

Improving Zero-Shot Phonetic Classification through Language-Agnostic Articulatory Features

Ryo Magoshi¹, Jaeyoung Lee^{1,2}, Shinsuke Sakai¹, Tatsuya Kawahara¹

¹ Graduate School of Informatics, Kyoto University, Japan

² NTT, Inc., Japan

magoshi@sap.ist.i.kyoto-u.ac.jp, jalpr16@gmail.com, sakai@sap.ist.i.kyoto-u.ac.jp,
kawahara@i.kyoto-u.ac.jp

Abstract

Recent Phonetic Foundation Models (PFMs) for Speech-to-IPA transcription rely on Grapheme-to-Phoneme (G2P) labels, but the phoneme labels are not necessarily phonetically faithful. To investigate this issue, we evaluate zero-shot phonetic classification on Chinese aspiration and Japanese moraic nasals. A PFM trained on G2P-labeled data excluding these two languages yields poor accuracy on both tasks, showing that multilingual coverage with discrete IPA tokens is not sufficient for unseen settings. To overcome this limitation, we propose a classification method based on continuous Articulatory Feature (AF) vectors extracted from each frame. This AF-based approach outperforms discrete token-based methods, particularly for rare phones. We further show that it is crucial to adopt the optimal temporal aggregation of AF vectors for the target distinction: single-frame classification is best for aspiration, while segmental classification substantially improves nasal classification.

Index Terms: zero-shot phonetic recognition, IPA transcription, articulatory features

1. Introduction

Automatic transcription of speech into the International Phonetic Alphabet (IPA), or Speech-to-IPA, is an important task. In principle, the IPA's universality makes it possible to represent all human speech sounds within a single framework. Thus, Speech-to-IPA is designed to transcribe acoustic reality directly beyond the limits of language-specific phonemic inventories. This objective distinguishes it from standard ASR, which is limited by the orthographic rules of specific writing systems, and opens up several applications. First, it is essential for the analysis of atypical speech, such as the pronunciation of L2 learners [1, 2] and the evaluation of pathological speech [3]. Second, it contributes to transcription of endangered languages that lack a written form, aiding in cultural preservation [4, 5]. Third, sharing an IPA token set across languages facilitates knowledge transfer from high-resource to low-resource languages [6–11].

Speech-to-IPA research based on deep learning has progressed rapidly. This progress has led to the emergence of Phonetic Foundation Models (PFMs) that claim universality through scaling [12, 13]. However, these models encounter a critical issue regarding the phonetic validity of their training data. Most current datasets for Speech-to-IPA rely on Grapheme-to-Phoneme (G2P) conversion to generate labels. As a language-dependent processing, G2P maps graphemes to IPA symbols without observing the actual acoustic signal. This process inherently introduces conversion errors and inconsistent IPA notations across languages, making the labels phonemic

rather than phonetic. Consequently, acoustic differences that helps distinguishing phonemes in one language may be treated as ignorable variations in another. For example, a G2P system for a language that does not distinguish aspiration will ignore these acoustic nuances and assign an identical label to different phones. These characteristics conflict with the goal of Speech-to-IPA, preventing the model from acquiring truly universal articulatory recognition.

While existing PFMs claim zero-shot performance through multilingual training [12, 13], these evaluations often rely on unreliable G2P-derived data and language-specific training. To investigate this issue, this study first evaluates Speech-to-IPA models in a zero-shot setting. We select evaluation tasks where rule-based G2P produces reliable labels for the target contrasts, and show that current PFMs perform poorly when discriminating unseen phones. These findings demonstrate that multilingual coverage alone is not sufficient for universal recognition in zero-shot settings, even if the target IPA tokens are included in the training data and are shared across languages.

To overcome the limitations, we propose a classification framework based on 24-dimensional Articulatory Features (AF) estimated by a large-scale self-supervised model. The Articulatory Feature Classification Module (AFCM) is jointly trained with the IPA tokens prediction and provides a phonetic-level representation. Therefore, it realizes language-agnostic classification by examining relevant speech frames.

This study makes two contributions. First, we reveal the limitations of current multilingual training approaches through zero-shot classification on distinct phones. Second, we propose an AF-based classification framework tailored to the acoustic and phonetic nature of the target distinction.

