Tone in Perspective: A Computational Typological Analysis of Tone Function in ASR

Siyu Liang and **Gina-Anne Levow**
University of Washington
liangsy, levow@uw.edu

Abstract

This study investigates the impact of pitch flattening on automatic speech recognition (ASR) performance across tonal and non-tonal languages. Using vocoder-based signal processing techniques, we created pitch-flattened versions of speech recordings and compared ASR performance against original recordings. Results reveal that tonal languages experience substantially larger performance degradation than non-tonal languages. Analysis of tone confusion matrices shows systematic patterns of misidentification where contour tones collapse toward level tones when pitch information is removed. Calculation of tone's functional load at syllable and word levels demonstrates that syllable-level functional load strongly predicts ASR vulnerability to pitch flattening, while word-level patterns reflect each language's morphological structure. These findings illuminate the differential importance of pitch information across languages and suggest that ASR systems for languages with high syllable-level functional load require more robust pitch modeling.

1 Introduction

Lexical tone, where pitch distinctions signal differences in word meaning, is a core feature of over half the world's languages (Yip, 2002). While tonal contrasts rely primarily on fundamental frequency (f_0), they also interact with duration, intensity, and voice quality. These complexities pose unique challenges for automatic speech recognition (ASR), particularly in tonal languages where pitch plays a central role in lexical identity.

Recent ASR models implicitly encode tonal information, but it remains unclear how critical pitch actually is for recognition across language types. To investigate this, we apply pitch flattening—a signal processing technique that removes f_0 contours—to speech recordings and compare ASR performance with and without flattened pitch contours across both tonal (Thai, Vietnamese, Mandarin) and non-tonal (Uzbek, Indonesian, Turkish) languages.

We find that tonal languages experience significantly larger degradation in ASR performance under pitch flattening, with systematic tone confusion patterns revealing that contour tones (e.g., falling, rising) tend to collapse toward level tones when f_0 contours are removed. To explain these differences, we compute the functional load of tone and show that syllable-level functional load is a strong predictor of ASR vulnerability, capturing cross-linguistic differences in tone dependency more effectively than word-level metrics.

2 Background and Related Work

2.1 Tone

Tone refers to the use of pitch patterns to distinguish lexical or grammatical meanings, and it appears in over half of the world's languages (Yip, 2002). At its core, tone is related to fundamental frequency (f_0), often supplemented by secondary cues such as duration or phonation type (e.g., creaky or breathy voice) (Garellek et al., 2013; Zhang and Kirby, 2020). While languages like Thai, Vietnamese, and Mandarin all employ pitch contrasts, each does so differently: Thai traditionally has five tones, Vietnamese features six, and Mandarin typically has four plus a neutral tone (Yip, 2002; Thurgood, 2002). The functional importance of tone also varies cross-linguistically; in some systems, pitch shapes nearly every syllable, whereas others use additional cues for lexical contrasts.

From a linguistic perspective, these pitch contrasts often evolve through *tonogenesis*—the historical development of tone from segmental distinctions such as voicing (Haudricourt, 1954). Once established, tone can become as critical as vowels or consonants in signaling word meaning (Surendran and Levow, 2004). This high informational

Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2025), pages 82–92

load means that even small shifts in f_0 may yield major changes in lexical interpretation. Yet tone is not always "standalone": interactions with intonation, stress, or morphology can influence its role within the broader phonological system.

2.2 Tone and ASR

The significance of pitch in tone perception poses unique challenges for ASR technology. Early systems for Chinese and Thai explicitly modeled pitch tracks alongside spectral features (Fu et al., 1998; Lei et al., 2006), while modern end-to-end frameworks often rely on learned representations (e.g., XLS-R (Babu et al., 2021)) to capture tonal nuances. Even so, how effectively these systems handle pitch remains an open question—particularly for low-resource tonal languages, where sparse training data compound recognition errors (Coto-Solano, 2021; Qin et al., 2022).

2.3 Pitch Manipulation

One way to isolate pitch's contribution is *pitch flattening*, which systematically removes f_0 contours while preserving segmental and temporal information (Valbret et al., 1992). This technique has informed both psycholinguistic studies—showing how listeners rely on other cues like duration or context when pitch is lost (Wang et al., 2013)—and ASR research, where drops in recognition accuracy can reveal a system's reliance on pitch. Related work has compared natural speech against flattened or synthesized stimuli for languages such as Mandarin and Thai (Liu and Samuel, 2004; Zsiga and Nitisaroj, 2007), demonstrating substantial performance declines in human perception when f_0 cues are removed or distorted.

2.4 Functional Load

To quantify how critical pitch distinctions are in any given language, researchers often invoke *functional load* (Hockett, 1967; Surendran and Levow, 2004). This information-theoretic metric captures the extent to which a contrast (e.g., a particular tone versus no tone) contributes to lexical distinctions. Languages with a high tonal load—where a substantial portion of the semantic space hinges on pitch—are predictably more vulnerable when pitch cues degrade. In contrast, languages whose words can be distinguished by segmental or morphological features may be less affected by pitch flattening.

