results show that our bottom-up approach achieves reasonable performance under this challenging setting. This demonstrates the potential for cross-lingual transfer to unseen languages and highlights the effectiveness of leveraging language-universal speech attributes combined with minimal language-specific knowledge sources. In addition, the proposed bottom-up framework is flexible and can be seamlessly integrated with large pre-trained models and advanced training techniques, as they are all compatible with our approach.

## 7.5 Summary

In summary, this chapter presented a bottom-up framework for continuous speech recognition in syllable-based languages by leveraging a language-universal set of articulatory attributes as the fundamental modeling unit. Through detailed analyses on Mandarin Chinese and Japanese, we demonstrated that the proposed approach effectively models fine-grained pronunciation, achieves competitive syllable recognition performance, and maintains robustness under limited supervision. Furthermore, our cross-lingual transfer exper-

iments confirmed the framework’s inherent ability to generalize to unseen syllable-based languages when supported by phonetic diversity in the training data. These findings underscore the potential of bottom-up attribute modeling as a versatile and scalable foundation for multilingual ASR systems, particularly in low-resource or zero-shot settings where conventional end-to-end methods face significant limitations.

## CHAPTER 8

### CONCLUSIONS AND FUTURE WORK

#### 8.1 Summary

This dissertation presents a unified framework and comprehensive study for multilingual automatic speech recognition (ASR) through modeling of language-universal articulatory attributes, diverging from traditional language-specific units such as characters, subwords, and phonemes. It is expected that our proposed research can contribute to the development of more robust, interpretable, and generalizable ASR systems, particularly in scenarios where linguistic resources are scarce or where the target languages differ significantly from those seen during training. By shifting the modeling paradigm toward universal phonetic representations grounded in articulatory theory, this work lays the foundation for scalable ASR architectures that better accommodate linguistic diversity and support broader speech applications.

To provide a compact, transferable, and interpretable representation of speech sounds, we construct a unified articulatory attribute inventory that maps phonemes across diverse languages into a shared set of features, including manner and place of articulation, voicing, vowel height and backness. We investigate two primary modeling strategies: separate attribute recognizers and joint modeling in a structured Cartesian product space. In the separate recognizer approach, each attribute category is treated as an independent prediction task, enabling the model to capture fine-grained articulatory patterns without relying on phoneme-level supervision. This design is motivated by the intrinsic sharing mechanism of speech attributes where each attribute feature is common to a group of phonemes, allowing the model to aggregate training data across multiple categories. As a result, speech attributes are less sensitive to individual phoneme-level labeling noise and can be trained

more robustly. Our experimental results support this preliminary assumption, demonstrating that this approach leads to more balanced token distributions and achieves improved recognition performance compared to traditional character- and phoneme-based models. Moreover, when attribute predictions are incorporated as knowledge-based constraints, further gains in phoneme recognition performance are observed, validating the effectiveness of the attribute modeling strategy.

In contrast, the product space approach models speech attributes jointly, capturing co-occurrence dependencies among features. By modeling attributes in a combined product space, such as manner-place or manner-place-voicing combinations, we extend the attribute-based framework beyond phoneme recognition to word-level tasks. Specifically, we demonstrate its effectiveness in zero-shot spoken keyword recognition by integrating it with a non-trainable pronunciation model and domain adversarial training (DAT). This system achieves strong cross-lingual transferability, performing well even in languages not seen during training. Furthermore, in continuous speech recognition scenarios, we show that attribute-based token recognition form a reliable foundation for keyword information extraction. When combined with utterance verification modules, including both statistical and neural embedding-based methods, the proposed approach substantially reduces false rejection rates and improves overall performance, demonstrating the robustness of attribute-based modeling in open-vocabulary and low-resource settings. Finally, we propose a bottom-up ASR system to syllable-based continuous ASR, demonstrating the applicability and adaptability of our proposed attribute modeling framework to general continuous speech recognition tasks.

## **8.2 Thesis Contributions**

Our contributions can be summarized as follows:

1. We construct a language-universal inventory of fundamental speech attributes. This inventory covers articulatory features, such as manner and place of articulation, that

are shared across many languages, based on linguistic knowledge from the International Phonetic Alphabet (IPA). We develop attribute recognition models trained on multilingual speech data to predict these features with high accuracy.

2. We propose a first step toward multilingual end-to-end automatic speech recognition (ASR) by integrating the above speech attributes as auxiliary predictions and knowledge-based constraints. We introduce a deterministic attribute-to-phoneme mapping, which projects the knowledge-rich articulatory attribute logits, into output phoneme logits.
3. We extend our language-universal attribute modeling framework to end-to-end spoken keyword recognition (SKR) by leveraging a self-supervised pretrained model and a non-trainable pronunciation model. The incorporation of domain adversarial training (DAT) significantly strengthens the framework, improving zero-shot recognition for unseen languages.
4. We develop a two-stage keyword information extraction (KIE) system that enhances keyword detection by leveraging language-universal speech attributes in the first-stage ASR. This compact attribute-based representation improves recall by capturing all putative keyword hits. In the second stage, two phoneme-level utterance verification systems are explored to refine these hits by validating keyword presence and boundaries, effectively reducing false alarms.
5. We introduce a bottom-up ASR framework for syllable-based speech recognition by leveraging articulatory and pronunciation knowledge sources. We further propose new evaluation metrics, including Pronunciation Error Rate (PrER) and Syllable Clause Error Rate (SCER), to assess both the accuracy of the phonetic representation and the ambiguity in syllable recovery. The proposed framework offers superior pronunciation modeling, especially in low-resource conditions, and provides strong language universality.