2. Related Work

Automatic transcription of speech into phonetic or phonemic representations can be categorized into two approaches.

The first approach employs either the IPA, or language-specific symbols such as Hiragana for Japanese or Pinyin for Chinese within a specific linguistic framework [14–17]. These models can be accurate enough to serve as reliable front-ends for text-to-speech systems [15–17]. Consequently, they remain constrained by language-specific phonology and lack the capacity for universal transcription across diverse languages. Even when the IPA is utilized, selection of the symbol set is typically biased toward the phonemic inventory of the target language.

The second approach utilizes a common symbol set, primarily the IPA, to achieve language-universal phonetic recognition [10, 12, 13, 18, 19]. Speech-conditioned G2P systems, which refine grapheme-to-phoneme conversion using acoustic information, also fall into this category [20–22]. While this

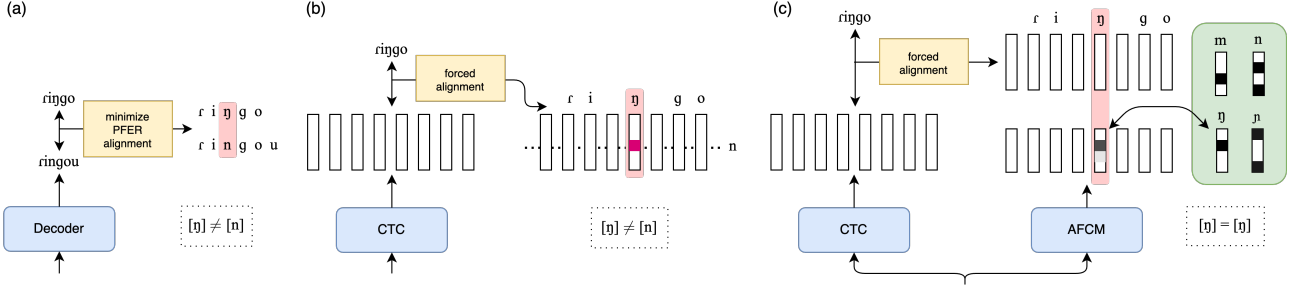


Figure 1: Overview of three classification methods, illustrated on a Japanese word “ringo” [riŋgo] (apple) for a nasal classification task. (a) Decoder-based classification aligns the decoded IPA sequence with the reference via PFER minimization. (b) CTC-based classification takes $\arg \max$ of the frame-level posteriors at the peak frame t^* identified by forced alignment. (c) AF-based classification (proposed) computes the L1 distance between the continuous AF vector at t^* and PanPhon template vectors, selecting the closest match.

approach can be applied to multilingual settings, it often suffers from degraded accuracy due to inconsistencies and errors in the underlying IPA labels. The primary bottleneck is that most large-scale datasets rely on G2P conversion tools rather than manual transcription, introducing the following issues:

1. Acoustic-Agnostic Labeling: Because G2P operates solely on orthographic input, the resulting labels represent abstract phonemes and fail to capture the natural phonetic variations present in actual speech.
2. Notational Divergence: Despite the IPA’s conceptual universality, notation conventions and symbol subsets vary significantly across languages and datasets, resulting in a lack of inter-dataset consistency.
3. Algorithmic Errors: G2P systems, whether rule-based [23–25], statistical [26–28] or neural [29–34], are prone to conversion errors that propagate into training labels.

These factors hamper phonetic faithfulness, hindering the development of models capable of universal recognition. On the other hand, AFs are phonetic descriptors that characterize each IPA symbol along dimensions such as nasality, voicing, and place of articulation, providing a continuous-valued and language-agnostic representation of speech sounds. To mitigate the limitations of discrete phonetic tokens, several studies have explored the use of AFs as an auxiliary or alternative representation for speech recognition.

Li et al. [35] utilized AFs as intermediate representations for grapheme-level output. However, their AFs are discrete and language-specific, precluding zero-shot application. Yen et al. [8] employed language-universal speech attribute constraints to improve multilingual phoneme recognition, demonstrating that AF-based regularization improves performance. Both approaches used AFs with only two dimensions, place and manner of articulation, which limits their phonetic expressiveness.