2.5 Tone and Typology

Because tone systems vary dramatically, from heavily monosyllabic languages like Vietnamese to those where multisyllabic words dilute the burden on pitch (Thurgood, 2002; Brunelle and Kirby, 2016), cross-linguistic experimentation is pivotal for robust ASR design. Studies have shown that, in some languages, phonation features may help compensate for reduced f_0 (Brunelle and Kirby, 2016), while in others, listeners (and ASR systems) default to level or "unmarked" tones when pitch is unavailable (Francis et al., 2003). By comparing both tonal and non-tonal languages under pitch-flattened conditions, we can pinpoint how different phonological structures handle the loss of f_0 cues and where ASR systems might fail. Insights from such comparisons suggest which modeling strategies, e.g., explicit pitch tracking, tone-based lexicons, or phonation-sensitive acoustic features, offer the most gains for languages heavily reliant on pitch.

3 Methods

We designed experiments to evaluate how pitch manipulation influences ASR performance across typologically diverse languages. Specifically, we investigate how removing lexical pitch cues via pitch flattening affects recognition accuracy in tonal versus non-tonal languages. By comparing ASR performance on original and pitch-flattened versions of the same utterances, we aim to quantify the importance of pitch information for recognition and identify the linguistic and structural factors that predict vulnerability to pitch manipulation.

3.1 Data

We selected six languages for our study: three tonal languages (Thai, Vietnamese, and Mandarin Chinese) and three non-tonal languages (Uzbek, Indonesian, and Turkish). Our selection of tonal languages was primarily constrained by data availability in the speech corpora and is typologically biased toward East and Southeast Asian tone systems. While these languages represent important tone types, they do not capture the full typological diversity of tone systems found worldwide, such as register tone languages of Africa or pitch-accent systems, which will be discussed in Section 7. All data were drawn from the Common Voice 17.0 corpus (Ardila et al., 2020). For each language, we used 2 hours of speech data for training and 30

Language	Original Text	Processed Text
Thai	ผมรักเธอ	phoom4 rak1 thoe0
Vietnamese	Tôi yêu bạn	tôi1 yêu1 ban6
Mandarin	我爱你	wo3 ai4 ni3

Table 1: Text preprocessing examples for “I love you” in the three tonal languages, showing original text and preprocessed text.

minutes for testing. All audio data were resampled at 16 kHz.

3.2 Preprocessing

For non-tonal languages, we applied minimal processing (standardized case and removed punctuation). For tonal languages, we applied specific preprocessing to ensure consistent transcription for tones. Table 1 shows examples of this preprocessing for each language.

For Thai, we used pythainlp.transliterate with engine=tltk_g2p, which converts Thai script to Latin characters with explicit tone marking (numbers 0–4). The numeric tone markers correspond to: 0 = mid tone, 1 = low tone, 2 = falling tone, 3 = high tone, and 4 = rising tone. Note that tone numbers used here follow a phonological convention rather than pitch height, where, for example, *rak1* (“love”) is a mid-tone syllable (not high), resulting from a low-class consonant with a dead syllable and no tone mark. In Vietnamese, we mapped diacritics denoting tone to numeric tone labels while keeping other diacritics for vowel contrast intact. Our mapping converted Vietnamese diacritics to numeric tone labels as follows: 1 = ngang (level/no diacritic), 2 = huyền (falling/grave accent), 3 = sắc (rising/acute accent), 4 = hỏi (dipping/hook), 5 = ngã (creaky/tilde), and 6 = nặng (heavy/dot below). For Mandarin Chinese, we used the pypinyin package with style=Style.TONE3. The numeric markers correspond to: 1 = high level tone (āi), 2 = rising tone (ái), 3 = falling-rising tone (ǎi), 4 = falling tone (ài), without explicitly including the neutral tone.

3.3 Pitch Flattening

Pitch flattening was performed using Praat’s Pitch-Synchronous OverLap and Add (PSOLA) algorithm (Valbret et al., 1992). This procedure effectively neutralizes lexical tone cues while maintaining other speech properties, including duration, intensity, and spectral envelope. In our implementation, the f_0 contour of each utterance was replaced with the utterance’s mean f_0 value. Figure 1 illustrates the process on a sample Thai utterance, showing the original and flattened pitch contours.

We should note that flattening the contour does not eliminate every trace of pitch, as micro-periodicity cues remain in the harmonic spectrum. Therefore, our results are a conservative estimate of tone dependence; a future experiment that additionally uses the interharmonic energy of low-pass filters would provide an even “cleaner” ablation.