### **8.3 Future Work**

Looking ahead, several promising directions for future work emerge from this dissertation. First, theoretical extensions of this work could include integrating the framework with advanced probabilistic modeling, investigating the theoretical limitations of mapping from attribute to higher level units, and incorporating acoustic knowledge to strengthen the foundations of this approach. Second, on the practical side, future work could integrate the attribute-based approach into large-scale multilingual ASR systems and speech processing pipelines, and explore its application in low-resource or domain-specific speech recognition settings. Finally, the approach could be extended to diverse languages by leveraging transfer learning to benefit low-resource cases and by exploring unsupervised attribute discovery to automatically adapt to arbitrary articulatory features.

## REFERENCES

- [1] J. Li, “Recent advances in end-to-end automatic speech recognition,” in *Proc. AP-SIPA*, 2022.
- [2] P. Cohen *et al.*, “Towards a universal speech recognizer for multiple languages,” in *Proc. ASRU*, 1997.
- [3] T. Schultz and A. Waibel, “Language-independent and language-adaptive acoustic modeling for speech recognition,” in *Proc. ICSLP*, 1998.
- [4] W. Byrne *et al.*, “Towards language independent acoustic modeling,” in *Proc. ICASSP*, 2000.
- [5] E. G. Garcia, E. Mengusoglu, and E. Janke, “Multilingual acoustic models for speech recognition in low-resource devices,” in *Proc. ICASSP*, 2007.
- [6] T. R. Niesler, “Language-dependent state clustering for multilingual acoustic modelling,” *Speech Commun.*, vol. 49, pp. 453–463, 2007.
- [7] C.-H. Lee and S. M. Siniscalchi, “An information-extraction approach to speech processing: Analysis, detection, verification, and recognition,” *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1089–1115, 2013.
- [8] S. Watanabe, T. Hori, and J. R. Hershey, “Language independent end-to-end architecture for joint language identification and speech recognition,” in *Proc. ASRU*, 2017.
- [9] L. Zhou, J. Li, E. Sun, and S. Liu, “A configurable multilingual model is all you need to recognize all languages,” in *Proc. ICASSP*, 2021.
- [10] C. Zhang, B. Li, T. N. Sainath, T. Strohmaier, and S.-y. Chang, “Uml: A universal monolingual output layer for multilingual asr,” in *Proc. ICASSP*, 2023.
- [11] H.-C. Lin, L. Deng, D. Yu, Y. Gong, A. Acero, and C.-H. Lee, “A study on multilingual acoustic modeling for large vocabulary asr,” in *Proc. ICASSP*, 2009.
- [12] A. Kannan *et al.*, “Large-scale multilingual speech recognition with a streaming end-to-end model,” in *Proc. Interspeech*, 2019.
- [13] C. L. Smith, “Handbook of the international phonetic association: A guide to the use of the international phonetic alphabet (1999). cambridge: Cambridge university press. pp. ix 204.,” *Phonology*, vol. 17, no. 2, pp. 291–295, 2000.

- [14] J. Kohler, “Multi-lingual phoneme recognition exploiting acoustic-phonetic similarities of sounds,” in *Proc. ICSLP*, 1996.
- [15] S. Gokcen and J. Gokcen, “A multilingual phoneme and model set: Toward a universal base for automatic speech recognition,” in *Proc. ASRU*, 1997.
- [16] F. Metze and A. H. Waibel, “A flexible stream architecture for asr using articulatory features,” in *Proc. Interspeech*, 2002.
- [17] S. M. Siniscalchi, D.-C. Lyu, T. Svendsen, and C.-H. Lee, “Experiments on cross-language attribute detection and phone recognition with minimal target-specific training data,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 3, pp. 875–887, 2012.
- [18] I.-F. Chen, C. Ni, B. P. Lim, N. F. Chen, and C.-H. Lee, “A novel keyword+lvcsr-filler based grammar network representation for spoken keyword search,” in *Proc. ISCSLP*, 2014.
- [19] W. Li, K. Li, S. M. Siniscalchi, N. F. Chen, and C.-H. Lee, “Detecting mispronunciations of 12 learners and providing corrective feedback using knowledge-guided and data-driven decision trees,” in *Proc. Interspeech*, 2016.
- [20] Y.-J. Lu *et al.*, “Improving speech enhancement performance by leveraging contextual broad phonetic class information,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 2738–2750, 2023.
- [21] Y. Wu and M. Huang, “Cnn-based articulatory feature recognition for kunqu-singing pronunciation evaluation,” in *Proc. CISCE*, 2023.
- [22] I. Bromberg *et al.*, “Detection-based asr in the automatic speech attribute transcription project,” in *Proc. Interspeech*, 2007.
- [23] M. Gales and S. Young, *Application of Hidden Markov Models in Speech Recognition*. Now Foundations and Trends, 2008.
- [24] M. Gales, “Acoustic modelling for speech recognition: Hidden markov models and beyond?” In *Proc. ASRU*, 2009.
- [25] D. Povey *et al.*, “The kaldi speech recognition toolkit,” in *Proc. ASRU*, 2011.
- [26] J. Chorowski, D. Bahdanau, K. Cho, and Y. Bengio, “End-to-end continuous speech recognition using attention-based recurrent nn: First results,” in *Proc. NeurIPS*, 2014.

