You Sound a Little Tense: L2 Tailored Clear TTS Using Durational Vowel Properties

Paige Tuttösí[1,2,3], *H. Henny Yeung*[4], *Yue Wang*[4], *Jean-Julien Aucouturier*[2], *Angelica Lim*[1]

[1]School of Computing Science, Simon Fraser University, Canada
[2]SUPMICROTECH, CNRS, institut FEMTO-ST, Université Marie et Louis Pasteur, France
[3]Enchanted Tools, France
[4]Department of Linguistics, Simon Fraser University, France

ptuttosi@sfu.ca

Abstract

We present the first text-to-speech (TTS) system tailored to second language (L2) speakers. We use duration differences between American English tense (longer) and lax (shorter) vowels to create a "clarity mode" for Matcha-TTS. Our perception studies showed that French-L1, English-L2 listeners the participants had fewer (at least 9.15%) transcription errors when using our clarity mode, and found it more encouraging and respectful than overall slowed down speech. Remarkably, listeners were not aware of these effects: despite the decreased word error rate in clarity mode, listeners still believed that slowing all target words was the most intelligible, suggesting that actual intelligibility does not correlate with perceived intelligibility. Additionally, we found that Whisper-ASR did not use the same cues as L2 speakers to differentiate difficult vowels and is not sufficient to assess the intelligibility of TTS systems for these individuals.
Index Terms: L2-tailored TTS, adaptive speech synthesis, L2 speech perception, accessible speech synthesis, clear speech

1. Introduction

Imagine your immigrant family arrives in a new country; you speak the language of your new country at a basic level, but all of a sudden, you need to understand public announcements and alerts on transit, or you need to call to set up your medical card and pass through an automated voice messaging system. In all of these cases, you will encounter the use of voice-based technology, in recognition, synthesis, and end-to-end systems, which is an area that is currently seeing expansive growth. While these text-to-speech (TTS) systems can help improve accessibility, such as in-home automation systems for the physically and visually impaired [1, 2], or as a socially supportive device to allow elderly individuals to maintain their autonomy [3], understanding TTS can be daunting for second-language (L2) speakers when synthesized voices are fast-speaking and directed toward first-language (L1) speakers.

There remains much work to ensure the ethical validity of such systems in terms of transparency, bias, fairness, and many other factors [4]. One aspect of fairness lacking in systems with synthesized speech is their inability to adapt to those with different hearing and comprehension abilities, which is the case for L2 speakers. Previous work in clear TTS has principally explored speech-in-noise [5–9], rather than factors related to speaker comprehension, like language ability. In human-to-human interaction, the ability of a speaker to adapt to an interlocutor is invaluable: humans modify their speech to L2 speakers [10]. Yet, evidence shows that human adaptation to L2 speakers does not consistently aid the L2 speakers in their comprehension [11, 12], therefore training on L2-directed speech may not be the ideal approach. To improve the fairness of voice technology and move towards adaptive synthesized speech, we must first understand how speech can be adjusted to facilitate comprehension in a data-driven, perception-based manner.

To improve TTS for second language speakers, we present a clarity mode for English TTS, built upon a perception-based approach. L2 speakers of English, particularly those whose L1 does not have vowel tensity contrast, often struggle to differentiate English lax (less extreme articulation and shorter) vowels with tense (more extreme articulation and longer) vowels [13–18]. For example, French speakers may replace the lax /ɪ/ (ship) by the tense /i/ (sheep) [16]. In our paper [19], psychophysical reverse-correlation [20] was used to reconstruct L1-English and L2-English listener's mental representation of duration and pitch in tense-lax vowel contrasts in a data-driven manner. For L2-English speakers who spoke French, Mandarin and Japanese as an L1, it was found that vowel duration, but not pitch, affected perception of tense/lax vowels. This suggests that we can apply a duration mechanism to improve comprehension of this English vowel contrast when an L2 listener struggles to use the primary formant cues [21]. Following these results, we create, for the first time, an L2 "clarity" mode within a TTS to improve L2 comprehension of these difficult American English vowels.

