Encoding of lexical tone in self-supervised models of spoken language

Gaofei Shen[1] **Michaela Watkins**[2] **Afra Alishahi**[1] **Arianna Bisazza**[3] **Grzegorz Chrupała**[1]
[1] Tilburg University [2] University of Amsterdam [3] University of Groningen
{g.shen, a.alishahi}@tilburguniversity.edu
m.m.watkins@uva.nl a.bisazza@rug.nl
grzegorz@chrupala.me

Abstract

Interpretability research has shown that self-supervised Spoken Language Models (SLMs) encode a wide variety of features in human speech from the acoustic, phonetic, phonological, syntactic and semantic levels, to speaker characteristics. The bulk of prior research on representations of phonology has focused on segmental features such as phonemes; the encoding of suprasegmental phonology (such as tone and stress patterns) in SLMs is not yet well understood. Tone is a suprasegmental feature that is present in more than half of the world's languages. This paper aims to analyze the tone encoding capabilities of SLMs, using Mandarin and Vietnamese as case studies. We show that SLMs encode lexical tone to a significant degree even when they are trained on data from non-tonal languages. We further find that SLMs behave similarly to native and non-native human participants in tone and consonant perception studies, but they do not follow the same developmental trajectory.

1 Introduction

Explaining the inner workings of self-supervised models of written and spoken language has been the focus of much recent work. Transformer-based (Vaswani et al., 2017) written language models have been shown to encode many types of linguistic information (Conneau et al., 2018; Hewitt and Manning, 2019). The analysis of self-supervised Spoken Language Models (SLMs) is also gaining traction: architectures such as wav2vec2 (Baevski et al., 2020) and HuBERT (Hsu et al., 2021) have been shown to encode linguistic information at the phonetic, phonological, syntactic and semantic levels of human speech without labeled data (Abdullah et al., 2021; Ma et al., 2021; de Seyssel et al., 2022; Bartelds et al., 2022; Martin et al., 2023; Shen et al., 2023; Pasad et al., 2024).

The majority of research on representations of phonetic and phonological information in SLMs focuses on the segmental level. **Segmental** refers to units of speech that do not spread but remain localized. Phonemes (e.g. vowels and consonants) are the smallest abstract units of sound that help to distinguish one unit from another (e.g. **p**at vs **b**at). **Suprasegmental**, in contrast, refers to features that are not necessarily limited to single units, but can spread across multiple phonemes or phrases. Examples include tone, stress patterns, and intonation, which can all entail syllable and phrase level changes (Singh and Fu, 2016). The representation of suprasegmental information in SLMs is important to study, as it is one of the main distinguishing features of speech compared to text: spoken utterances use suprasegmental cues to convey information that is sometimes not explicitly marked in a corresponding written sentence.[1] As a first step, in this work, we focus on lexical tone as a highly constrained, relatively well-understood example of a suprasegmental feature.

We firstly examine to what extent SLMs trained on tonal and non-tonal languages encode tone information in their internal representations. We find that SLMs are capable of capturing tonal information, regardless of whether they are trained on tonal or non-tonal languages.

Secondly, we investigate the impact of supervised fine-tuning on the automatic speech recognition (ASR) task. We find that fine-tuning *enhances* tone representations for models trained on tonal languages, but *reduces* them for models trained on non-tonal languages.

Thirdly, we explore whether SLMs exhibit the same perceptual patterns as native and non-native human listeners. We find that models show patterns similar to humans in discrimination of Mandarin tones and consonants, but find no evidence that they follow a similar developmental trajectory.

[1] Mandarin does not explicitly mark tone in writing, but Vietnamese does.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 4250–4261

2 Tones

Estimates suggest that more than 60% of the world's languages use some degree of tonal contrast (Yip, 2002). The primary focus of this work is on lexical tone, the process by which lexical items are distinguished from one another primarily by pitch cues (Chen et al., 2022). We focus primarily on pitch cues: while there are suggestions that articulatory cues alone are not sufficient for tonal distinction in languages such as Vietnamese, with socio-linguistic register also being important if present in the language in question (Brunelle, 2009), this is still subject to further research in perception and production studies to validate as the body of work remains small, particularly regarding perception. We therefore state here that non-tonal phonemic units (e.g. vowels, consonants) can be defined primarily by **non-pitch articulatory cues**, such as vowel height, voicing, and duration. In contrast, tonal units make use of **pitch cues**, with F0 (fundamental frequency) contour usually considered to be the primary cue (Rhee et al., 2021). In ambiguous contexts, other pitch cues can be used in combination with non-pitch cues such as amplitude, voice quality (e.g. breathy vs creaky), and spectral tilt (Rhee et al., 2021).

The Tone Bearing Unit (TBU) is the segment containing tone; this is typically, although not always, found on the syllable level. In the case of the primary language investigated in this paper, Mandarin, syllables are morpheme based (as in, Mandarin syllables *are* morphemes). Thus in Mandarin every morpheme contains obligatory tone, and is in of itself the TBU (Jun and Kubozono, 2020).

We compare SLMs trained on non-tonal languages as well as three fully lexical tonal languages: Mandarin, Cantonese and Vietnamese. The models are tested primarily on Mandarin data. Mandarin demonstrates full tonality, with tone found on each morpheme (Hyman, 2018), and has been widely studied for tone perception as well as acquisition. Secondarily, we also test on data from another lexical tone language, Vietnamese, to assess if our results generalize.

Mandarin Chinese is typically described as containing four (lexical) tones and one neutral tone that only occurs in unstressed syllables (Wu et al., 2020). The tones are conventionally assigned the labels 1-4 (T1-4); Figure 1 illustrates the four Mandarin tones.