- [27] C. Lüscher *et al.*, “Rwth asr systems for librispeech: Hybrid vs attention - w/o data augmentation,” in *Proc. Interspeech*, 2019.
- [28] T. K. Jaeyoung Lee Masato Mimura, “Embedding articulatory constraints for low-resource speech recognition based on large pre-trained model,” in *Proc. Interspeech*, 2023.
- [29] K. Glocker, A. Herygers, and M. Georges, “Allophant: Cross-lingual phoneme recognition with articulatory attributes,” in *Proc. Interspeech*, 2023.
- [30] X. Li, S. Dalmia, D. R. Mortensen, F. Metze, and A. W. Black, *Zero-shot learning for speech recognition with universal phonetic model*, 2019.
- [31] S. Li, C. Ding, X. Lu, P. Shen, T. Kawahara, and H. Kawai, “End-to-end articulatory attribute modeling for low-resource multilingual speech recognition,” in *Proc. Interspeech*, 2019.
- [32] N. M. J. Basil Abraham S. Umesh, “Articulatory feature extraction using ctc to build articulatory classifiers without forced frame alignments for speech recognition,” in *Proc. Interspeech*, 2016.
- [33] M. Müller, S. Stüker, and A. H. Waibel, “Towards improving low-resource speech recognition using articulatory and language features,” in *Proc. SLT*, 2016.
- [34] L. Qu, C. Weber, E. Lakomkin, J. Twiefel, and S. Wermter, “Combining articulatory features with end-to-end learning in speech recognition,” in *Proc. ICANN*, 2018.
- [35] S. Toshniwal *et al.*, “Multilingual speech recognition with a single end-to-end model,” in *Proc. ICASSP*, 2017.
- [36] J. T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, “Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers,” in *Proc. ICASSP*, 2013.
- [37] S. Dalmia, R. Sanabria, F. Metze, and A. W. Black, “Sequence-based multi-lingual low resource speech recognition,” in *Proc. ICASSP*, 2018.
- [38] G. Heigold *et al.*, “Multilingual acoustic models using distributed deep neural networks,” in *Proc. ICASSP*, 2013.
- [39] A. Ghoshal, P. Swietojanski, and S. Renals, “Multilingual training of deep neural networks,” in *Proc. ICASSP*, 2013.

- [40] K. Veselý, M. Karafiát, F. Grézl, M. Janda, and E. Egorova, “The language-independent bottleneck features,” in *Proc. SLT*, 2012.
- [41] V. Pratap *et al.*, “Massively multilingual asr: 50 languages, 1 model, 1 billion parameters,” in *Proc. Interspeech*, 2020.
- [42] B. Li *et al.*, “Scaling end-to-end models for large-scale multilingual asr,” in *Proc. ASRU*, 2021.
- [43] A. Das, K. Kumar, and J. Wu, “Multi-dialect speech recognition in english using attention on ensemble of experts,” in *Proc. ICASSP*, 2021.
- [44] N. Gaur *et al.*, “Mixture of informed experts for multilingual speech recognition,” in *Proc. ICASSP*, 2021.
- [45] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, “End-to-end attention-based large vocabulary speech recognition,” in *Proc. ICASSP*, 2016.
- [46] A. Graves, S. Fernandez, F. Gomez, and J. Schmidhuber, “Towards end-to-end speech recognition with recurrent neural networks,” in *Proc. ICML*, 2006.
- [47] J. Chorowski, D. Bahdanau, K. Cho, and Y. Bengio, “Attention-based models for speech recognition,” in *Proc. NeurIPS*, 2015.
- [48] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, “Listen, attend and spell,” in *Proc. ICASSP*, 2016.
- [49] A. Y. Hannun *et al.*, “Deep speech: Scaling up end-to-end speech recognition,” *ArXiv*, vol. abs/1412.5567, 2014.
- [50] S. M. Siniscalchi, P. Schwarz, and C.-H. Lee, “High-accuracy phone recognition by combining high-performance lattice generation and knowledge based rescoring,” in *Proc. ICASSP*, 2007.
- [51] G. Hinton *et al.*, “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.
- [52] I.-F. Chen and C.-H. Lee, “A study on using word-level hmms to improve asr performance over state-of-the-art phone-level acoustic modeling for lvcsr,” in *Proc. Interspeech*, 2012.
- [53] A.-r. Mohamed, G. E. Dahl, and G. E. Hinton, “Acoustic modeling using deep belief networks,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 14–22, 2012.