Magoshi et al. [10] adopted the AF system defined by PanPhon [36], which converts each IPA token into a 24-dimensional AF vector whose dimensions encode phonetic properties such as syllabic, sonorant, nasal, and continuant. Each dimension takes a value of +1 (present), −1 (absent), or 0 (not applicable), covering IPA symbols including those with diacritics such as aspirated stops. This representation is richer than the two-dimensional characterizations used in prior work. They proposed a frame-level AF classification method as an auxiliary task within a CTC-based Speech-to-IPA model, producing continuous AF vectors alongside discrete IPA posteriors. Our work builds upon [10] but uses the AF vectors directly for phonetic classification, rather than relying on discrete IPA token outputs.

3. Methods

We compare two Speech-to-IPA models (§3.1) and design three complementary classification methods to assess their zero-shot phonetic discrimination capability (§3.2).

3.1. Models

POWSM [13] is a hybrid CTC/Attention model [37]. An Attention-based Encoder-Decoder (AED) autoregressively generates an IPA token sequence, while a Connectionist Temporal Classification (CTC) [38] branch provides frame-level posterior distributions over the IPA token set. This architecture is representative of recent large-scale PFMs [13].

XLS-R + AFCM is a CTC-based model that uses XLS-R [39] as a pre-trained encoder, augmented with an Articulatory Feature Classification Module (AFCM) [10] which estimates a 24-dimensional AF vector at each frame in addition to the standard CTC posteriors. We propose using this AF output for zero-shot phonetic classification. The AFCM is trained with cross-entropy loss to predict the PanPhon [36] AF vector corresponding to each frame’s IPA token. The resulting logits constitute the continuous AF vectors used for classification. It is jointly trained with CTC-based IPA token recognition. Although the AF labels inherit errors from the IPA labels, the AFCM learns distributed phonetic properties rather than discrete tokens, making it more robust to individual labeling errors.

3.2. Classification Methods

To assess whether each model correctly identifies the target IPA symbol, we design three classification methods (Figure 1) that share a common two-step procedure: (1) locate the segment in which the target IPA is expected to occur, and (2) examine the model’s output within that segment to determine correctness. Decoder-based and CTC-based classifications serve as baselines that operate on discrete IPA token outputs.

Decoder-based classification aligns the decoded IPA sequence $\hat{\mathbf{Y}} = (\hat{y}_1, \dots, \hat{y}_N)$ to the reference sequence $\mathbf{Y}^* = (y_1^*, \dots, y_M^*)$ via $\pi^* = \arg \min_{\pi} \text{PFER}(\hat{\mathbf{Y}}, \mathbf{Y}^*; \pi)$, minimizing the Phoneme-Feature Error Rate (PFER) [18]. Here, π denotes an alignment mapping from the reference indices to the decoded indices. For the target position k , the prediction is deemed correct if the aligned token $\hat{y}_{\pi^*(k)}$ matches the ground truth y_k^* . If the alignment maps no decoded token at the position k (i.e., deletion), the prediction is treated as incorrect.

CTC-based classification utilizes the CTC log-posterior probabilities $\log p(c \mid \mathbf{x}_t)$, where c denotes an IPA token and

$\mathbf{x}_t$ is the encoder output at frame t . Given the interval (t_s, t_e) of the target segment obtained by forced alignment, the peak frame t^* is identified as:

$$t^* = \arg \max_{t_s \leq t \leq t_e} \log p(c^{\text{ref}} | \mathbf{x}_t) \quad (1)$$

where c^{ref} is the reference token. The estimated token is then $\hat{c} = \arg \max_c \log p(c | \mathbf{x}_{t^*})$.

Building on the CTC framework, we propose AF-based classification, which replaces discrete token posteriors with continuous AFs. Let $\mathbf{a}_t$ denote the AFCM output at frame t , and let t^* be the peak frame obtained as in CTC. We estimate the phonetic identity by computing the L1 distance between the AF vector $\mathbf{a}_{t^*}$ and the template vector derived from each candidate category $c \in \mathcal{C}$:

$$\hat{c} = \arg \min_{c \in \mathcal{C}} \sum_{i \in \mathcal{D}} |a_{t^*,i} - f_{c,i}| \quad (2)$$

where $\mathbf{f}_c$ is the PanPhon AF template vector for IPA token c , $\mathcal{C}$ is the set of candidate IPA tokens, and $\mathcal{D}$ is the set of AF dimensions that are discriminative among $\mathcal{C}$. For example, the four Japanese nasals [m], [n], [ŋ], [ɲ] share identical values on all AF dimensions except five place features (anterior, coronal, labial, high, and back), yielding $|\mathcal{D}| = 5$.