Since one morpheme (one character) corresponds to one tone in Mandarin Chinese, we can use the Pinyin transcription to obtain our tone labels easily (see Figure 1 for notations); for example:

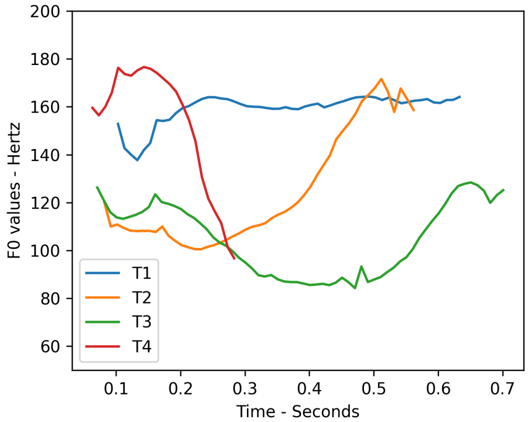

Figure 1: F0 contours of the four Mandarin tones measured from pronunciations recorded by one of the co-authors, a native speaker of Mandarin Chinese. The four syllables are pronounced in isolation (notation: mā T1, má T2, mǎ T3, mà T4).

(1) 今 天 天 气 很 好
Jīn tiān tiān qì hěn hǎo
T1 T1 T1 T4 T3 T3
'The weather today is very good.'

The tone label corresponds to the tone of the character when it is pronounced in isolation (base form). However, Mandarin features *tone sandhi*, i.e. the tone assigned to individual morphemes can change in pronunciation based on the tone of the adjacent morpheme (sandhi form). One instance of tone sandhi rules in Mandarin is T3 sandhi (Chen, 2000): if two T3 ('dipping') tones occur next to one another, the first will adjust to T2 ('rising') to avoid two consecutive T3 tones, as can be seen in examples 2 and 3, after Chen (2000). Tone labels obtained from Pinyin transcriptions only take the base form into account.

(2) 小
xiǎo
T3
'small'

	小	狗	
	xiǎo	gǒu	
(3)	T3	T3	*base form*
	T2	T3	*sandhi form*
	'small	dog'	

The primary pitch cue that distinguishes the individual Mandarin tones from each other is F0; however, secondary pitch cues are also present such as voice quality and spectral tilt (Belotel-Grenie and Grenie, 1994; Huang, 2020).

Vietnamese also has obligatory tones on every syllable, similar to Mandarin's morphemic TBU (Kirby, 2011). We adhere to the eight tone system described by Kirby (2011) in our experiment setup.

Cantonese is a Sinitic language related to Mandarin, and also features lexical tone, with six tonal distinctions (Zee, 1991) as opposed to Mandarin's four.

3 Related work

The present paper builds both on works interpreting the inner workings of SLMs and on experiments on perception of aspects of human speech.

3.1 Analyzing SLMs

The transformer architecture (Vaswani et al., 2017) has dominated the SLM realm. Researchers have developed many methods to analyze the inner-working of these models. Pasad et al. (2021) provide an overview on the variety of linguistic featurs encoded by self-supervised SLMs. The models tend to follow an autoencoder-like behavior with the middle layers showing the strongest encoding of a variety of linguistic features.

More recently, research has focused on specific properties of the input audio that is being encoded by the models. Martin et al. (2023) tested whether SLMs can distinguish between voiced and voiceless consonants. Shen et al. (2023) showed that self-supervised as well as visually-supervised SLMs are capable of encoding syntactic properties to some extent. Some prior works in the field have touched on the encoding of suprasegmental features in SSL speech models. Bartelds et al. (2022) showed the hidden state activations of SLMs are capable of capturing intonational and durational information on the phrase level, indicating that they can encode non-segmental information to a significant degree.

Many recent interpretability studies are inspired by psycholinguistics and child language development research. With the rise of probing and other interpretability methods, researchers replicated experimental paradigms in psychology and linguistics to better understand the capabilities of models compared to humans. For example, Wilcox et al. (2023) tested text language models using psycholinguistic experimental paradigms, showing that they are capable of learning syntactic dependencies with relatively little input data.

On the speech side, Lavechin et al. (2023) presented evidence that self-supervised SLMs can develop limited language-specific perception. Cruz Blandón et al. (2023) proposed comparing model behavior using checkpoints in the SLM pretraining process with data in child language development. They showed that computational language models can be a valuable resource in testing or confirming linguistic theories in the language development field. The methodology is mostly concerned with the overall learning of the language model in the output stage. Our work contributes to the explanation of the inner workings of SLMs.

3.2 Human perception experiments

In terms of tone perception, F0 is a clear primary cue (Ryant et al., 2014b; Rhee et al., 2021; Chen et al., 2022), but other secondary pitch cues serve to assist when speech is ambiguous and/or disrupted. Given that conversational speech contains non-trivial speech recognition difficulties such as e.g. tone sandhi and coarticulation, individual variation, and context omission (Ryant et al., 2014b), secondary cues play a role in the distinction of tones. An example of this is voice quality, where for example lowering F0 (introducing 'creaky' voice) increased perceptual saliency for T3, whereas T1 and T4 accuracy decreased and T2 remained unaffected (Huang, 2020; Chai, 2019; Kuang, 2017). This emphasises the fact that F0 does not operate in isolation, but that covariation between pitch and voice quality is inherent in Mandarin. Spectral cues (e.g. amplitude differences, spectral tilt) have also been suggested to be sufficient for adult speakers in tone production, while children are thought to hyperarticulate the tonal differences in speech (Rhee et al., 2021).

Suprasegmental cues appear to be preferred in experiments that compare segmental and suprasegmental cues against each other. Human infants are more sensitive to suprasegmental cues, with even newborns showing the same preference (Mehler et al., 1988; Nazzi et al., 1998). Several studies observe that tonal sensitivity develops earlier than perception of vowels and consonants (Xi et al., 2009; Yeung et al., 2013), with sensitivity to non-native tonal distinctions remaining longer than perception of non-tonal non-native phoneme categories (Liu and Kager, 2014; Shi et al., 2017).

Comparing vowels, consonants and tones, Singh et al. (2015) show that Mandarin learning children's sensitivity to consonants and vowels develop at a similar rate and shows departure from tones. The effect of tone mispronunciation is much larger than that of vowel or consonant mispronunciation for toddlers, but the pattern is reversed in preschoolers (Singh et al., 2015).