- [54] W. Zhou, M. Zeineldeen, Z. Zheng, R. Schlüter, and H. Ney, “Acoustic data-driven subword modeling for end-to-end speech recognition,” in *Proc. Interspeech*, 2021.
- [55] K. Livescu, E. Fosler-Lussier, and F. Metze, “Subword modeling for automatic speech recognition: Past, present, and emerging approaches,” *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 44–57, 2012.
- [56] T. Hori, J. Cho, and S. Watanabe, “End-to-end speech recognition with word-based rnn language models,” in *Proc. ICASSP*, 2018.
- [57] Z. Xiao, Z. Ou, W. Chu, and H.-C. Lin, “Hybrid ctc-attention based end-to-end speech recognition using subword units,” in *Proc. ICSLP*, 2018.
- [58] R. Sennrich, B. Haddow, and A. Birch, “Neural machine translation of rare words with subword units,” *ArXiv*, vol. abs/1508.07909, 2015.
- [59] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language models are unsupervised multitask learners,” 2019.
- [60] International Phonetic Association, “The international phonetic association (revised to 1993) – ipa chart,” *Journal of the International Phonetic Association*, vol. 23, 1993.
- [61] T. Schultz and A. H. Waibel, “Language-independent and language-adaptive acoustic modeling for speech recognition,” *Speech Commun.*, vol. 35, no. 1, pp. 31–51, 2001.
- [62] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, “Automatic speech recognition for under-resourced languages: A survey,” *Speech Commun.*, vol. 35, pp. 31–51, 2014.
- [63] A. Gutkin, “Uniform multilingual multi-speaker acoustic model for statistical parametric speech synthesis of low-resourced languages,” in *Proc. Interspeech*, 2017.
- [64] S. Feng, M. Tu, R. Xia, C. Huang, and Y. Wang, “Language-universal phonetic representation in multilingual speech pretraining for low-resource speech recognition,” in *Proc. Interspeech*, 2023.
- [65] C. Taguchi, Y. Sakai, P. Haghani, and D. Chiang, “Universal automatic phonetic transcription into the international phonetic alphabet,” *ArXiv*, 2023.
- [66] D. R. Mortensen *et al.*, “Allovera: A multilingual allophone database,” in *Proc. LREC*, 2020.

- [67] X. Li *et al.*, “Universal phone recognition with a multilingual allophone system,” in *Proc. ICASSP*, 2020.
- [68] X. Li, J. Li, F. Metze, and A. W. Black, “Hierarchical phone recognition with compositional phonetics,” in *Proc. Interspeech*, 2021.
- [69] B. Yan, S. Dalmia, D. R. Mortensen, F. Metze, and S. Watanabe, “Differentiable allophone graphs for language-universal speech recognition,” in *Proc. Interspeech*, 2021.
- [70] S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 28, no. 4, pp. 357–366, 1980.
- [71] H. Hermansky, “Perceptual linear predictive (plp) analysis of speech.,” *The Journal of the Acoustical Society of America*, vol. 87 4, pp. 1738–52, 1990.
- [72] P. Boersma, “Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound,” in *Proceedings of the Institute of Phonetic Sciences, University of Amsterdam*, 1993.
- [73] E. Shriberg, “Spontaneous speech: How people really talk and why engineers should care,” in *Proc. Interspeech*, 2005.
- [74] J. P. Teixeira, C. Oliveira, and C. Lopes, “Vocal acoustic analysis – jitter, shimmer and hnr parameters,” *Procedia Technology*, vol. 9, pp. 1112–1122, 2013.
- [75] K. N. Stevens, *Acoustic Phonetics*. The MIT Press, 1999.
- [76] I. Goodfellow *et al.*, “Generative adversarial nets,” in *Proc. NeurIPS*, 2014.
- [77] J. Yi, J. Tao, Z. Wen, and Y. Bai, “Adversarial multilingual training for low-resource speech recognition,” in *Proc. ICASSP*, 2018.
- [78] K. Hu, H. Sak, and H. Liao, “Adversarial training for multilingual acoustic modeling,” *ArXiv*, 2019.
- [79] H. Yen, P.-J. Ku, S. M. Siniscalchi, and C.-H. Lee, “Language-universal speech attributes modeling for zero-shot multilingual spoken keyword recognition,” in *Proc. Interspeech*, 2025.
- [80] S. Schneider, A. Baevski, R. Collobert, and M. Auli, “Wav2vec: Unsupervised pre-training for speech recognition,” *ArXiv*, vol. abs/1904.05862, 2019.

- [81] A. Baevski, S. Schneider, and M. Auli, “Vq-wav2vec: Self-supervised learning of discrete speech representations,” in *Proc. ICLR*, 2020.
- [82] A. Baevski, H. Zhou, A.-r. Mohamed, and M. Auli, “Wav2vec 2.0: A framework for self-supervised learning of speech representations,” in *Proc. NeurIPS*, 2020.
- [83] W. Hsu, B. Bolte, Y. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, “Hubert: Self-supervised speech representation learning by masked prediction of hidden units,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3451–3460, 2021.
- [84] C. Chiu, J. Qin, Y. Zhang, J. Yu, and Y. Wu, “Self-supervised learning with random-projection quantizer for speech recognition,” *Proc. ICML*, 2022.
- [85] S. Chen *et al.*, “Wavlm: Large-scale self-supervised pre-training for full stack speech processing,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.
- [86] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” *ArXiv*, 2022.
- [87] Y. Zhang *et al.*, “Google usm: Scaling automatic speech recognition beyond 100 languages,” *ArXiv*, vol. abs/2303.01037, 2023.
- [88] S. Communication *et al.*, “Seamlessm4t: Massively multilingual & multimodal machine translation,” *ArXiv*, 2023.
- [89] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, “Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 9, pp. 2251–2265, 2017.
- [90] C.-K. Yang, K.-P. Huang, K.-H. Lu, C.-Y. Kuan, C.-Y. Hsiao, and H.-y. Lee, “Investigating zero-shot generalizability on mandarin-english code-switched asr and speech-to-text translation of recent foundation models with self-supervision and weak supervision,” in *Proc. ICASSP*, 2023.
- [91] A. Rouditchenko *et al.*, “Comparison of multilingual self-supervised and weakly-supervised speech pre-training for adaptation to unseen languages,” in *Proc. Interspeech*, 2023.
- [92] Q. Xu, J. Chen, A. Sriram, and G. Synnaeve, “Improving multilingual asr by fine-tuning whisper for low-resource languages,” in *Proc. ICASSP*, 2023.
- [93] Y. Wang, Y. Li, Z. Guo, and M. Chen, “Multilingual fine-tuning of whisper for improved cross-lingual asr,” in *Proc. EMNLP*, 2023.