2. Mechanism validation

In [19] we showed that shortening a word with an ambiguous vowel sound resulted in an English-L2 listener hearing the lax vowel, and lengthening the word resulted in hearing the tense vowel. We also explored whether these duration cues could be used to control the perception of *non-ambiguous* vowels in English-L2 [22]. Overall, we observed a bias towards the lax vowel, yet English-L2 speakers (with French, Chinese, and Japanese L1s) could be convinced to hear the tense vowel if the duration of a word containing a lax vowel became too long [22]. These results suggested that the usual approach for L2-directed speech, slowing either the entire phrase or difficult words for emphasis, can result in lax vowels being mistaken as tense vowels by L2 speakers from a variety of L1 backgrounds, as also shown in clear speech [23]. Instead, to improve L2 comprehension, the duration properties of tense/lax vowels needed to be maintained, applying the emphasis only to tense vowels.

3. L2 clarity TTS

Given the improved perception of tense/lax vowels with simple, linguistically-driven duration changes, we added this as a new 'clarity mode' in Matcha-TTS [24] for L2 speech.

To enable L2 clarity mode in Matcha-TTS, we added a clarity flag that can be set to "True" or "False" at synthesis, along

with a markup to control which words are emphasized. The user (or large language model driving a dialogue system) surrounds difficult words with exclamation points, e.g., "!peel!", allowing the TTS to parse the words to be treated for clarity through several steps: 1) Parse each flagged word to see if it contains a tense or lax vowel, 2) If the word contains tense vowels but no lax vowels, the clarity modification is applied to the tense vowel containing word, 3) If the word contains both tense and lax vowels, and if the tense vowel has primary stress, the clarity modification is applied to the tense vowel containing word.

The modifications are made by applying an array ($carray$) with the same length as the phonemized phrase to scale the predicted duration of each phoneme (w), a method shown to result in natural and effective duration changes in TTS systems containing a phonemizer [25]. This *clarity duration* multiplier is applied after the base speech rate multiplier, which is an array containing the speech rate provided at synthesis (in our case, 0.75 which in our pilot was found to have a natural, conversational speech rate for English L1 listeners). The predicted durations (w) result from the Hadamard product of the text encoder outputs (Equation 1): $\log w$: the log duration predicted by the duration predictor as in [26], and x_{mask}: the mask for the text input indicating which values are valid [27]. In Matcha-TTS this is then scaled to the input duration the $speechrate$ array. The resulting y_{lengths} and $y_{\text{max\_length}}$ are then used to calculate the attention alignment map for the text encoder.

$$\begin{aligned} w &= e^{\log w} \odot x_{\text{mask}} \\ w_{\text{ceil}} &= (\lceil w \rceil \odot \text{speechrate}) \odot \text{c\_array} \\ y_{\text{max\_length}} &= \max\left(\max\left(1, \sum_{i,j} w_{\text{ceil}}[i,j]\right)\right) \end{aligned} \tag{1}$$

A 1.6x stretch is applied across the entire word with a gradual ramp up and down to the base speech rate over the 6 phonemized items (if there are 6 phonemized items between target words, otherwise as many as are available) preceding and following the target word. Six items were chosen as this encompasses two phonemes preceding and following the target word (approximately 200-300ms [28] as per [19]).We also apply a 1x stretch (the base speech rate) to lax-vowel-containing words simultaneously, following the same parsing above to ensure that lax vowels are not stretched by surrounding tense vowel words. A 1.6x stretch was chosen to minimize duration changes. In our validation [22], we saw that perception performance quickly improved from the baseline by increasing the duration of the target, tense vowel-containing word. There was minimal improvement in performance between a 1.6x and 2.0x stretch.