3.3 Automatic classification of tones

Automatic tone classification in Mandarin traditionally uses F0 contour and mel-frequency cepstral coefficients (MFCC) features. Advances in deep learning brought improvements in performance of tone classification. Ryant et al. (2014a) compare MFCC features and F0 contour as input to a neural tone classifier. MFCC features, while not explicitly encoding the F0 contour information, achieve an error rate of 15.56% for Tone 1-4 classification. The combination of MFCC features and F0 contours extracted with different methods did not see an improvement in the classifier's performance, indicating that the classifier was able to extract F0 contour from the MFCC features, or it was able to predict Mandarin tones reliably without F0 contour information. However, it is possible that the classifier was able to exploit associations between specific phonemes strings and tone labels, and hence avoid learning to detect tone based on pitch and voice quality cues.

After the introduction of self-supervised SLMs, Yuan et al. (2021) fine-tuned an English pre-trained wav2vec2 model (Baevski et al., 2020) for Mandarin tone classification and achieved a tone error rate of 6% on the same dataset as (Ryant et al., 2014a). Clearly, SLMs can handle the task of classifying Mandarin lexical tone with labeled fine-tuning. The aim of the present paper is not to compete with the existing implementations of Mandarin tone classifiers; rather we aim to uncover the tone encoding capabilities that emerge without explicit supervision.

4 Methodology

We use a number of wav2vec2-based (Baevski et al., 2020) models pre-trained and fine-tuned on datasets of different languages for our investigation. As examples of tonal languages, we choose Mandarin, Vietnamese and Cantonese, whereas English and French serve as non-tonal language examples. The models trained in the languages above are then tested on test data from Mandarin and Vietnamese.

To examine the encoding of tone, we train linear probing classifiers on the hidden state activations extracted from the aforementioned models for every morpheme in our testing datasets. [2]

4.1 Datasets

Training data. We examine SLMs that were trained on datasets of the following languages:

Mandarin pre-trained with AISHELL-2 (Du et al., 2018) and fine-tuned with AISHELL-1 (Bu et al., 2017). **English** pre-trained and fine-tuned with LibriSpeech (Panayotov et al., 2015). **Vietnamese** pre-trained with unlabelled YouTube audio and fine-tuned with the VLSP dataset for ASR (Nguyen, 2021). **Cantonese** pre-trained on a combined dataset of older Cantonese adult speech and YouTube audio (Huang and Mak, 2023). **French** pre-trained on MLS French (Pratap et al., 2020). Table 1 summarizes the characteristics of these datasets.

Test data. We primarily use the Mandarin Chinese THCHS-30 dataset (Wang and Zhang, 2015) for testing models' encoding of Mandarin tone. THCHS-30 consists of 30 hours of Mandarin speech recorded in a laboratory environment. The dataset is transcribed into both Chinese characters and Mandarin Pinyin. We also obtain character-level forced alignment with the Charsiu aligner (Zhu et al., 2022).

To test the generalizability of our results, we also use the Vietnamese VIVOS dataset (Luong and Vu, 2016), which consists of 15 hours of Vietnamese read speech recorded in a laboratory environment. The dataset is transcribed into Vietnamese orthography. We then convert the transcription into International Phonetic Alphabet (IPA) with tone labels with vPhon (Kirby, 2008). We use the Montreal Forced Aligner (McAuliffe et al., 2017) to obtain a syllable-level forced alignment.

Pre-training data. For the experiments on SLM's learning trajectory and perceptual patterns (see Section 5.3), we pre-train SLMs from scratch on the following datasets:

- MAGICDATA (Magic Data Technology Co., 2019), containing 755 hours of read Mandarin Chinese. The dataset was pre-split into a 712-hour training set and a 28-hour validation set.

[2] We release the codebase for our experiments at `https://github.com/techsword/tone-encoding-in-speech-model`

Training language	Tonality	Size (hours) Pre-training	Fine-tuning	Speech type
English (Baevski et al., 2020)	Non-tonal	960	960	Read
French (Parcollet et al., 2023)	Non-tonal	1,000	-	Read
Mandarin (Lu and Chen, 2022)	Tonal	1,000	178	Read
Vietnamese (Nguyen, 2021)	Tonal	13,000	250	YouTube audio/Read
Cantonese (Huang and Mak, 2023)	Tonal	2,800	-	Spotaneous + Read

Table 1: Description of the datasets used in pre-training/fine-tuning models.

- LibriSpeech (Panayotov et al., 2015), see details in Table 1. We split a subset of the LibriSpeech dataset into a 710-hour training set and a 29-hour validation set.

4.2 Spoken Language Models

Architecture. With the exception of the Cantonese model, all models investigated in this paper are based on the base configuration of wav2vec2 (Baevski et al., 2020). Wav2vec2-base consists of five convolutional feature encoder and twelve transformer layers. The feature encoder processes the audio waveform input into latent speech representations, and the transformer layers encode the feature encoder output into contexual representations. The wav2vec2-base models has 95M parameters. The Cantonese model uses the wav2vec2-conformer architecture with 180M parameters.

Training objectives. The fully self-supervised *pre-training* objective in wav2vec2 consists in discriminating between the matched and unmatched segment representations for a masked portion of the latent speech representation. The ASR *fine-tuning* objective consists in transcribing the audio input into output tokens in the orthography of the target language and is realized by adding a linear layer on top of a pre-trained wav2vec2 model.

Checkpoints. For the experiments in Section 5.3 we pre-train two SLMs with the fairseq toolkit (Ott et al., 2019) on LibriSpeech for English and MAGICDATA for Mandarin; we train both models for 85,000 steps using 8 Nvidia A100-40GB GPU with update frequency = 8 to simulate training with 64 GPUs. Each model finished training in approximately 96 hours. We save checkpoints every 5,000 steps.

4.3 Probing classifiers

Preprocessing. We follow previous work (Ryant et al., 2014a) in removing segments transcribed with the neutral tone from the Mandarin tone classification task. Mandarin neutral tones primarily appear in unstressed syllables (cf. Section 2) and hence are more susceptible to variations.

Language	Split	Samples
Mandarin	Train	223,851
Mandarin	Test	45,772
Vietnamese	Train	124,248
Vietnamese	Test	29,629

Table 2: Train/test splits for the tone probing classifier, for the Mandarin and Vietnamese data.