- [94] A. Kumar, R. Singh, and D. Patel, “Fine-tuning google usm for robust asr in under-represented languages,” *ArXiv*, 2023.
- [95] X. Li, “Low-resource speech recognition for thousands of languages,” Ph.D. dissertation, Carnegie Mellon University, Pittsburgh, PA, USA, 2023.
- [96] D. M. Eberhard, G. F. Simons, and C. D. Fennig, *Ethnologue: Languages of the World*. SIL International, 2023.
- [97] F. Coulmas, *Written and unwritten language (Key Topics in Sociolinguistics)*. Cambridge University Press, 2013, pp. 39–59.
- [98] J. Lu, P. Gong, J. Ye, J. Zhang, and C. Zhang, “A survey on machine learning from few samples,” *Pattern Recognition*, vol. 139, pp. 109–480, 2023.
- [99] S. J. Pan and Q. Yang, “A survey on transfer learning,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.
- [100] D. Wang and T. F. Zheng, “Transfer learning for speech and language processing,” in *Proc. APSIPA*, 2015.
- [101] H. Kheddar, Y. Himeur, S. Al-maadeed, A. Amira, and F. Bensaali, “Deep transfer learning for automatic speech recognition: Towards better generalization,” *ArXiv*, vol. abs/2304.14535, 2023.
- [102] Y.-A. Chung *et al.*, “W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training,” in *Proc. ASRU*, 2021.
- [103] A. Conneau *et al.*, “Fleurs: Few-shot learning evaluation of universal representations of speech,” *ArXiv*, vol. abs/2205.12446, 2022.
- [104] N. Chomsky and M. Halle, *The Sound Pattern of English*. New York: Harper and Row, 1968.
- [105] G. N. Clements, “The geometry of phonological features,” *Phonology Yearbook*, pp. 225–252, 1985.
- [106] W. Li, S. M. Siniscalchi, N. F. Chen, and C.-H. Lee, “Improving non-native mispronunciation detection and enriching diagnostic feedback with dnn-based speech attribute modeling,” in *Proc. ICASSP*, 2016.
- [107] W. Li, “Improving mispronunciation detection and enriching diagnostic feedback for non-native learners of mandarin,” Ph.D. dissertation, Georgia Institute of Technology, Atlanta, GA, USA, 2019.

- [108] J. Lin *et al.*, “Improving mandarin tone recognition based on dnn by combining acoustic and articulatory features using extended recognition networks,” *Journal of Signal Processing Systems*, vol. 90, pp. 1077–1087, 2018.
- [109] I.-F. Chen, S. M. Siniscalchi, and C.-H. Lee, “Attribute based lattice rescoring in spontaneous speech recognition,” in *Proc. ICASSP*, 2014.
- [110] J. Li, Y. Tsao, and C.-H. Lee, “A study on knowledge source integration for candidate rescoring in automatic speech recognition,” in *Proc. ICASSP*, 2005.
- [111] C.-Y. Lin and H.-C. Wang, “Attribute-based mandarin speech recognition using conditional random fields,” in *Proc. Interspeech*, 2007.
- [112] V. Mitra *et al.*, “Joint modeling of articulatory and acoustic spaces for continuous speech recognition tasks,” in *Proc. ICASSP*, 2017.
- [113] V. Mitra, W. Wang, C. Bartels, H. Franco, and D. Vergyri, “Articulatory information and multiview features for large vocabulary continuous speech recognition,” in *Proc. ICASSP*, 2018.
- [114] S. M. Siniscalchi and C.-H. Lee, “An attribute detection based approach to automatic speech processing,” *Loquens*, 2014.
- [115] A. Graves, S. Fern´andez, F. J. Gomez, and J. Schmidhuber, “Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks,” in *Proc. ICML*, 2006.
- [116] F. Jelinek, “Continuous speech recognition by statistical methods,” *Proceedings of the IEEE*, vol. 64, no. 4, pp. 532–556, 1976.
- [117] J. B. Allen, *How do Humans Process and Recognize Speech?* R. P. Ramachandran and R. J. Mammone, Eds. Boston, MA: Springer US, 1995, pp. 251–275.
- [118] R. P. Lippmann, “Speech recognition by machines and humans,” *Speech Communication*, vol. 22, no. 1, pp. 1–15, 1997.
- [119] K. Kirchhoff, “Combining articulatory and acoustic information for speech recognition in noisy and reverberant environments,” in *Proc. ICSLP*, 1998.
- [120] H.-W. Hon, “Vocabulary-independent speech recognition: The vocind system,” Ph.D. dissertation, Carnegie Mellon University, Pittsburgh, PA, USA, 1992.
- [121] S. A. Liu, “Landmark detection for distinctive feature-based speech recognition,” Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, USA, 1996.