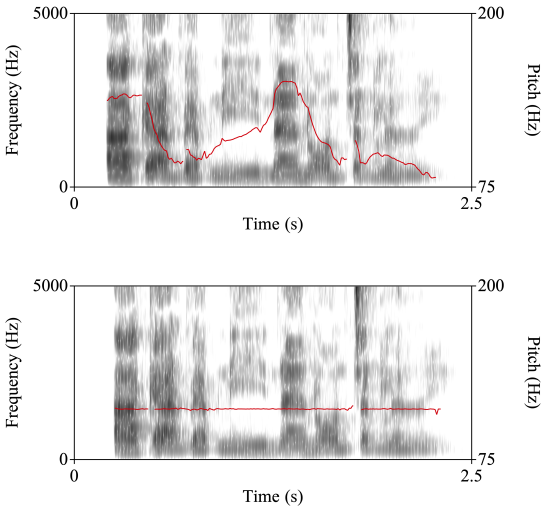

Figure 1: Example of pitch flattening on a Thai utterance "This kind of weather makes me feel sleepy." The top panel shows the original spectrogram overlaid with pitch contour; the bottom panel shows the flattened version of the same audio.

3.4 ASR Model Training

We fine-tuned individual XLS-R 300m models (Babu et al., 2021) for each language. Specifically, we trained the model on 2 hours of speech from Common Voice 17.0 and tested on 30 minutes. Additionally, for each tonal language, we ran the ASR model on pitch flattened test data too. Hyperparameters and training details are included in the Appendix (see Appendix A.1 for complete hyperparameter settings).

3.5 Evaluation Metrics

We evaluated ASR performance using multiple metrics to capture different aspects of recognition accuracy. In addition to WER (Word Error Rate) and CER (Character Error Rate), we also use additional metrics given in Table 2.

Metric	Description
TER	Tone Error Rate: percentage of syllables with correctly recognized segments but incorrectly identified tones
ConER	Consonant Error Rate: errors in consonant recognition
VER	Vowel Error Rate: errors in vowel recognition
WER-T	Modified version of WER where tone markers were ignored
CER-T	Modified version of CER where tone markers were ignored
Δ	Absolute difference between pitch-flattened and original speech for each metric

Table 2: Evaluation metrics used to assess ASR performance across languages.

3.6 Tone Confusion Analysis

For tonal languages, we constructed tone confusion matrices to analyze specific patterns of tone misidentification when pitch information was removed. These matrices recorded the counts of each reference tone (true label) being recognized as each possible tone (predicted label) in both original and pitch-flattened conditions. We then calculated difference matrices (flattened minus original) to identify which tonal confusions increased most dramatically after pitch flattening.

3.7 Functional Load Calculation

To quantify the information-theoretic contribution of tone in each language, we calculated the functional load (FL) of tonal contrasts at both syllable and word levels, following the methodology of Surendran and Levow (2004):

$$FL = \frac{H_{with} - H_{without}}{H_{with}} \quad (1)$$

where H_{with} represents the Shannon entropy of the distribution with tonal contrasts maintained, and $H_{without}$ represents the entropy after neutralizing tonal distinctions.

For syllable-level calculations, we extracted syllable frequencies from our corpus, maintaining or neutralizing tone distinctions to compute the respective entropies. For word-level calculations, we employed language-specific tokenization tools: PyThaiNLP with the newmm engine for Thai, Jieba for Mandarin, and underthesea for Vietnamese. These tools provided morphological segmentation used for analyzing the relationship between tone and word structure.

We also calculated the average number of syllables per word for each language to understand how morphological characteristics might influence the relationship between syllable-level and word-level functional loads. These calculations allowed us to quantitatively assess whether languages with higher functional load of tone would show greater vulnerability to pitch flattening in ASR performance.

4 Results

4.1 Impact of Pitch Flattening on ASR Performance

Table 3 presents our baseline ASR outcomes for six languages (three tonal, three non-tonal), comparing recognition on the original recordings vs. pitch-flattened audio that removes f_0 contours. As expected, the tonal languages (Vietnamese, Mandarin, Thai) experience substantially larger performance drops than the non-tonal ones (Uzbek, Indonesian, Turkish), confirming that pitch serves as a crucial contrastive cue for tone-based systems.

In particular, Thai displays the highest jump in WER upon flattening (+0.232), with Mandarin and Vietnamese also incurring significant degradations (+0.194 and +0.118). By contrast, pitch removal in Uzbek, Indonesian, and Turkish increases WER by only 5–8 points, indicating that segmental cues alone largely suffice for lexical discrimination in these atonal settings.

4.2 Tone Dependence and Detailed Phonetic Metrics

To examine tone-dependence in further detail, Table 4 shows additional metrics for the three tonal languages, including *tone error rate* (TER), *consonant error rate* (ConER), *vowel error rate* (VER), and error rates when ignoring tone markers (WER-T, CER-T). Thai exhibits the largest TER increase ($+0.2543$), reflecting its strong reliance on f_0 cues. Mandarin and Vietnamese also display pronounced TER jumps of +0.2009 and +0.1837, respectively.

Although consonant and vowel error rates increase less dramatically, they still reveal that pitch flattening affects the broader phonetic structure, not only the tonal dimension. When ignoring tone, i.e., disregarding tone output in error rate calculation, the error rates CER-T and WER-T of the three tonal languages are very similar to the non-tonal languages in Table 3.