Generating classifier input. We extract the hidden state activations of models as a response to audio samples in the test data. We average-pool the hidden state output corresponding to the duration of individual syllables to obtain a vector using forced alignment timestamps. The resulting 768-dimensional vectors are input to the classifiers. To control for the influence of lexical cues on tone detection, we construct an exclusive train-test-split such that phoneme strings appearing in the test set do not appear in the training set. This setup prevents the probing classifier from exploiting associations between tones and phoneme sequences. We employ a randomized 80:20 train-test split with the split sizes shown in Table 2.

F0 and MFCC baselines. We closely follow Ryant et al. (2014a) and use F0 contours and 40-dimensional mel-frequency cepstral coefficients (MFCC) features as baselines. We use Librosa (McFee et al., 2023) to extract the MFCC features and Praat (Jadoul et al., 2018; Boersma and Weenink, 2021) to extract the F0 contours from the audio samples. We then find the center frame for each word using the alignment timestamps and concatenate all frames in a 21 frame window (10-

Split	Samples
Train	92,413
Test	15,688

Table 3: Train/test split for the consonant probing classifier, for the Mandarin data.

1-10) for both F0 and MFCC features. We end up with a 21-dimensional vector for F0 contours and 840-dimensional vector for MFCC features as our baseline classifier inputs.

Text baseline. In addition to audio baselines, we also include a text-based transformer model in our comparison. BERT (Devlin et al., 2019) serves as a reference point to show how much information is encoded in the speech signal as opposed to what can be guessed from pure text. We use a Chinese pre-trained BERT[3] that encodes Chinese characters into vectors. We extract per-word hidden state outputs with a resulting 768-dimensional vector.

Tone classifiers. We use the syllable activation vectors as input to a Ridge linear classifier that predicts the lexical tone of the input morpheme. We select the final model via 5-fold cross-validation, and report the classification accuracy on the test split. The regularization strength α was tuned for values $\{10^n \mid n \in \{-4, -3, -2, -1, 0, 1, 2\}\}$.

Consonant classifiers. When comparing tone to consonant classification, we employ the same classifier setup for consonant and replicate the perception experiment in Wang and Chen (2020) in Section 5.3.1. Since we only investigate consonants that appear solely in the onset position and the rest of the phonemes are not relevant to our task, we use the same syllable vectors as above instead of obtaining a phoneme vector with using phoneme level alignment. We construct exclusive train-test-splits that contain unique rimes (nucleus + coda) of the syllables. Specific details of the train/test split for this experiment can be found in Table 3.

5 Results

In this section, we present a series of experiments for analyzing the encoding of tone in SLMs.

5.1 Tone encoding across languages

Figure 2 shows the tone classification accuracy using the layer-wise representations of all models

[3]https://huggingface.co/bert-base-chinese

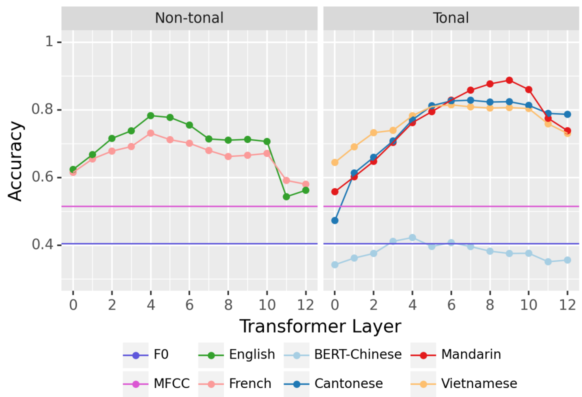

Figure 2: Classification accuracy of Mandarin lexical tones using layer-wise representations from models pre-trained on tonal and non-tonal languages.

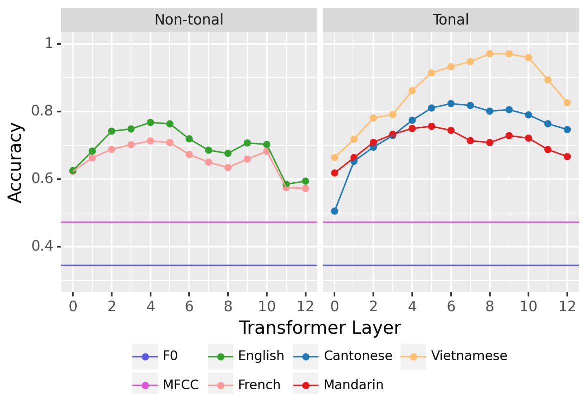

Figure 3: Classification accuracy of Vietnamese lexical tones with hidden-state activations from models pre-trained on tonal and non-tonal languages.

pre-trained on non-tonal (left) versus tonal (right) languages. We see that all layers of all models perform better than the F0 and MFCC baselines, which themselves outperform the text-based BERT baseline. The classification accuracy for tonal language models is overall higher, and increases in the higher layers of the models. Models trained on non-tonal languages also show substantial encoding of tone; but remarkably, there is a substantial drop in classification accuracy in their final layers while the corresponding decrease is much less pronounced in tonal language models.

We repeat the tone classification experiment for Vietnamese tones. Results in Figure 3 show the Cantonese model performs slightly better than the English model, especially towards the later layers; the Mandarin model, however, patterns similar to the English model. This is could be due to the fact that Mandarin has fewer tonal contrasts than Vietnamese and Cantonese (cf. Section 2), or, more likely, that Vietnamese uses different acoustic cues such as phonation type and voice quality in tonal perception than f0 contours or height in Mandarin (Brunelle, 2009).

Studies on human participants show that speakers of other tonal languages can perform better at identifying Mandarin lexical tones compared to non-tonal language speakers (So and Best, 2010), but other literature suggest that listeners are more sensitive to specific cues in tonal perception that are present in their native languages (DiCanio, 2012; Schaefer and Darcy, 2014). The SLMs we tested show the former pattern for Mandarin tone classification. Regarding Vietnamese tones, the result is more equivocal suggesting that Cantonese tone representations generalize to Vietnamese to some extent, while Mandarin ones do not.

5.2 Impact of ASR fine-tuning

We examine how fine-tuning for ASR impacts the encoding of tone in SLMs. Since tonal information is crucial for correctly transcribing tonal language input, SLMs fine-tuned for tonal languages are expected to perform better at our tone classification task. Figure 4 compares the tone classification accuracy of the English and Mandarin models, pre-trained only (left) versus pre-trained and then fine-tuned (right); Figure 5 shows the corresponding results for English and Vietnamese.