- [122] A. Juneja, O. Deshmukh, and C. Espy-Wilson, “An event-based acoustic-phonetic approach to speech segmentation and e-set recognition,” in *Proc. ICASSP*, 2002.
- [123] G. Fant, *Speech Sounds and Features*. The MIT Press, 1973.
- [124] P. Ladefoged and S. Disner, *Vowels and Consonants*. Wiley, 2012.
- [125] M. Tang, S. Seneff, and V. Zue, “Modeling linguistic features in speech recognition,” in *Proc. Eurospeech*, 2003.
- [126] S. Moran and D. McCloy, Eds., *PHOIBLE 2.0*. Jena: Max Planck Institute for the Science of Human History, 2019.
- [127] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An asr corpus based on public domain audio books,” in *Proc. ICASSP*, 2015.
- [128] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, “Mls: A large-scale multilingual dataset for speech research,” *ArXiv*, 2020.
- [129] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in *Proc. ICLR*, 2019.
- [130] H. Ji, T. Patel, and O. Scharenborg, “Predicting within and across language phoneme recognition performance of self-supervised learning speech pre-trained models,” *ArXiv*, 2022.
- [131] R. Ardila *et al.*, “Common voice: A massively-multilingual speech corpus,” in *Proc. LREC*, 2020.
- [132] A. Gulati *et al.*, “Conformer: Convolution-augmented transformer for speech recognition,” in *Proc. Interspeech*, 2020.
- [133] P. Ramachandran, B. Zoph, and Q. V. Le, “Searching for activation functions,” *ArXiv*, vol. abs/1710.05941, 2018.
- [134] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. Le, and R. Salakhutdinov, “Transformer-xl: Attentive language models beyond a fixed-length context,” in *Proc. ACL*, 2019.
- [135] Y. Dauphin, A. Fan, M. Auli, and D. Grangier, “Language modeling with gated convolutional networks,” in *Proc. ICML*, 2017.
- [136] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *CoRR*, vol. abs/1412.6980, 2014.
- [137] A. Vaswani *et al.*, “Attention is all you need,” in *Proc. NeurIPS*, 2017.

- [138] Y. Ganin *et al.*, “Domain-adversarial training of neural networks,” *Journal of Machine Learning Research*, vol. 17, no. 59, pp. 1–35, 2016.
- [139] T. Zeppenfeld and A. Waibel, “A hybrid neural network, dynamic programming word spotter,” in *Proc. ICASSP*, 1992.
- [140] J. Rohlicek, W. Russell, S. Roukos, and H. Gish, “Continuous hidden markov modeling for speaker-independent word spotting,” in *Proc. ICASSP*, 1989.
- [141] S. Bajpai and D. Radha, “Smart phone as a controlling device for smart home using speech recognition,” in *Proc. ICASSP*, 2019.
- [142] H. Kavya and V. Karjigi, “Sensitive keyword spotting for crime analysis,” in *Proc. NCCSN*, 2014.
- [143] T. Dinushika, L. Kavmini, P. Abeyawardhana, U. Thayasivam, and S. Jayasena, “Speech command classification system for sinhala language based on automatic speech recognition,” in *Proc. IALP*, 2019.
- [144] D. Seo, H.-S. Oh, and Y. Jung, “Wav2KWS: Transfer learning from speech representations for keyword spotting,” *IEEE Access*, pp. 1–1, 2021.
- [145] D. C. de Andrade, S. Leo, M. Viana, and C. Bernkopf, “A neural attention model for speech command recognition,” *ArXiv*, vol. abs/1808.08929, 2018.
- [146] S. Majumdar and B. Ginsburg, “Matchboxnet: 1d time-channel separable convolutional neural network architecture for speech commands recognition,” in *Proc. Interspeech*, 2020.
- [147] A. Berg, M. O’Connor, and M. T. Cruz, “Keyword Transformer: A Self-Attention Model for Keyword Spotting,” in *Proc. Interspeech*, 2021.
- [148] S. R. Shahamiri and S. S. B. Salim, “Artificial neural networks as speech recognisers for dysarthric speech: Identifying the best-performing set of mfcc parameters and studying a speaker-independent approach,” *Adv. Eng. Informatics*, vol. 28, pp. 102–110, 2014.
- [149] M. Mazumder *et al.*, “Multilingual spoken words corpus,” in *Proc. NeurIPS*, 2021.
- [150] K. P. F.R.S., “Liii. on lines and planes of closest fit to systems of points in space,” *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, pp. 559–572, 1901.

- [151] B. R. Chiswick and P. W. M. and, “Linguistic distance: A quantitative measure of the distance between english and other languages,” *Journal of Multilingual and Multicultural Development*, vol. 26, no. 1, pp. 1–11, 2005.
- [152] M. Dryer, M. Haspelmath, M.-P.-I. für Evolutionäre Anthropologie, and M. P. D. Library, *The World Atlas of Language Structures Online*. Max Planck Digital Library, 2011.
- [153] P. Gamallo, J. R. Pichel, and I. Alegria, “From language identification to language distance,” *Physica A: Statistical Mechanics and its Applications*, vol. 484, pp. 152–162, 2017.
- [154] K.-F. Lee and H.-W. Hon, “Large-vocabulary speaker-independent continuous speech recognition using hmm,” in *Proc. ICASSP*, 1988.
- [155] I.-F. Chen and C.-H. Lee, “A hybrid hmm/dnn approach to keyword spotting of short words,” in *Proc. Interspeech*, 2013.
- [156] R. Rose and D. Paul, “A hidden markov model based keyword recognition system,” in *Proc. ICASSP*, 1990.
- [157] J. Wilpon, L. Rabiner, C.-H. Lee, and E. Goldman, “Automatic recognition of keywords in unconstrained speech using hidden markov models,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 38, no. 11, pp. 1870–1878, 1990.
- [158] J. Mamou, B. Ramabhadran, and O. Siohan, “Vocabulary independent spoken term detection,” in *Proc. SIGIR*, 2007.
- [159] D. R. H. Miller *et al.*, “Rapid and accurate spoken term detection,” in *Proc. Interspeech*, 2007.
- [160] D. Vergyri *et al.*, “The sri/ogi 2006 spoken term detection system,” in *Proc. Interspeech*, 2007.
- [161] B. Wei *et al.*, “End-to-End Transformer-Based Open-Vocabulary Keyword Spotting with Location-Guided Local Attention,” in *Proc. Interspeech*, 2021.
- [162] I. Szöke *et al.*, “Comparison of keyword spotting approaches for informal continuous speech,” in *Proc. Interspeech*, 2005.
- [163] Y. Zhuang, X. Chang, Y. Qian, and K. Yu, “Unrestricted vocabulary keyword spotting using lstm-ctc,” in *Proc. Interspeech*, 2016.
- [164] T. Kawahara, C.-H. Lee, and B.-H. Juang, “Key-phrase detection and verification for flexible speech understanding,” in *Proc. ICSLP*, 1996.