4.3 Tone Confusion

Figure 2 illustrates the changes in tone confusion patterns after pitch flattening. More details about the values can be found in Appendix A.2. Each

Language	WER (orig.)	WER (flat.)	Δ_{WER}	CER (orig.)	CER (flat.)	Δ_{CER}
Tonal						
Vietnamese	0.715	0.833	0.118	0.312	0.380	0.068
Mandarin	0.478	0.672	0.194	0.209	0.283	0.074
Thai	0.288	0.520	0.232	0.082	0.154	0.072
Non-Tonal						
Uzbek	0.782	0.857	0.075	0.247	0.288	0.041
Indonesian	0.599	0.668	0.069	0.193	0.232	0.039
Turkish	0.743	0.816	0.073	0.240	0.292	0.052

Table 3: WER and CER results under original vs. pitch-flattened conditions, grouped by tonal and non-tonal categories. The Δ columns show $(\text{Flattened - Original})$.

Language	Version	TER	Δ_{TER}	ConER	Δ_{ConER}	VER	Δ_{VER}	WER-T	$\Delta_{\text{WER-T}}$	CER-T	$\Delta_{\text{CER-T}}$
Vietnamese	original	0.3954		0.3525		0.3739		0.6430		0.3063	
	flattened	0.5791	0.1837	0.3929	0.0404	0.4199	0.0460	0.6932	0.0502	0.3408	0.0345
Mandarin	original	0.3430		0.4300		0.3287		0.6169		0.4646	
	flattened	0.5439	0.2009	0.4658	0.0358	0.3686	0.0399	0.6838	0.0669	0.5066	0.0420
Thai	original	0.1266		0.0981		0.0864		0.2465		0.0810	
	flattened	0.3809	0.2543	0.1279	0.0298	0.1205	0.0341	0.3099	0.0634	0.1087	0.0277

Table 4: Comparison of tone error rate (TER), consonant error rate (ConER), vowel error rate (VER), and ignoring-tone WER/CER for Vietnamese, Mandarin, and Thai.

heatmap plots the *difference* (flattened minus original counts), where red regions indicate increased confusion and blue regions show decreased confusion. Analysis of these patterns reveals specific directional shifts in tone recognition after f_0 removal.

Across all three languages, diagonal elements (representing correct tone identification) show substantial negative values, indicating significantly reduced accuracy. Thai exhibits the largest average diagonal decrease (-146.40 per tone), followed by Mandarin (-232.25) and Vietnamese (-94.50). Conversely, off-diagonal elements show positive values (Thai: +29.28, Vietnamese: +15.36, Mandarin: +56.75), reflecting increased confusion between different tones.

The most pronounced confusion patterns are highly directional. In Thai, flattened audio led to falling tone being misidentified as mid tone (+246 instances), followed by rising tone confused with mid tone (+111). This suggests that without f_0 contours, the distinctive falling and rising patterns collapse toward the perceptually less marked mid tone. Thai's falling tone showed the largest proportional decrease in correct identification (-55.2%), followed by rising tone (-43.1%).

Vietnamese exhibited a striking trend where multiple tones were confused with ngang (level) tone after flattening: huyền (falling) $\rightarrow$ ngang (+312), sắc (rising) $\rightarrow$ ngang (+259), hỏi (dipping) $\rightarrow$ ngang (+92), and nặng (heavy) $\rightarrow$ ngang (+56). This systematic shift toward the unmarked ngang tone demonstrates how pitch flattening neutralizes the distinctive contour features of Vietnamese tones. The huyền tone showed the most dramatic reduction in correct identification (-46.9%), while the ngang tone was least affected.

For Mandarin, the most significant confusion was falling tone misidentified as high tone (+306), followed by rising tone confused with high tone (+129). Without pitch cues, distinctive contour tones (falling, rising, fall-rise) are increasingly confused with the level high tone. The falling tone experienced the largest decrease in accuracy (-30.6%), consistent with its heavily pitch-dependent contour.

These directional confusions reveal a general pattern: in the absence of f_0 contrast, contour tones (those with dynamic pitch movements such as falling, rising, or complex contours) collapse toward level tones (mid tone in Thai, ngang in Vietnamese, and high tone in Mandarin). While the results are consistent with the idea that level tones function as unmarked defaults, they could equally reflect an artefact of the acoustic manipulation: the loss of dynamic contour cues renders rising, falling, and dipping tones indistinguishable. We caution, however, that flattened utterances are acoustically

Language	Syllable FL	Word FL	Avg. Syll./Word	Δ_{WER}	Δ_{TER}
Thai	0.1243	0.0189	1.86	0.232	0.2543
Mandarin	0.0597	0.0336	1.15	0.194	0.2009
Vietnamese	0.0530	0.0517	0.99	0.118	0.1837

Table 5: Functional load (FL) of tone at syllable and word levels, with average syllables per word and ASR performance degradation metrics.

atypical for any training distribution. Some of the observed errors may thus reflect domain mismatch rather than pure loss of lexical information.