For both CTC-based and AF-based methods, we compare two temporal aggregation methods. The *single-frame* classification directly uses the representation at the peak frame t^* . The *segmental* classification instead averages the frame-level representations from t^* to the onset of the subsequent token:

$$\bar{\mathbf{v}} = \frac{1}{T_{\text{next}} - t^*} \sum_{t=t^*}^{T_{\text{next}}-1} \mathbf{v}_t \quad (3)$$

where T_{next} is the onset frame of the subsequent IPA token and $\mathbf{v}_t$ denotes the CTC log-posterior vector or the AF vector $\mathbf{a}_t$. Then, respective classification method is applied to $\bar{\mathbf{v}}$ in place of the single-frame representation.

4. Experiments

4.1. Experimental Settings

The two models are trained on the Common Voice [40] and FLEURS [41] portions of IPAPack++ [12], which provides G2P-derived IPA labels for multiple speech corpora. After excluding Chinese and Japanese, the training set covers 78 languages and approximately 3,000 hours. IPA sequences in the training data are tokenized using PanPhon [36], yielding a vocabulary of 280 tokens. Table 1 shows the language family distribution of the training data. The ‘‘Other’’ category groups six families each consisting of five or fewer languages: Austronesian, Uralic, Mongolic, Kartvelian, Constructed (Interlingua), and one language isolate (Basque).

POWSM follows the architecture of [13] but is trained from scratch only with the IPA tokens in our training set, resulting in 252M parameters due to the reduced vocabulary size.¹ XLS-R + AFCM combines a pre-trained XLS-R encoder with a CTC head and AFCM [10], totaling 318M parameters. Both models are trained with AdamW [42] at a learning rate of 5×10^{-4} , with 10k warmup steps and a batch size of 10 minutes.

¹The original POWSM is not used because its training data includes Chinese and Japanese, which are used for the zero-shot evaluation.

Table 1: Language family distribution of the training data.

Language family	# Lang.	Hours
Indo-European	40	2115.0
Atlantic-Congo	8	568.2
Dravidian	3	85.4
Turkic	6	79.9
Kra-Dai	2	46.2
Sino-Tibetan	3	35.5
Afro-Asiatic	3	16.4
Austroasiatic	1	8.2
Other (6 families)	12	45.1
Total	78	3000

Table 2: Distribution of target phones in the training data. ‘‘Ratio’’ denotes the proportion of aspirated tokens within each pair of voiceless stops (top) and the proportion of each phone in the nasal group (bottom).

Phone	Count	Ratio
[p ^h] [p]	192k 1.80M	9.6%
[t ^h] [t]	264k 5.12M	4.9%
[k ^h] [k]	267k 3.22M	7.7%
[m]	3.06M	31.6%
[n]	5.63M	58.3%
[ŋ]	722k	7.5%
[ɲ]	253k	2.6%

We design two zero-shot tasks with Chinese and Japanese, which are excluded from the training dataset. The first task classifies Chinese voiceless stops ([p^h]/[p], [t^h]/[t], [k^h]/[k]) as aspirated or unaspirated, evaluated on the Chinese FLEURS test set from IPAPack++ (652 utterances, 2.0 hours). The second task classifies Japanese moraic nasals ([m], [n], [ŋ], [ɲ]), evaluated on the CSJ [43] eval2 and eval3 (586 utterances, 52 minutes) with human-verified IPA labels [25]. In both cases, a rule-based G2P produces faithful labels: Chinese aspiration is a phonemic contrast and the Japanese nasals are conditional allophones which are predictable from the following consonant [44]. In AF-based classification, the PanPhon AF template vectors $\mathbf{f}_c$ are binarized from $\{-1, +1\}$ to $\{0, 1\}$ to match the binary output of the AFCM. For evaluation, we report balanced accuracy, defined as the macro-average of per-IPA recall.