- [165] R. Sukkar and C.-H. Lee, “Vocabulary independent discriminative utterance verification for nonkeyword rejection in subword based speech recognition,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 4, no. 6, pp. 420–429, 1996.
- [166] J. Richards, M. Ma, and A. Rosenberg, “Using word burst analysis to rescore keyword search candidates on low-resource languages,” in *Proc. ICASSP*, 2014.
- [167] M. Ma, J. Richards, V. Soto, J. Hirschberg, and A. Rosenberg, “Strategies for rescoring keyword search results using word-burst and acoustic features,” in *Proc. Interspeech*, 2014.
- [168] V. Soto, E. Cooper, L. Mangu, A. Rosenberg, and J. Hirschberg, “Rescoring confusion networks for keyword search,” in *Proc. ICASSP*, 2014.
- [169] V. Soto, L. Mangu, A. Rosenberg, and J. Hirschberg, “A comparison of multiple methods for rescoring keyword search lists for low resource languages,” in *Proc. Interspeech*, 2014.
- [170] O. Vinyals and S. Wegmann, “Chasing the metric: Smoothing learning algorithms for keyword detection,” in *Proc. ICASSP*, 2014.
- [171] A. E. Rosenberg, J. DeLong, C.-H. Lee, B.-H. Juang, and F. K. Soong, “The use of cohort normalized scores for speaker verification,” in *Proc. ICSLP*, 1992.
- [172] Y. Zigel and A. Cohen, “On cohort selection for speaker verification,” in *Proc. Interspeech*, 2003.
- [173] T. Kinnunen and E. Karpov, “Efficient online cohort selection method for speaker verification,” in *Proc. Interspeech*, 2004.
- [174] S. Settle, K. Audhkhasi, K. Livescu, and M. Picheny, “Acoustically grounded word embeddings for improved acoustics-to-word speech recognition,” in *Proc. ICASSP*, 2019.
- [175] Y. Hu, S. Settle, and K. Livescu, “Multilingual jointly trained acoustic and written word embeddings,” in *Proc. Interspeech*, 2020.
- [176] W. Jeon, “Acoustic neighbor embeddings,” *ArXiv*, vol. abs/2007.10329, 2022.
- [177] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” in *Proc. ICLR*, 2013.
- [178] J. Pennington, R. Socher, and C. Manning, “GloVe: Global vectors for word representation,” in *Proc. EMNLP*, 2014.

- [179] B. Shi, S. Settle, and K. Livescu, “Whole-word segmental speech recognition with acoustic word embeddings,” in *Proc. SLT*, 2020.
- [180] H. Yen and W. Jeon, “Improvements to embedding-matching acoustic-to-word asr using multiple-hypothesis pronunciation-based embeddings,” in *Proc. ICASSP*, 2022.
- [181] W. He, W. Wang, and K. Livescu, “Multi-view recurrent neural acoustic word embeddings,” in *Proc. ICLR*, 2017.
- [182] F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embedding for face recognition and clustering,” in *Proc. CVPR*, 2015.
- [183] J. G. Fiscus, J. Ajot, J. S. Garofolo, and G. Doddington, “Results of the 2006 spoken term detection evaluation,” in *Proc. SIGIR*, 2006.
- [184] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, “Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi,” in *Proc. Interspeech*, 2017.
- [185] G. Forney, “The viterbi algorithm,” *Proceedings of the IEEE*, vol. 61, no. 3, pp. 268–278, 1973.
- [186] J. K. Omura, “On the viterbi decoding algorithm,” *IEEE Trans. Inf. Theory*, vol. 15, no. 1, pp. 177–179, 1969.
- [187] A. Ganapathiraju, J. Hamaker, J. Picone, M. Ordowski, and G. Doddington, “Syllable-based large vocabulary continuous speech recognition,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 9, no. 4, pp. 358–366, 2001.
- [188] H. Wu and X. Wu, “Context-dependent syllable acoustic model for continuous chinese speech recognition,” in *Proc. Interspeech*, 2007.
- [189] Z. Qu, P. Haghani, E. Weinstein, and P. Moreno, “Syllable-based acoustic modeling with ctc-smbr-lstm,” in *Proc. ASRU*, 2017.
- [190] S. Zhou, L. Dong, S. Xu, and B. Xu, “A comparison of modeling units in sequence-to-sequence speech recognition with the transformer on mandarin chinese,” *ArXiv*, vol. abs/1805.06239, 2018.
- [191] S. Zhou, L. Dong, S. Xu, and B. Xu, “Syllable-based sequence-to-sequence speech recognition with the transformer in mandarin chinese,” in *Proc. Interspeech*, 2018.
- [192] J. Yuan, X. Cai, D. Gao, R. Zheng, L. Huang, and K. W. Church, “Decoupling recognition and transcription in mandarin asr,” in *Proc. ASRU*, 2021.