3.1. Stimulus generation

We tested 4 different TTS styles: 1. **Base**: the base Matcha-TTS (0.75x speech rate), 2. **Stretch**: 1.2x (0.75×1.6) speech rate applied across the entire phrase, 3. **Emphasis**: 1.6x stretch applied across all target words (0.75x speech rate elsewhere), and 4. **Clarity**: 1.6x stretch applied across the target words containing a tense vowel (0.75x speech rate elsewhere).

Previously, in [19, 22], only a **single target word**, always at the end of a phrase, was tested. To test the robustness of the L2 clarity TTS when the target words occurred in other parts of the phrase, we tested clarity mode in several contexts. First, single target word phrases where the target word was in the middle of the phrase were tested. We then tested **double target word** phrases (e.g., "Write down dull and doll"), where the targets could be at the end of the phrase, in the middle of the phrase, or at the beginning of the phrase. Additionally, we explored the performance when the TTS had a combination of tense and lax vowels in a single phrase; we tested a tense and a lax minimal pair, a tense and a lax that are not minimal pairs, two tense and two lax vowels. We aimed to use a variety of starting and ending consonants surrounding the target vowels, and all words had a minimal pair in English. We tested all tense/lax vowel pairs: /i/ (peel) and /ɪ/ (pill), /u/ (fool) and /ʊ/ (full), /ɑ/ (cot) and /ʌ/ (cut). Lastly, we aimed not to semantically bias phrase meanings towards any word, that is, given the context outside of the target word neither word in the minimal pair made more logical or semantic sense. To do this, we asked ChatGPT-4o to provide a starting list of neutral phrases with varying lengths. These were then selected and modified to create our final list of 16 phrases and insert our target words (Table 1).

Table 1: *List of phrases for the experiment*

Phrases
The word cut seemed important to the instructions.
She kept mentioning cot during the conversation.
The speaker mentioned pill, or at least something similar.
The word pill was what she was trying to write.
The phrase had fool somewhere in the middle of it.
I saw full written on the note pad.
The sign mentioned sin, but the person said scene.
He wrote down bought, but remembered it as but.
In his talk he kept using could, but I am pretty sure he meant cooed.
The paper mentioned kid, yet he is telling me knot.
There was confusion between pull and bean in their speech.
I am not sure if the word was pool or if cup was the right one.
Sheep goes on the top of the page and dull goes on the bottom.
Bit was the first word he said, then nut followed.
Actually hut is the correct word, it was replaced with should by accident.
Maybe he said hot, but I really thought keyed was what he said.
Reap was a more important word in the story than wooed.
Confusion Phrases
The speaker mentioned pill, or at least something similar.
He wrote down but, but remembered it as bought.
There was confusion between pool and bin in their speech.
Maybe he said hut, but I really thought kid was what he said.
The word cot seemed important to the instructions.
The sign mentioned scene, but the person said sin.
Ship goes on the top of the page and doll goes on the bottom.
Rip was a more important word in the story than wood.
The phrase had full somewhere in the middle of it.
In his talk he kept using cooed, but I am pretty sure he meant could.
I am not sure if the word was pull or if cop was the right one.
Actually hot is the correct word, it was replaced with shooed by accident.
She kept mentioning cut during the conversation.
The word peel was what she was trying to write.
The paper mentioned keyed, yet he is telling me nut.
Beat was the first word he said, then knot followed.
I saw fool written on the note pad.

Because participants would hear each of the phrases 4 times (each TTS style had the same phrases), we added 1 confusion phrase (randomly assigning a TTS style) for each of the phrases. A confusion phrase was the same phrase context using the opposite minimal pair, for example: "She kept mentioning cot during the conversation." and the confusion phrase "She kept mention-