Because phone frequencies in natural language are inherently imbalanced, the training data reflects this skew. Table 2 shows the distribution of the target phones in the training data. Aspirated stops account for fewer than 10% of the voiceless stop tokens at every place of articulation, and [ɲ] constitutes only 2.6% of nasal tokens.

4.2. Experimental Results

Table 3 presents the per-IPA recall and balanced accuracy for each task. POWSM yields poor accuracy on both tasks, indicating that training on G2P-labeled multilingual data is not sufficient for zero-shot phonetic discrimination. The proposed method of using XLS-R + AFCM with CTC achieves 94.5% on aspiration and 59.0% on nasal classification, demonstrating the benefit of finetuning with PanPhon-tokenized labels.

For aspiration (Table 3(a)), POWSM with CTC exhibits a strong bias toward unaspirated stops, recalling aspirated stops at only 10–20% while unaspirated stops exceed 98%. XLS-R

Table 3: Zero-shot phonetic classification results. (a) Three binary classifications of Chinese voiceless stops (aspirated vs. unaspirated) at each place of articulation. (b) A single four-way classification of Japanese nasals. All values are recalls in percentages and balanced accuracy is the macro-average of per-IPA recall. “AF segmental” denotes segmental AF classification.

(a) Chinese aspiration: binary classification per place of articulation									
Model	#Params	Method	[p ^h]	[p]	[t ^h]	[t]	[k ^h]	[k]	Balanced Acc.
POWSM	252M	Decoder	58.8	43.9	69.0	37.7	39.9	63.3	53.1
		CTC single-frame	10.2	99.5	16.1	99.0	12.7	98.8	56.7
		CTC segmental	11.9	99.3	20.5	98.3	16.1	98.5	58.3
XLS-R + AFCM	318M	CTC single-frame	92.7	98.4	90.9	98.8	89.3	95.7	94.5
		CTC segmental	82.5	96.5	96.2	80.5	85.0	91.9	87.5
		AF single-frame	93.2	97.9	91.4	99.3	93.6	94.1	95.4
		AF segmental	2.8	100.0	1.9	99.6	1.7	100.0	50.8

(b) Japanese nasal: four-way classification							
Model	#Params	Method	[m]	[n]	[ɲ]	[ɳ]	Balanced Acc.
POWSM	252M	Decoder	53.7	68.1	63.9	1.0	46.7
		CTC single-frame	73.1	87.2	32.7	1.0	48.5
		CTC segmental	74.6	88.4	30.2	4.2	49.3
XLS-R + AFCM	318M	CTC single-frame	65.7	97.9	69.3	3.1	59.0
		CTC segmental	62.7	99.7	46.5	6.2	53.8
		AF single-frame	74.6	92.8	83.2	12.5	65.8
		AF segmental	86.6	71.6	93.6	38.5	72.6

+ AFCM with CTC single-frame resolves this bias, achieving over 89% recall for all categories (balanced accuracy 94.5%). The single-frame AF method further improves to 95.4%, with over 91% recall across all categories. However, segmental AF collapses to near-zero recall on aspirated stops ([p^h] 2.8%, [t^h] 1.9%, [k^h] 1.7%), as the transient aspiration burst is diluted by averaging over the segmental window (Figure 2(a)).

For nasal classification (Table 3(b)), CTC-based methods achieve high recall on [m] and [n] but systematically misclassify [ɳ]. The advantage of AF is pronounced on the rare categories: for [ɳ], segmental AF achieves 38.5% recall, whereas all non-AF methods remain below 7%. Since AF labels provide supervision for phonetic features shared across many phonemes, the model can learn acoustic-articulatory mappings even for phones that are rare in the training data. For [ɲ], segmental AF reaches 93.6% compared to 69.3% with CTC single-frame. As shown in Figure 2(b), the place-of-articulation cues of [ɲ] persist throughout the segment, enabling segmental aggregation to improve AF-based balanced accuracy from 65.8% to 72.6%.