- [193] H. Kubozono, “The mora and syllable structure in Japanese: Evidence from speech errors,” *Language and Speech*, vol. 32, pp. 249–278, 1989.
- [194] L. M. Tomokiyo and K. Ries, “What makes a word: Learning base units in Japanese for speech recognition,” in *CoNLL97: Computational Natural Language Learning*, 1997.
- [195] N. Takahashi and S. Nakagawa, “Syllable recognition using syllable-segment statistics and syllable-based HMM,” in *Proc. ICSLP*, 2002.
- [196] H. Seki, K. Yamamoto, and S. Nakagawa, “Comparison of syllable-based and phoneme-based DNN-HMM in Japanese speech recognition,” in *Proc. ICAICTA*, 2014.
- [197] W. Hadamitzky and M. Spahn, *Romanization systems*, Wolfgang Hadamitzky: Japan-related Textbooks, Dictionaries, and Reference Works, 2005.
- [198] T. Akamatsu, *Japanese phonetics : theory and practice / Tsutomu Akamatsu* (LINCOM studies in Asian linguistics ; 3). München ; Lincom Europa, 1997.
- [199] J. Gao and T. Arai, “Acoustic cues of the stop voicing contrast in modern Tokyo Japanese,” in *Proc. Spring Meeting of the Acoustical Society of Japan*, 2018.
- [200] J. Ogata and Y. Ariki, “Syllable-based acoustic modeling for Japanese spontaneous speech recognition,” in *Proc. Eurospeech*, 2003.
- [201] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, “Aishell-1: An open-source Mandarin speech corpus and a speech recognition baseline,” in *Proc. O-COCOSDA*, 2017.
- [202] R. Sonobe, S. Takamichi, and H. Saruwatari, “JSUT corpus: Free large-scale Japanese speech corpus for end-to-end speech synthesis,” *ArXiv*, vol. abs/1711.00354, 2017.
- [203] S. Watanabe *et al.*, “ESPnet: End-to-end speech processing toolkit,” in *Proc. Interspeech*, 2018.
- [204] A. Lee, T. Kawahara, and K. Shikano, “Julius—an open source real-time large vocabulary recognition engine,” in *Proc. Eurospeech*, 2001.
- [205] A. Lee, T. Kawahara, and K. Shikano, “Recent development of open-source speech recognition engine Julius,” in *Proc. APSIPA*, 2009.
- [206] A. Conneau, A. Baevski, R. Collobert, A.-r. Mohamed, and M. Auli, “Unsupervised cross-lingual representation learning for speech recognition,” *ArXiv*, vol. abs/2006.13979, 2020.

## VITA

Hao Yen is currently pursuing a Ph.D. degree at the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA. He received his bachelor's degree in Electrical Engineering from National Taiwan University, Taipei, Taiwan, in 2020. He was a AIML ASR understanding intern at Apple Siri, Seattle, WA, in 2022, a research intern at Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, in 2022, and an applied scientist intern at Microsoft Research, Redmond, WA, in 2024.

The publications during his Ph.D. study at Georgia Institute of Technology are listed below.

1. **Hao Yen**, Shaoshi Ling, Guoli Ye, “Efficient Long-form Speech Recognition for General Speech In-context Learning,” *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.
2. Pin-Jui Ku, Chun-Wei Ho, **Hao Yen**, Sabato Marco Siniscalchi, Chin-Hui Lee, “An Explicit Consistency-Preserving Loss Function for Phase Reconstruction and Speech Enhancement,” *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.
3. **Hao Yen**, Pin-Jui Ku, Sabato Marco Siniscalchi, Chin-Hui Lee, “Language-Universal Speech Attributes Modeling for Zero-Shot Multilingual Spoken Keyword Recognition,” *Proc. Interspeech*, 2024.
4. **Hao Yen**, Sabato Marco Siniscalchi, Chin-Hui Lee, “Boosting End-to-End Multilingual Phoneme Recognition through Exploiting Universal Speech Attributes Constraints,” *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

5. **Hao Yen**, Pin-Jui Ku, Chao-Han Huck Yang, Hu Hu, Sabato Marco Siniscalchi, Pin-Yu Chen, Yu Tsao, “Neural Model Reprogramming with Similarity Based Mapping for Low-Resource Spoken Command Recognition,” *Proc. Interspeech*, 2023.
6. **Hao Yen**, Woojay Jeon, “Improvements to Embedding-Matching Acoustic-to-Word ASR Using Multiple-Hypothesis Pronunciation-Based Embeddings,” *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.
7. **Hao Yen**, François G. Germain, Gordon Wichern, Jonathan Le Roux, “Cold Diffusion for Speech Enhancement,” *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.
8. **Hao Yen**, Chao-Han Huck Yang, Hu Hu, Chin-Hui Lee, “A Lottery Ticket Hypothesis Framework for Low-Complexity Device-Robust Neural Acoustic Scene Classification,” *Proc. Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2021.