We find that fine-tuning affects the encoding of tone for non-tonal vs tonal language models in opposite ways: classification accuracy benefits from fine-tuning for Mandarin, but is harmed by it for the English model. The same pattern holds for English vs Vietnamese on the Vietnamese tone data in Figure 5.

These results likely reflect the fact that ASR fine-tuning encourages the SLM to increase its specialization in identifying the language-specific information needed to output the written form of the language. Tonal information may not contribute much to this objective in non-tonal languages, and thus fine-tuning would tend to remove it. In tonal-language ASR however, tone information may be crucial to correctly transcribe the input audio, for example, when disambiguating Mandarin syllables that consist of the same segmental phonemes and only differ in tone, in order to output the correct Chinese character.

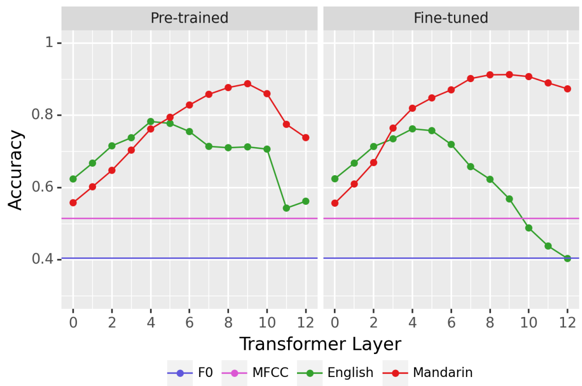

Figure 4: Classification accuracy of Mandarin lexical tones using layer-wise representations from models pre-trained and fine-tuned on Mandarin and English.

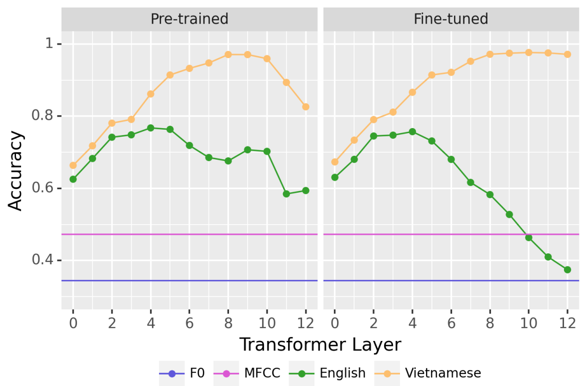

Figure 5: Classification accuracy of Vietnamese lexical tones using layer-wise representations from models pre-trained and fine-tuned on Vietnamese and English.

5.3 Comparison to human perception

In this section we report the results motivated by tone and consonant perception patterns in humans.

5.3.1 Learning trajectory

Children have a higher sensitivity to tone than consonant distinctions early on. For children speaking a non-tonal language, this sensitivity towards tone continues longer than sensitivity towards non-native segmental features, i.e. consonants and vowels (Shi et al., 2017; Liu and Kager, 2014).

Here we aim to determine the corresponding learning trajectory in SLMs by testing them during pre-training. Figure 6 shows the accuracy of classifying Mandarin consonants and tones in the best performing layer of SLMs trained on English and Mandarin as a function of the number of training steps.

Although we observe classification accuracy of the SLMs quickly surpasses the F0 and MFCC baselines after 10,000 steps, we do not detect an obvious difference in the overall pattern between the case of consonants and tones. This suggest that SLMs do not follow the same differential trajectory as children, at least as measured via our methodology.

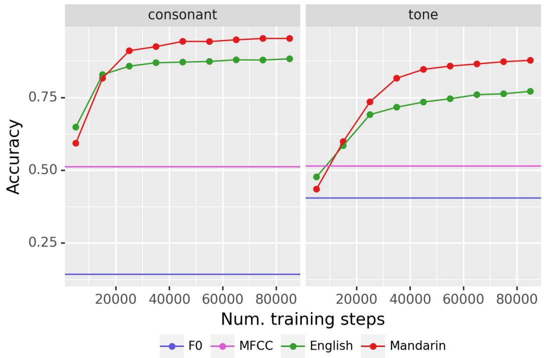

Figure 6: Classification accuracy of Mandarin lexical tones versus consonants for models pre-trained on English and Mandarin.

5.3.2 Tone and consonant contrasts

Non-native speakers can have difficulty distinguishing between T2-T3 and T1-T4 tone pairs in Mandarin (Hao, 2012). While native adult listeners have shown near perfect tone identification accuracy (So and Best, 2010; Tsukada and Kondo, 2019), there is literature documenting native Mandarin-learning toddlers and also some adults having more difficulty in distinguishing the tone pair T2-3 compared to other combinations of Mandarin lexical tone. Huang and Johnson (2011) showed that the tone pair T2-3 is the most confusable for both native speakers of American English and Mandarin while T1-4 also being confusable for native Mandarin speakers. We investigate this pattern in pre-trained SLMs via a dedicated probing experiment, using the final (85,000 steps) checkpoint of the pre-trained models in Section 5.3.1. As can be seen in Figure 7, tone pairs T1-4 and T2-3 show the largest differences in the best classification accuracy between the Mandarin and English models, which roughly matches human perceptual pattern.

We complement the results on the development of tone contrast with a parallel experiment on those Mandarin consonant contrasts which are challenging for speakers of English. Each member of a contrasting group is perceived as the same phoneme by English speakers due to perceptual assimilation (Wang and Chen, 2020). Table 4 displays the resulting mapping to English phoneme categories.

	Mandarin		English	
Group	Pinyin	IPA	Alphabet	IPA
1	sh, x	ʂ, ɕ	sh	ʃ
2	ch, zh, q	tʂ^{h}, tʂ, tɕ^{h}	ch	tʃ
3	s, z, c	s, ts, ts^{h}	s	s

Table 4: Perceptual mapping of Mandarin consonants onto English consonants (Wang and Chen, 2020).

Figure 8 shows that accuracy for consonant groups 2 and 3 match closely for the two models. Group 1 shows a discrepancy, possibly due to the potential mapping of Mandarin x /ɕ/ into two English consonants sh /ʃ/ and z /z/, as hypothesized by Wang and Chen (2020).