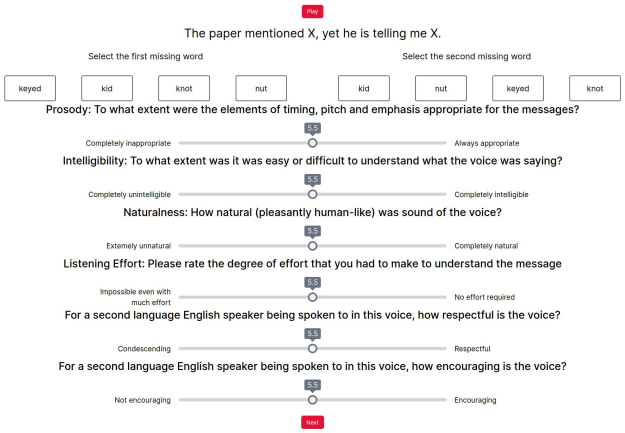

Figure 1: *L2-TTS double-word experiment set up in Gorilla.*

ing cut during the conversation." Lastly, for the single target word TTS, "clarity" TTS was the same as "emphasis" for a single tense vowel target and "base" for a single lax vowel target. As such, we assessed the TTS in terms of the length of the target words rather than TTS styles and ensured no repeated phrases with the same treatment.

3.2. Experimental procedure

We conducted an experiment to assess the objective (through word error rate) and subjective (through mean opinion scores) performance of French L1 English L2 listeners using our "clarity mode" compared to the baseline models (i.e., "base", "stretch", "emphasis"). In an online experiment, participants chose to receive instructions in either English or French. They then provided demographic information on the language they first learned, their most commonly used daily language, their age and gender, and their self-rated English proficiency. Each participant was randomly assigned to start in either the single- or double-word trial.

The participants were shown the phrase with the target words removed, e.g., "Write down the word X followed by the word X on the paper," and could play the audio only once. They then selected which words they heard and were told (in the double-word case) that it was possible to hear the same word twice. It was never the case that they would hear the same word, however, this instruction was added to encourage participants not to base their choice for the second word on what they believe they heard in the first word. The missing word was selected from a list of 4 words: for the single-word trial, 2 words were the tense/lax vowel minimal pair (e.g., beat, bit); the third word was a minimal pair word with another vowel (e.g., bat); and the final word was dissimilar from the other three choices (e.g., shop). The dissimilar choice functioned as an attention check. In the double word case, when the two words were a minimal pair, the selection for the remaining two words was as per the single-word trial; if the two words were not a minimal pair, the 4 choices were the two words and their minimal pairs. The order of the phrases was randomized for each participant. The experiment set up can be seen in Fig. 1.

Once the participants selected which word they heard, they responded to a questionnaire containing the MOS-X2 [29] naturalness (nMOS), intelligibility (iMOS), and prosody (pMOS) scores, as well as an intelligibility question for listening effort (eMOS) ("Please rate the degree of effort you had to make to understand the message") from MOS-X [29]. The listening effort question was added to understand if there is a difference between being able to understand the phrase and how much effort was required to understand the phrase. They then responded to two questions from L2-directed speech research [12]: "For an English second language speaker being spoken to with this voice, how respectful is the voice?" (Resp.: condescending - respectful), "For an English second language speaker being spoken to with this voice, how encouraging is the voice?" (Enc: not encouraging - encouraging). These questions were to help us understand if slowing down or adding emphasis makes the TTS sound condescending, as can be the case with L2-directed speech [11, 12]. All scores were on a 10-point Likert scale.

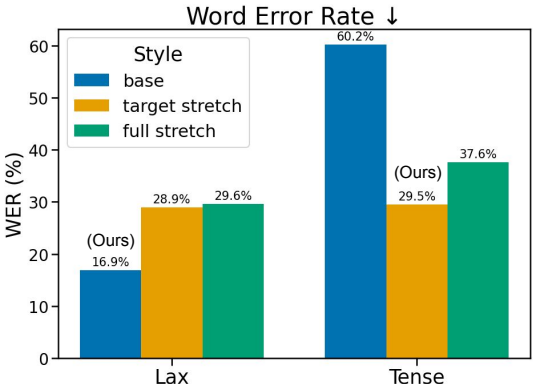

Figure 2: *Single target word results: our clarity mode has the lowest WER for lax-base (left) and tense-target stretch (right).*

3.2.1. Participants

We recruited N=56 participants (29F, age = 34.59±10.67) via Prolific. French was our target L1 population and the participant language demographics were as follows: daily language = French (40), English (14), French and English (2), Italian (1); English proficiency (1-5) = 5 (29), 4 (20), 3 (5), 2 (2); 71.4% of participants chose to have the instructions in French. The study received internal ethics approval.