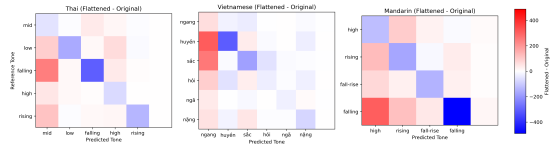

Figure 2: Confusion matrices based on tone count difference comparing flattened to original audio for Thai, Vietnamese, and Mandarin. Red cell marks increased prediction in that category, and blue cell marks decreases. Off-diagonal hotspots reveal a consistent drift of contour tones toward each language's level tone (mid, ngang, and high, respectively) in the left column.

4.4 Functional Load and Tone Dependency

To better understand the relationship between tone importance and ASR degradation, we calculated the functional load (FL) of tone at both syllable and word levels across the three tonal languages based on 32k tokens from the transcripts of the same Common Voice database, with the data scarcity of Vietnamese as the lower bound. Table 5 summarizs the results and reveals an interesting pattern: syllable-level functional load aligns remarkably well with ASR performance degradation. Thai exhibits the highest syllable-level FL (0.1243), followed by Mandarin (0.0597) and Vietnamese (0.0530), a ranking that precisely mirrors the order of WER increase under pitch flattening (Thai: +0.232, Mandarin: +0.194, Vietnamese: +0.118) and TER increase (Thai: +0.2543, Mandarin: +0.2009, Vietnamese: +0.1837). This strong correlation (Pearson's $r = 0.91$ for syllable FL vs. WER degradation) suggests that syllable-level functional load effectively predicts a language's ASR vulnerability to pitch flattening.

Interestingly, word-level functional load presents a different pattern. Vietnamese maintains nearly all of its tonal information at the word level (word FL: 0.0517, 97.5% of its syllable FL), while Mandarin preserves about half (word FL: 0.0336, 56.3% of syllable FL), and Thai retains only 15.2% (word FL: 0.0189). These proportions directly reflect each language's morphological structure: Vietnamese's predominantly monosyllabic words (average 0.99 syllables per word) necessitate tone distinctions for lexical identity, whereas Thai's higher proportion of multisyllabic words (average 1.86 syllables per word) allows tone to function more as one feature among many for word identification.

This morphological analysis complements our earlier confusion matrix findings. In Vietnamese, where tone information remains critical at the word level, confusion patterns show tones collapsing toward the less marked ngang (level) tone, but overall ASR degradation is less severe than in languages with higher syllable-level functional load. Thai, despite maintaining less tone information at the word level, experiences the largest performance drop precisely because its syllable-level tone distinctions carry substantial information that cannot be compensated for by other features when pitch is removed.

The pattern of flattening-induced confusion (contour tones collapsing toward level tones) observed in Figure 2 offers additional insight into why languages with higher syllable-level functional load suffer greater ASR degradation. Languages where tone carries more syllable-level information typically employ more distinctive contour tones, which are particularly vulnerable to pitch flattening. This vulnerability manifests in the dramatic decreases in recognition accuracy for falling (-55.2%) and rising (-43.1%) tones in Thai.

Taken together, these findings suggest that syllable-level functional load offers a more effective predictor of ASR vulnerability to pitch degradation than word-level measures. This has important implications for speech technology development across tonal languages: systems for languages with high syllable-level functional load will require more robust pitch modeling and may benefit from explicit tone-specific accommodations, while those for languages with lower tone dependency might be more resilient to noisy pitch environments.

5 Discussion

Our results reveal significant differences in how the ASR results of tonal and non-tonal languages respond to pitch flattening, with systematic patterns that illuminate the relationship between tone, speech perception, and ASR performance. These findings have important implications for both lin-

guistic theory and speech technology development.

5.1 Differential Impact of Pitch Flattening

The substantially larger ASR performance degradation observed in tonal languages (Thai: +23.2%, Mandarin: +19.4%, Vietnamese: +11.8% WER) compared to non-tonal languages (5-8% WER increase) confirms the critical role of f_0 information in tonal language processing. However, the non-zero impact on non-tonal languages indicates that pitch also contributes to speech recognition even when not lexically contrastive, likely through prosodic cues that help segment and identify words.

The varying degrees of degradation among tonal languages suggest differences in tone dependency. Thai showed the highest vulnerability to pitch flattening. This could be explained by our functional load analysis revealed Thai has a higher syllable-level tonal information density. These results align with Surendran and Levow (2004), who found language-specific differences in tone's functional load, but extend their work by demonstrating a direct relationship between this information-theoretic measure and ASR vulnerability.

The relatively smaller impact on Vietnamese (+11.8% WER) despite its complex six-tone system suggests that Vietnamese ASR benefits from additional disambiguating cues. As noted by Brunelle and Kirby (2016), Vietnamese tones involve substantial phonation contrasts (creaky, breathy voice) that may provide redundant information when pitch cues are removed. This phonation-based redundancy appears to partially compensate for the loss of f_0 information in Vietnamese, unlike in Thai and Mandarin where pitch plays a more singular role.