Overall, these results highlight two findings. First, temporal aggregation using XLS-R + AFCM outperforms existing discrete CTC posteriors, suggesting that existing methods can further be improved. Second, the effectiveness of the aggregation varies with the acoustic nature of the target distinction. Single-frame classification is preferable for temporally localized cues such as aspiration, while segmental aggregation is beneficial for sustained cues such as nasal place of articulation. The choice can be determined by the characteristics of target phones.

5. Conclusions

This study focuses on zero-shot phonetic classification on Chinese aspiration and Japanese moraic nasals. Our results show that relying solely on discrete IPA tokens fails to discriminate unseen phones, while classification using continuous AF vectors substantially improves performance, particularly for underrepresented categories. The optimal temporal aggregation varies with whether the target cue is transient or sustained.

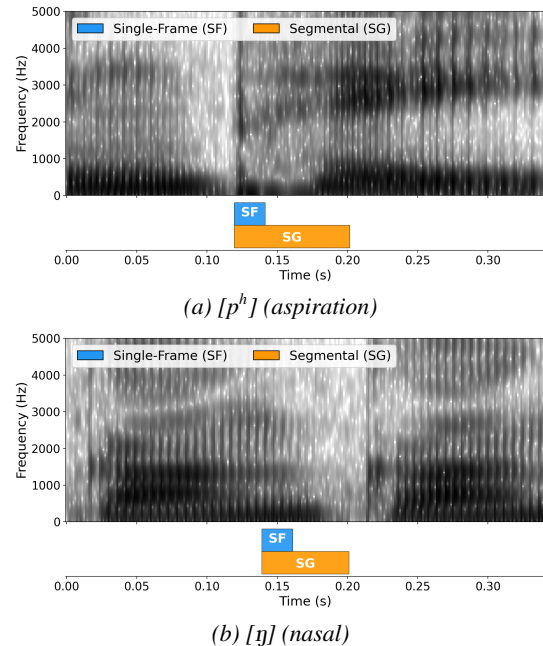


Figure 2: Spectrograms with single-frame (SF) and segmental (SG) windows. The aspiration burst in (a) is localized near S, while the nasal place cues in (b) persist across N.

These findings have implications beyond the two tasks studied here. First, classifying conditioned allophones such as Japanese moraic nasals shows that AF-based methods can capture sub-phonemic variation. Second, our framework can facilitate perceptual training of phones absent from the learner’s L1 without native-speaker supervision, as it operates in a language-agnostic manner using articulatory representations. Third, the proposed approach provides a step toward truly language-agnostic zero-shot ASR, where unseen phones can potentially be handled without language-specific adaptation.

6. Acknowledgments

This work was supported by JST NEXUS 251043539.

7. Generative AI Use Disclosure

The authors used generative AI tools for editing and polishing the manuscript. The authors take full responsibility for all content.