6 Conclusion

We analyze the tone encoding capabilities of spoken language models trained on three tonal and two non-tonal languages, using classifier probes with data from two tonal languages: Mandarin and Vietnamese. We find that SLMs trained on either tonal or non-tonal languages encode tonal information in Mandarin and Vietnamese to a significant degree.

We also find that fine-tuning for the speech recognition task enhances the tone encoding capabilities of models trained on tonal languages but reduces them for models trained on non-tonal languages. While we see evidence suggesting that the learning trajectories of SLMs in pre-training do not follow the same developmental trajectories found in human language acquisition, we find that SLMs show patterns similar to that of human listeners in tone and consonant perception experiments.

While this paper focused on investigating the encoding of lexical tones with Mandarin and Vietnamese as case studies, Mandarin and Vietnamese are in no way representative of all the diverse tonal languages that exist in the world. The encoding of other suprasegmental features such as stress patterns and intonation is equally important to study in future work. We hope that explaining self-supervised models of spoken languages provides a unique perspective for us to contribute to a better understanding of how languages work. Given the rise of SLMs trained with multilingual datasets (Conneau et al., 2021; Radford et al., 2022), it would be interesting to investigate if the multilingual SLMs encode segmentals and suprasegmentals from different languages more robustly. This paper serves as a starting point in the research of

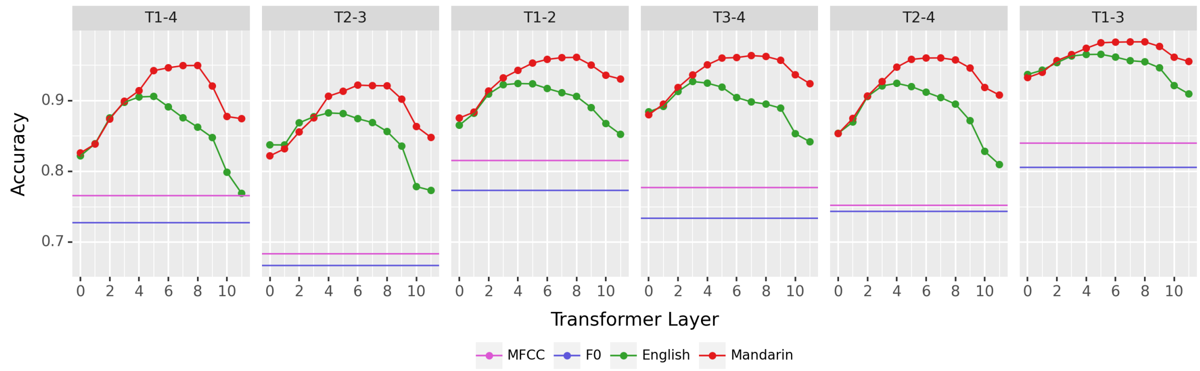

Figure 7: Binary classification accuracy for Mandarin tonal pairs, for English and Mandarin models.

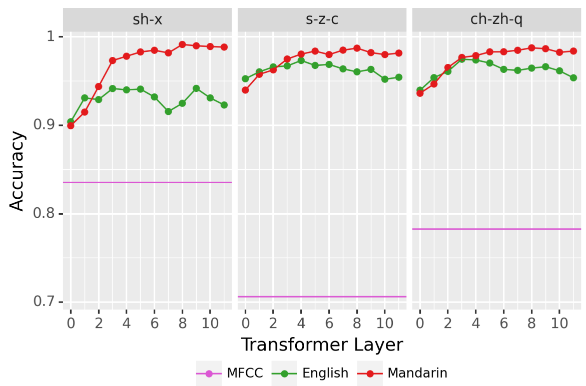

Figure 8: Classification accuracy for Mandarin consonant groups, for English and Mandarin models. The F0 baseline with its much lower classification accuracy is omitted from this figure for clarity.

the encoding of suprasegmental cues in spoken language models.

7 Limitations

We selected SLMs based on the wav2vec2 architecture in our experimental design, but we acknowledge that the training data of the models selected is quite varied in their size and quality (noisy vs clean speech) as described in Section 4.1. This is partially due to the scarce availability of (high-quality) speech data for underrepresented languages, especially the many tonal languages of the world. Hence SLMs pre-trained on monolingual datasets of these languages are also sparse. The Vietnamese wav2vec2 model (Nguyen, 2021) was trained on a significantly larger amount of data (13k hours) than the other models tested (around 1000 hours). It is possible that in addition to the inclusion of tonal languages in training, the amount of training data also played a role in increasing the tonal encoding capabilities of SLMs. However, literature has shown that more training data does not always have a positive impact on the models performance if the additional data is noisy (Parcollet et al., 2023). At the same time, we note that the Cantonese model (Huang and Mak, 2023), in addition to being pre-trained on a larger dataset, is also different in architecture. Additionally, our use of two read speech datasets as test data does not fully reflect the linguistic diversity of different accents and dialects in Mandarin and Vietnamese. Future work needs to go wider and deeper in both model architecture and dataset diversity in order to uncover more generalizable patterns in different languages.

8 Acknowledgements

This publication is part of the project *InDeep: Interpreting Deep Learning Models for Text and Sound* (with project number NWA.1292.19.399) of the National Research Agenda (NWA-ORC) program. Funding by the Dutch Research Council (NWO) is gratefully acknowledged. We also thank SURF (www.surf.nl) for the support in using the National Supercomputer Snellius.

References

Badr M. Abdullah, Iuliia Zaitova, Tania Avgustinova, Bernd Möbius, and Dietrich Klakow. 2021. How Familiar Does That Sound? Cross-Lingual Representational Similarity Analysis of Acoustic Word Embeddings.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations.

Martijn Bartelds, Wietse de Vries, Faraz Sanal, Caitlin Richter, Mark Liberman, and Martijn Wieling. 2022. Neural representations for modeling variation in speech. *Journal of Phonetics*, 92:101137.

Agnes Belotel-Grenie and Michel Grenie. 1994. Phonation types analysis in standard chinese. In *3rd International Conference on Spoken Language Processing, ICSLP 1994*, pages 343–346. The International Society for Computers and Their Applications (ISCA).