5.2 Tone Confusion Patterns and Perceptual Structure

The tone confusion analysis revealed striking directional patterns across all three tonal languages. In Thai, falling and rising tones were frequently confused with mid tone; in Vietnamese, multiple tones collapsed toward ngang (level) tone; and in Mandarin, contour tones were often misidentified as high tone. This systematic shift of confusion from contour tones toward level tones suggests that with neutralized f_0 cues, ASR systems default to perceptually unmarked tonal categories, a finding that parallels observations in human speech perception studies (Francis et al., 2003; Khouw and Ciocca, 2007).

It should be noted that pitch-flattened syllables are not strictly equivalent to natural 'level tones' in these languages. Natural level tones in East and Southeast Asian languages also include pitch movements, such as a slight fall or rise at the end, and are produced with specific phonation characteristics (Yip, 2002). Despite this distinction, our results show that when pitch information is neutralized through flattening, ASR systems consistently default to categorizing these flattened stimuli as level tones, suggesting that level tones serve as defaults in the absence of distinctive pitch movement.

These directional confusions have both acoustic and phonological implications. Acoustically, contour tones (with dynamic pitch movements) are more dependent on f_0 information than level tones. Phonologically, the patterns align with markedness theory: level tones typically function as unmarked categories in tonal systems (Yip, 2002), serving as defaults when distinctive features are unavailable. Importantly, this pattern is not simply a frequency effect, such as evident in our Mandarin data (see Appendix A.2) where the falling tone (4) is actually the most frequent in our dataset, yet confusion still predominantly shifts toward the high level tone (1) rather than following raw frequency distributions.

The diagonal values in the confusion matrices (representing correct identification) showed the largest decreases for tones with substantial pitch movement: falling tone in Thai (-55.2%), huyền tone in Vietnamese (-46.9%), and falling tone in Mandarin (-30.6%). This suggests that the perceptual distance between tones is not uniform but depends on their phonetic realization, with contour tones being perceptually more distant from other categories and thus more vulnerable to pitch flattening.

5.3 Functional Load and Language Structure

Our functional load analysis provides a quantitative framework for understanding cross-linguistic differences in tone dependency. The strong correlation between syllable-level functional load and ASR degradation (Thai: 0.1243/+23.2% WER, Mandarin: 0.0597/+19.4% WER, Vietnamese: 0.0530/+11.8% WER) suggests that this information-theoretic measure effectively predicts a language's vulnerability to pitch flattening.

The differences between syllable-level and word-level functional load reflect each language's morphological structure. Vietnamese maintained nearly all its tonal information at the word level (97.5%

of syllable-level FL), consistent with its predominantly monosyllabic nature. By contrast, Thai preserved only 15.2% of its tonal information at the word level, reflecting its higher proportion of multisyllabic words where tone distinctions on individual syllables become less critical for overall word identification.

These patterns highlight an important insight: a language's dependency on tone is not solely determined by the number of tonal contrasts or their acoustic properties, but also by the information-theoretic role of tone within the broader phonological and morphological system. Languages with high syllable-level functional load, especially those with significant proportions of monosyllabic words, are inherently more vulnerable to pitch perturbations.

5.4 Implications for ASR Development

Our findings have several practical implications for ASR system development. First, they suggest that pitch modeling requirements differ substantially across languages, even among those classified as tonal. Languages with high syllable-level functional load (like Thai) would benefit from explicit modeling of pitch contours, while those with redundant cues (like Vietnamese) might achieve acceptable performance with less sophisticated pitch representations.

Second, the systematic tone confusion patterns identified could inform error correction strategies in ASR systems. By understanding the likely confusion directions when pitch information is degraded (e.g., contour tones being misidentified as level tones), post-processing algorithms could apply targeted corrections based on contextual and acoustic cues.

Third, our results suggest that ASR robustness for tonal languages could be improved through explicit modeling of phonation cues, particularly for languages like Vietnamese where voice quality provides redundant information. Integrating both pitch and phonation features would create systems more resilient to acoustic degradations affecting either dimension.

Fourth, language modeling capabilities could potentially compensate for degraded tonal information. Our experiments used a basic CTC-based approach without additional language modeling, but we hypothesize that stronger language models could help recover tone information from context in pitch-degraded scenarios. This could be particularly effective in languages with higher word-level redundancy, where contextual cues might disambiguate tonally similar syllables.

Finally, the functional load framework offers a principled approach for predicting a priori which languages will require more sophisticated tone modeling in ASR systems. Rather than treating all tonal languages uniformly, developers could allocate resources based on information-theoretic measures of tone's importance in each language.

6 Conclusion

This study investigated the impact of pitch flattening on ASR performance across tonal and non-tonal languages, revealing several key insights about the role of pitch in speech recognition. Our findings demonstrate that tonal languages experience substantially greater performance degradation when pitch information is removed, but with significant variations that correlate with the functional load of tone in each language. The systematic patterns of tone confusion observed—where contour tones collapse toward level tones—highlight fundamental aspects of tonal perceptual structure.