8. References

- [1] G. Zhao, S. Sonsaat, A. Silpachai, I. Lucic, E. Chukharev-Hudilainen, J. Levis, and R. Gutierrez-Osuna, “L2-ARCTIC: A Non-native English Speech Corpus,” in *Interspeech*, 2018, pp. 2783–2787.
- [2] M. Shahin and B. Ahmed, “Phonological-Level Mispronunciation Detection and Diagnosis,” in *Interspeech*, 2024, pp. 307–311.
- [3] L. J. Mallaband, “The agreement of phonetic transcriptions between paediatric speech and language therapists transcribing a disordered speech sample,” *International Journal of Language & Communication Disorders*, vol. 59, no. 5, pp. 1981–1995, 2024. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1111/1460-6984.13043>
- [4] B. Foley, J. Arnold, R. Coto-Solano, G. Durantin, T. M. Ellison, D. van Esch, S. Heath, F. Kratochvíl, Z. Maxwell-Smith, D. Nash, O. Olsson, M. Richards, N. San, H. Stoakes, N. Thieberger, and J. Wiles, “Building Speech Recognition Systems for Language Documentation: The CoEDL Endangered Language Pipeline and Inference System (ELPIS),” in *SLTU Workshop*, 2018, pp. 205–209.
- [5] J. Shi, J. D. Amith, R. Castillo García, E. Guadalupe Sierra, K. Duh, and S. Watanabe, “Leveraging end-to-end ASR for endangered language documentation: An empirical study on yolóxochitl Mixtec,” in *EACL*. Association for Computational Linguistics, 2021, pp. 1134–1145.
- [6] X. Li, S. Dalmia, J. Li, M. Lee, P. Littell, J. Yao, A. Anastasopoulos, D. R. Mortensen, G. Neubig, A. W. Black, and F. Metze, “Universal phone recognition with a multilingual allophone system,” in *ICASSP*, 2020, pp. 8249–8253.
- [7] Q. Xu, A. Baevski, and M. Auli, “Simple and Effective Zero-shot Cross-lingual Phoneme Recognition,” in *Interspeech*, 2022, pp. 2113–2117.
- [8] H. Yen, S. M. Siniscalchi, and C.-H. Lee, “Boosting end-to-end multilingual phoneme recognition through exploiting universal speech attributes constraints,” in *ICASSP*, 2024, pp. 11 876–11 880.
- [9] J. Lee, M. Mimura, and T. Kawahara, “Leveraging ipa and articulatory features as effective inductive biases for multilingual asr training,” in *ICASSP*, 2025, pp. 1–5.
- [10] R. Magoshi, S. Sakai, J. Lee, and T. Kawahara, “Multi-lingual and Zero-Shot Speech Recognition by Incorporating Classification of Language-Independent Articulatory Features,” in *Interspeech*, 2025, pp. 91–95.
- [11] M. Mimura, J. Lee, and T. Kawahara, “Switch Conformer with Universal Phonetic Experts for Multilingual ASR,” in *Interspeech*, 2025, pp. 1128–1132.
- [12] J. Zhu, F. Samir, E. Chodroff, and D. R. Mortensen, “ZIPA: A family of efficient models for multilingual phone recognition,” in *ACL*. Association for Computational Linguistics, 2025, pp. 19 568–19 585.
- [13] C.-J. Li, K. Chang, S. Bharadwaj, E. Yeo, K. Choi, J. Zhu, D. Mortensen, and S. Watanabe, “POWSM: A Phonetic Open Whisper-Style Speech Foundation Model,” 2026. [Online]. Available: <https://arxiv.org/abs/2510.24992>
- [14] M. Řezáčková, J. Švec, and D. Tihelka, “T5G2P: Using Text-to-Text Transfer Transformer for Grapheme-to-Phoneme Conversion,” in *Interspeech*, 2021, pp. 6–10.
- [15] H. Ohnaka, Y. Shirahata, B. Park, and R. Yamamoto, “Grapheme-Coherent Phonemic and Prosodic Annotation of Speech by Implicit and Explicit Grapheme Conditioning,” in *Interspeech*, 2025, pp. 2525–2529.
- [16] T. S. H. Mak, K. Y. Suen, and A. Y. Lam, “Speech-guided Grapheme-to-Phoneme Conversion for Cantonese Text-to-Speech,” in *Interspeech*, 2025, pp. 2535–2539.
- [17] R. Hu, X. Lin, J. Liu, S. Huang, and Z. Zhan, “Transcript-Prompted Whisper with Dictionary-Enhanced Decoding for Japanese Speech Annotation,” in *Interspeech*, 2025, pp. 2540–2544.
- [18] C. Taguchi, Y. Sakai, P. Haghani, and D. Chiang, “Universal Automatic Phonetic Transcription into the International Phonetic Alphabet,” in *Interspeech*, 2023, pp. 2548–2552.
- [19] J. L. Suchardt, H. El-Shazli, and P. Cassotti, “Towards language-agnostic STIPA: Universal phonetic transcription to support language documentation at scale,” in *EMNLP*. Association for Computational Linguistics, 2025, pp. 31 423–31 439.
- [20] J. Route, S. Hillis, I. Czeresnia Etinger, H. Zhang, and A. W. Black, “Multimodal, Multilingual Grapheme-to-Phoneme Conversion for Low-Resource Languages,” in *DeepLo Workshop*. Association for Computational Linguistics, 2019, pp. 192–201.
- [21] S. Ribeiro, G. Comini, and J. Lorenzo-Trueba, “Improving grapheme-to-phoneme conversion by learning pronunciations from speech recordings,” in *Interspeech*, 2023, pp. 999–1003.
- [22] H. Gao, M. Hasegawa-Johnson, and C. D. Yoo, “G2pu: Grapheme-to-phoneme transducer with speech units,” in *ICASSP*, 2024, pp. 10 061–10 065.
- [23] D. R. Mortensen, S. Dalmia, and P. Littell, “Epitran: Precision G2P for many languages,” in *LREC*. European Language Resources Association (ELRA), 2018.
- [24] M. Bernard and H. Titeux, “Phonemizer: Text to phones transcription for multiple languages in python,” *Journal of Open Source Software*, vol. 6, p. 3958, 2021.
- [25] Y. Matogawa, Y. Sakai, T. Watanabe, and C. Taguchi, “Japanese rule-based grapheme-to-phoneme conversion system and multilingual named entity dataset with international phonetic alphabet,” in *SIGMORPHON Workshop*. Association for Computational Linguistics, 2024, pp. 77–86.
- [26] M. Bisani and H. Ney, “Joint-sequence models for grapheme-to-phoneme conversion,” *Speech Communication*, vol. 50, no. 5, pp. 434–451, 2008.
- [27] A. Deri and K. Knight, “Grapheme-to-phoneme models for (almost) any language,” in *ACL*. Association for Computational Linguistics, 2016, pp. 399–408.
- [28] J. R. Novak, N. Minematsu, and K. Hirose, “Phonetisaurus: Exploring grapheme-to-phoneme conversion with joint n-gram models in the wfst framework,” *Natural Language Engineering*, vol. 22, no. 6, p. 907–938, 2016.
- [29] K. Rao, F. Peng, H. Sak, and F. Beaufays, “Grapheme-to-phoneme conversion using long short-term memory recurrent neural networks,” in *ICASSP*, 2015, pp. 4225–4229.
- [30] S. Yolchuyeva, G. Németh, and B. Gyires-Tóth, “Transformer Based Grapheme-to-Phoneme Conversion,” in *Interspeech*, 2019, pp. 2095–2099.
- [31] K. Park and S. Lee, “g2pM: A Neural Grapheme-to-Phoneme Conversion Package for Mandarin Chinese Based on a New Open Benchmark Dataset,” in *Interspeech*, 2020, pp. 1723–1727.
- [32] J. Zhu, C. Zhang, and D. Jurgens, “ByT5 model for massively multilingual grapheme-to-phoneme conversion,” in *Interspeech*, 2022, pp. 446–450.
- [33] X. Li, F. Metze, D. Mortensen, S. Watanabe, and A. Black, “Zero-shot learning for grapheme to phoneme conversion with language ensemble,” in *Findings of ACL*. Association for Computational Linguistics, 2022, pp. 2106–2115.