Paul Boersma and David Weenink. 2021. Praat: Doing phonetics by computer [Computer program].

Marc Brunelle. 2009. Tone perception in northern and southern vietnamese. *Journal of Phonetics*, 37(1):79–96.

Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. 2017. AISHELL-1: An Open-Source Mandarin Speech Corpus and A Speech Recognition Baseline.

Yuan Chai. 2019. The source of creak in Mandarin utterances. In *Proceedings of ICPhS 2019*.

Matthew Y. Chen. 2000. *Tone Sandhi: Patterns across Chinese Dialects*. Cambridge Studies in Linguistics. Cambridge University Press, Cambridge.

Yue Chen, Yingming Gao, and Yi Xu. 2022. Computational Modelling of Tone Perception Based on Direct Processing of f0 Contours. *Brain Sciences*, 12(3):337.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised Cross-Lingual Representation Learning for Speech Recognition. In *Interspeech 2021*, pages 2426–2430. ISCA.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

María Andrea Cruz Blandón, Alejandrina Cristia, and Okko Räsänen. 2023. Introducing Meta-analysis in the Evaluation of Computational Models of Infant Language Development. *Cognitive Science*, 47(7):e13307.

Maureen de Seyssel, Marvin Lavechin, Yossi Adi, Emmanuel Dupoux, and Guillaume Wisniewski. 2022. Probing phoneme, language and speaker information in unsupervised speech representations. In *Interspeech*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Christian T. DiCanio. 2012. Cross-linguistic perception of Itunyoso Trique tone. *Journal of Phonetics*, 40(5):672–688.

Jiayu Du, Xingyu Na, Xuechen Liu, and Hui Bu. 2018. AISHELL-2: Transforming Mandarin ASR Research Into Industrial Scale.

Yen-Chen Hao. 2012. Second language acquisition of Mandarin Chinese tones by tonal and non-tonal language speakers. *Journal of Phonetics*, 40(2):269–279.

John Hewitt and Christopher D. Manning. 2019. A Structural Probe for Finding Syntax in Word Representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units.

Ranzo Huang and Brian Mak. 2023. Wav2vec 2.0 ASR for Cantonese-Speaking Older Adults in a Clinical Setting. In *INTERSPEECH 2023*, pages 4958–4962. ISCA.

Tsan Huang and Keith Johnson. 2011. Language Specificity in Speech Perception: Perception of Mandarin Tones by Native and Nonnative Listeners. *Phonetica*, 67(4):243–267.

Yaqian Huang. 2020. Different attributes of creaky voice distinctly affect Mandarin tonal perception. *The Journal of the Acoustical Society of America*, 147(3):1441–1458.

Larry M. Hyman. 2018. What tone teaches us about language. *Language*, 94(3):698–709.

Yannick Jadoul, Bill Thompson, and Bart de Boer. 2018. Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, 71:1–15.

Sun-Ah Jun and Haruo Kubozono. 2020. Asian Pacific Rim. In Carlos Gussenhoven and Aoju Chen, editors, *The Oxford Handbook of Language Prosody*, Oxford Handbooks. Oxford University Press, Oxford, New York.

James Kirby. 2008. vPhon: A Vietnamese phonetizer (version 2.1.1).

James P. Kirby. 2011. Vietnamese (hanoi vietnamese). *Journal of the international phonetic association*, 41(3):381–392.

Jianjing Kuang. 2017. Covariation between voice quality and pitch: Revisiting the case of Mandarin creaky voice. *The Journal of the Acoustical Society of America*, 142(3):1693–1706.

Marvin Lavechin, Maureen De Seyssel, Marianne Métais, Florian Metze, Abdelrahman Mohamed, Hervé Bredin, Emmanuel Dupoux, and Alejandrina Cristia. 2023. Statistical learning models of early phonetic acquisition struggle with child-centered audio data.

Liquan Liu and René Kager. 2014. Perception of tones by infants learning a non-tone language. *Cognition*, 133(2):385–394.

Ke-Han Lu and Kuan-Yu Chen. 2022. A context-aware knowledge transferring strategy for CTC-based ASR.

Hieu-Thi Luong and Hai-Quan Vu. 2016. A non-expert Kaldi recipe for Vietnamese speech recognition system. In *Proceedings of the Third International Workshop on Worldwide Language Service Infrastructure and Second Workshop on Open Infrastructures and Analysis Frameworks for Human Language Technologies (WLSI/OIAF4HLT2016)*, pages 51–55, Osaka, Japan. The COLING 2016 Organizing Committee.

Danni Ma, Neville Ryant, and Mark Liberman. 2021. Probing Acoustic Representations for Phonetic Properties. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 311–315.

Ltd. Magic Data Technology Co. 2019. MAGICDATA Mandarin Chinese Read Speech Corpus.

Kinan Martin, Jon Gauthier, Canaan Breiss, and Roger Levy. 2023. Probing Self-supervised Speech Models for Phonetic and Phonemic Information: A Case Study in Aspiration. In *INTERSPEECH 2023*, pages 251–255. ISCA.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Interspeech 2017*, pages 498–502. ISCA.

Brian McFee, Matt McVicar, Daniel Faronbi, Iran Roman, Matan Gover, Stefan Balke, Scott Seyfarth, Ayoub Malek, Colin Raffel, Vincent Lostanlen, Benjamin Van Niekirk, Dana Lee, Frank Cwitkowitz, Frank Zalkow, Oriol Nieto, Dan Ellis, Jack Mason, Kyungyun Lee, Bea Steers, Emily Halvachs, Carl Thomé, Fabian Robert-Stöter, Rachel Bittner, Ziyao Wei, Adam Weiss, Eric Battenberg, Keunwoo Choi, Ryuichi Yamamoto, CJ Carr, Alex Metsai, Stefan Sullivan, Pius Friesch, Asmitha Krishnakumar, Shunsuke Hidaka, Steve Kowalik, Fabian Keller, Dan Mazur, Alexandre Chabot-Leclerc, Curtis Hawthorne, Chandrashekhar Ramaprasad, Myungchul Keum, Juanita Gomez, Will Monroe, Viktor Andreevitch Morozov, Kian Eliasi, Nullmightybofo, Paul Biberstein, N. Dorukhan Sergin, Romain Hennequin, Rimvydas Naktinis, Beantowel, Taewoon Kim, Jon Petter Åsen, Joon Lim, Alex Malins, Darío Hereñú, Stef Van Der Struijk, Lorenz Nickel, Jackie Wu, Zhen Wang, Tim Gates, Matt Vollrath, Andy Sarroff, Xiao-Ming, Alastair Porter, Seth Kranzler, Voodoohop, Mattia Di Gangi, Helmi Jinoz, Connor Guerrero, Abduttayyeb Mazhar, Toddrme2178, Zvi Baratz, Anton Kostin, Xinlu Zhuang, Cash TingHin Lo, Pavel Campr, Eric Semeniuc, Monsij Biswal, Shayenne Moura, Paul Brossier, Hojin Lee, and Waldir Pimenta. 2023. Librosa/librosa: 0.10.1. Zenodo.