Beyond documenting these effects, we established a quantitative relationship between information-theoretic measures of tone importance and ASR vulnerability. Languages with high syllable-level functional load proved most susceptible to pitch flattening, while word-level functional load patterns reflected each language's morphological characteristics. This framework offers a principled approach for predicting which languages will require more sophisticated tone modeling in speech technology applications.

Our findings have implications for both linguistic theory and ASR system development. Theoretically, they support models of tone perception where unmarked level tones serve as default categories when distinctive pitch information is unavailable. Practically, they suggest that ASR systems for tonal languages should be designed with language-specific considerations of tone's functional load and the availability of redundant acoustic cues.

Future work could extend this analysis to a wider typological range of tone systems. For instance, examining Cantonese, which features a more complex inventory of level tones, could test whether our observed pattern of confusion toward level tones holds in languages where multiple level tones must

be distinguished. Similarly, investigating Bantu languages, which feature tonal contrasts that are often analyzed differently from East Asian systems despite having contour properties, would broaden our typological understanding of how different tone systems respond to pitch degradation.

7 Limitations

While providing valuable insights, our study has several limitations that suggest directions for future research. First, our analysis focused on ASR performance rather than human perception. Parallel studies with human listeners would clarify whether the confusion patterns observed are specific to machine learning systems or reflect broader perceptual principles.

Second, our pitch flattening approach, while effective at isolating the contribution of f_0, represents an extreme case of pitch degradation. Future work could explore more nuanced manipulations, such as partial flattening or targeted disruption of specific pitch features, to identify which aspects of the pitch contour are most critical for recognition.

Third, our functional load calculations were limited to tone's contribution and did not address interactions with other phonological features. Expanding this analysis to include phonation, vowel quality, and other features would provide a more comprehensive understanding of how different dimensions contribute to lexical contrasts across languages.

Fourth, our ASR system used basic CTC-based decoding without sophisticated language modeling. A stronger language model would likely improve overall performance and might partially compensate for pitch flattening through contextual prediction. Future work should investigate the degree to which language modeling can mitigate the effects of degraded tonal information in various languages.

Finally, while we included three major tonal languages, our study does not capture the full typological diversity of tone systems. Extending this work to include languages with different tonal inventories (e.g., Cantonese with its multiple level tones), register tone languages (e.g., Hmong), pitch-accent languages (e.g., Japanese), and languages with different tone systems like those found in Bantu languages would provide a more complete picture of how pitch information contributes to speech recognition across language types.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus. *arXiv preprint*. ArXiv:1912.06670 [cs].

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. *arXiv preprint*. ArXiv:2111.09296 [cs, eess].

Marc Brunelle and James Kirby. 2016. Tone and Phonation in Southeast Asian Languages: Tone and Phonation in Southeast Asian Languages. *Language and Linguistics Compass*, 10(4):191–207.

Rolando Coto-Solano. 2021. Explicit Tone Transcription Improves ASR Performance in Extremely Low-Resource Languages: A Case Study in Bribri. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 173–184, Online. Association for Computational Linguistics.

Alexander L. Francis, Valter Ciocca, and Brenda Kei Chit Ng. 2003. On the (non)categorical perception of lexical tones. *Perception & Psychophysics*, 65(7):1029–1044.

Qian-Jie Fu, Fan-Gang Zeng, Robert V Shannon, and Sigfrid D Soli. 1998. Importance of tonal envelope cues in Chinese speech recognition. *The Journal of the Acoustical Society of America*, 104(1):505–510. Publisher: Acoustical Society of America.

Marc Garellek, Patricia Keating, Christina M. Esposito, and Jody Kreiman. 2013. Voice quality and tone identification in White Hmong. *The Journal of the Acoustical Society of America*, 133(2):1078–1089.

André-Georges Haudricourt. 1954. De l'origine des tons en vietnamien. *Journal Asiatique*, 242:69–82.

Charles F. Hockett. 1967. The Quantification of Functional Load. *WORD*, 23(1-3):300–320. Publisher: Routledge _eprint: https://doi.org/10.1080/00437956.1967.11435484.

Edward Khouw and Valter Ciocca. 2007. Perceptual correlates of Cantonese tones. *Journal of Phonetics*, 35(1):104–117.

Xin Lei, Manhung Siu, Mei-Yuh Hwang, Mari Ostendorf, and Tan Lee. 2006. Improved tone modeling for Mandarin broadcast news speech recognition. In *Interspeech 2006*, pages paper 1752–Tue3A2O.4–0. ISCA.

Siyun Liu and Arthur G. Samuel. 2004. Perception of Mandarin lexical tones when F0 information is neutralized. *Language and Speech*, 47(Pt 2):109–138.