- [34] M. F. Qharabagh, Z. Dehghanian, and H. R. Rabiee, “Llm-powered grapheme-to-phoneme conversion: Benchmark and case study,” in *ICASSP*, 2025, pp. 1–5.
- [35] 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 *Interspeech*, 2019, pp. 2145–2149.
- [36] D. R. Mortensen, P. Littell, A. Bharadwaj, K. Goyal, C. Dyer, and L. Levin, “PanPhon: A resource for mapping IPA segments to articulatory feature vectors,” in *COLING*. The COLING 2016 Organizing Committee, 2016, pp. 3475–3484.
- [37] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, “Hybrid CTC/attention architecture for end-to-end speech recognition,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.
- [38] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks,” in *ICML*, vol. 2006, 2006, pp. 369–376.
- [39] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, “XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale,” in *Interspeech*, 2022, pp. 2278–2282.
- [40] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, “Common voice: A massively-multilingual speech corpus,” in *LREC*. European Language Resources Association, 2020, pp. 4218–4222.
- [41] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, and A. Bapna, “Fleurs: Few-shot learning evaluation of universal representations of speech,” in *SLT*, 2023, pp. 798–805.
- [42] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in *ICLR*, 2017.
- [43] K. Maekawa, “Corpus of spontaneous japanese: its design and evaluation,” in *ISCA/IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003, p. paper MMO2.
- [44] T. J. Vance, *The sounds of Japanese with audio CD*. Cambridge University Press, 2008.