J. Mehler, P. Jusczyk, G. Lambertz, N. Halsted, J. Bertoncini, and C. Amiel-Tison. 1988. A precursor of language acquisition in young infants. *Cognition*, 29(2):143–178.

T. Nazzi, J. Bertoncini, and J. Mehler. 1998. Language discrimination by newborns: Toward an understanding of the role of rhythm. *Journal of Experimental Psychology. Human Perception and Performance*, 24(3):756–766.

Thai Binh Nguyen. 2021. Vietnamese end-to-end speech recognition using wav2vec 2.0.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. Fairseq: A Fast, Extensible Toolkit for Sequence Modeling. pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Titouan Parcollet, Ha Nguyen, Solene Evain, Marcely Zanon Boito, Adrien Pupier, Salima Mdhaffar, Hang Le, Sina Alisamir, Natalia Tomashenko, Marco Dinarelli, Shucong Zhang, Alexandre Allauzen, Maximin Coavoux, Yannick Esteve, Mickael Rouvier, Jerome Goulian, Benjamin Lecouteux, Francois Portet, Solange Rossato, Fabien Ringeval, Didier Schwab, and Laurent Besacier. 2023. LeBenchmark 2.0: A Standardized, Replicable and Enhanced Framework for Self-supervised Representations of French Speech.

Ankita Pasad, Chung-Ming Chien, Shane Settle, and Karen Livescu. 2024. What Do Self-Supervised Speech Models Know About Words?

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. 2021. Layer-wise Analysis of a Self-supervised Speech Representation Model.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. MLS: A Large-Scale Multilingual Dataset for Speech Research. In *Interspeech 2020*, pages 2757–2761.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision.

Nari Rhee, Aoju Chen, and Jianjing Kuang. 2021. Going beyond F0: The acquisition of Mandarin tones. *Journal of Child Language*, 48(2):387–398.

Neville Ryant, Malcolm Slaney, Mark Liberman, Elizabeth Shriberg, and Jiahong Yuan. 2014a. Highly Accurate Mandarin Tone Classification In The Absence of Pitch Information. In *Speech Prosody 2014*, pages 673–677. ISCA.

Neville Ryant, Jiahong Yuan, and Mark Liberman. 2014b. Mandarin tone classification without pitch tracking. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4868–4872, Florence, Italy. IEEE.

Vance Schaefer and Isabelle Darcy. 2014. Lexical function of pitch in the first language shapes cross-linguistic perception of Thai tones. *Laboratory Phonology*, 5(4):489–522.

Gaofei Shen, Afra Alishahi, Arianna Bisazza, and Grzegorz Chrupała. 2023. Wave to Syntax: Probing spoken language models for syntax. In *INTERSPEECH 2023*, pages 1259–1263.

Rushen Shi, Jun Gao, André Achim, and Aijun Li. 2017. Perception and Representation of Lexical Tones in Native Mandarin-Learning Infants and Toddlers. *Frontiers in Psychology*, 8.

Leher Singh and Charlene S. L. Fu. 2016. A New View of Language Development: The Acquisition of Lexical Tone. *Child Development*, 87(3):834–854.

Leher Singh, Hwee Hwee Goh, and Thilanga D. Wewalaarachchi. 2015. Spoken word recognition in early childhood: Comparative effects of vowel, consonant and lexical tone variation. *Cognition*, 142:1–11.

Connie K. So and Catherine T. Best. 2010. Cross-language Perception of Non-native Tonal Contrasts: Effects of Native Phonological and Phonetic Influences. *Language and speech*, 53(Pt 2):273–293.

Kimiko Tsukada and Mariko Kondo. 2019. The Perception of Mandarin Lexical Tones by Native Speakers of Burmese. *Language and Speech*, 62(4):625–640.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need.

Dong Wang and Xuewei Zhang. 2015. THCHS-30 : A Free Chinese Speech Corpus.

Xinchun Wang and Jidong Chen. 2020. The Acquisition of Mandarin Consonants by English Learners: The Relationship between Perception and Production. *Languages*, 5(2):20.

Ethan Gotlieb Wilcox, Richard Futrell, and Roger Levy. 2023. Using Computational Models to Test Syntactic Learnability. *Linguistic Inquiry*, pages 1–44.

Yaru Wu, Martine Adda-Decker, and Lori Lamel. 2020. Mandarin lexical tones: A corpus-based study of word length, syllable position and prosodic position on duration. pages 1908–1912.

Jie Xi, Wei JIANG, Linjun Zhang, and Hua Shu. 2009. Categorical Perception of VOT and Lexical Tones in Chinese and the Developmental Course. *Acta Psychologica Sinica*, 41:572–579.

H. Henny Yeung, Ke Heng Chen, and Janet F. Werker. 2013. When does native language input affect phonetic perception? The precocious case of lexical tone. *Journal of memory and language*, 68(2):123–139.

Moira Yip. 2002. *Tone*. Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge.

Jiahong Yuan, Neville Ryant, Xingyu Cai, Kenneth Church, and Mark Liberman. 2021. Automatic recognition of suprasegmentals in speech.

Eric Zee. 1991. Chinese (Hong Kong Cantonese). *Journal of the International Phonetic Association*, 21(1):46–48.

Jian Zhu, Cong Zhang, and David Jurgens. 2022. Phone-to-audio alignment without text: A Semi-supervised Approach. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.