Siqing Qin, Longbiao Wang, Sheng Li, Jianwu Dang, and Lixin Pan. 2022. Improving low-resource Tibetan end-to-end ASR by multilingual and multilevel unit modeling. *EURASIP Journal on Audio, Speech, and Music Processing*, 2022(1):2.

Dinoj Surendran and Gina-Anne Levow. 2004. The functional load of tone in Mandarin is as high as that of vowels. In *Speech Prosody 2004*, pages 99–102. ISCA.

Graham Thurgood. 2002. Vietnamese and tonogenesis: Revising the model and the analysis. *Diachronica*, 19(2):333–363.

H. Valbret, E. Moulines, and J. P. Tubach. 1992. Voice transformation using PSOLA technique. *Speech Communication*, 11(2):175–187.

Jiuju Wang, Hua Shu, Linjun Zhang, Zhaoxing Liu, and Yang Zhang. 2013. The roles of fundamental frequency contours and sentence context in Mandarin Chinese speech intelligibility. *The Journal of the Acoustical Society of America*, 134(1):EL91–EL97.

Moira Jean Winsland Yip. 2002. *Tone*. Cambridge textbooks in linguistics. Cambridge University Press, Cambridge ; New York.

Yubin Zhang and James Kirby. 2020. The role of F0 and phonation cues in Cantonese low tone perception. *The Journal of the Acoustical Society of America*, 148(1):EL40–EL45.

Elizabeth Zsiga and Rattima Nitisaroj. 2007. Tone Features, Tone Perception, and Peak Alignment in Thai. *Language and Speech*, 50(3):343–383. Publisher: SAGE Publications Ltd.

A Appendix

This appendix provides additional details on our fine-tuning hyperparameters for XLS-R 300m in both experiments.

A.1 XLS-R Fine-Tuning Hyperparameters

All training runs (for both Common Voice and TIBMD@MUC data) used the same set of essential hyperparameters, with only minor adjustments for batch size depending on GPU memory:

- **Model:** facebook/wav2vec2-xls-r-300m
- **Batch Size:** 8
- **Learning Rate:** 3×10^{-4}
- **Warmup Steps:** 500
- **Max Steps:** 2000
- **Vocabulary Size:** based on unique characters in the training corpus (including space or | as word delimiter).

A.2 Tone Confusion Results

The results of tone confusion are as follows:

	0	1	2	3	4	none
0	912	25	25	33	13	3
1	36	473	17	16	15	1
2	35	10	496	14	3	0
3	35	12	15	287	3	0
4	12	23	5	12	274	1
none	0	0	0	0	0	0

Table 6: Thai tone confusion (**Original**). Rows = reference tone (0 = Mid, 1 = Low, 2 = Falling, 3 = High, 4 = Rising, none = no assigned tone), columns = predicted tone.

	0	1	2	3	4	none
0	862	21	65	49	9	3
1	130	309	30	81	6	3
2	281	24	188	54	9	3
3	82	25	21	218	5	1
4	123	19	19	30	133	3
none	0	0	0	0	0	0

Table 7: Thai tone confusion (**Flattened**). Rows = reference tone (0=Mid, 1=Low, 2=Falling, 3=High, 4=Rising, none=no tone), columns = predicted tone.

	1	2	3	4	5	6	none
1	751	32	172	16	11	25	10
2	188	354	42	23	4	46	6
3	73	20	440	68	20	34	7
4	33	58	18	99	21	40	3
5	10	2	29	21	71	7	3
6	35	30	60	26	24	184	4
none	0	0	0	0	0	0	0

Table 8: Vietnamese tone confusion (**Original**). Tones: 1 = mid, 2 = huyền (falling), 3 = sắc (rising), 4 = hỏi (dipping), 5 = ngã (creaky), 6 = nặng (heavy), none = no tone. Rows = reference, columns = predicted.

	1	2	3	4	5	6	none
1	815	7	135	14	13	15	9
2	500	43	78	14	6	13	6
3	332	9	260	13	12	29	5
4	125	3	49	68	14	11	3
5	49	1	14	21	41	16	2
6	91	9	109	22	20	105	5
none	0	0	0	0	0	0	0

Table 9: Vietnamese tone confusion (**Flattened**). Tones: 1=mid, 2=falling, 3=rising, 4=dipping, 5=creaky, 6=heavy, none=no tone. Rows = reference, columns = predicted.

	1	2	3	4	none
1	678	64	42	164	13
2	78	700	97	163	34
3	56	99	434	104	22
4	145	130	97	1192	34
none	12	31	14	26	121

Table 10: Mandarin Chinese tone confusion (**Original**). Tones: 1=high-level, 2=rising, 3=dipping, 4=falling, none=no tone. Rows = reference, columns = predicted.

	1	2	3	4	none
1	553	163	66	154	21
2	207	529	85	200	40
3	130	126	290	134	29
4	451	252	140	703	52
none	26	25	9	25	116

Table 11: Mandarin Chinese tone confusion (**Flattened**). Tones: 1=high-level, 2=rising, 3=dipping, 4=falling, none=no tone. Rows = reference, columns = predicted.