3.3. Results

3.3.1. Single word

To confirm our previous findings, we expect improved performance in WER with lengthened tense vowels but decreased performance for the same treatment in lax vowels. We computed one-way ANOVAs (Type II) followed by a post-hoc Tukey HSD as well as word error rates (WER) as the proportion of incorrect target words for each type of target word treatment. In the single word case "Clarity" is the same as "Base" for lax vowel target words, and "Emphasis" for tense vowel target words. As such, we present the results in the three defined states existing for the single word case: "base" has the baseline duration on the target, "target stretch" is 1.6x the duration of the baseline on the target, and "full stretch" is 1.2x the duration on the entire phrase.

We observed that, through our "clarity" mode stretch applied to tense-vowel-containing words, we could overcome the bias towards lax vowels, i.e., the fact that participants were more likely to respond, for example, "pill" than "peel," which we observed in our validation study. By lengthening the target tense-vowel-containing word, the WER on the baseline ("base tense") was reduced from 60.23% to 29.48% for "target stretch tense" (Table 2, Fig. 2). Importantly, we also confirmed that, indeed, stretching lax vowels, as is typical in L2-directed speech, re-

Table 2: *Single target word results: word error rate, intelligibility (iMOS), naturalness (nMOS), effort (eMOS), prosody (pMOS), encouragement (Enc) and respect (Resp.) Our clarity mode, using base lax and target stretch tense, achieves the lowest WER.*

TTS Style	WER ↓	iMOS ↑	nMOS ↑	eMOS ↑	pMOS ↑	Enc. ↑	Resp. ↑
Base lax (Ours)	**16.86**%	7.63	7.11/***	7.33	7.20/***	6.44/***	6.63/**
Target stretch lax	28.90%	**8.18***/**	**7.91****/***	**7.73**	**7.87***/***	**7.03***/***	**7.32***/***
Full stretch lax	29.65%	7.53	4.18	7.40	5.83	5.54	5.76
Base tense	60.23%	7.57	7.07/***	7.18	7.22/***	6.45/***	6.53/**
Target stretch tense (Ours)	**29.48**%	**8.18***/	**7.71**/***	**7.69**	**7.80**/***	**6.84**/***	**7.00**/***
Full stretch tense	37.57%	7.68	4.27	7.61	6.13	5.49	5.68

Significance values are presented as: Significance compared to base or target stretch / Significance compared to full stretch TTS
* = 0.05, ** = 0.01, *** = 0.001

Table 3: *Single target word results: ANOVA and Tukey test statistics on MOS Likert scores.*

TTS Style	F-Statistic (5,1029)	P-Value	Significant Tukey Results
iMOS	5.12	<.001	Lax: targetS/base,fullS p=.043, p=.008; Tense: targetS/base p=.016
nMOS	110.83	<.001	Lax: all/fullS p<.001 targetS/base p=.006; Tense: all/fullS p<.001
eMOS	2.21	0.051	N/A
pMOS	30.24	<.001	Lax: all/fullS p<.001 targetS/base p=.028; Tense: all/fullS p <.001
Enc.	21.45	<.001	Lax: all/fullS p<.001 targetS/base p=.036; Tense: all/Stretch p<.001
Resp.	17.90	<.001	Lax: targetS/fullS,base p<.001, p=0.022 base/fullS p=0.001; targetS/fullS p<.001 base/fullS p=0.001

Tukey results are presented as: Higher value / Lower value
fullS = full stretch, targetS = target stretch

sults in more errors (28.90% vs. 16.86%). Lastly, the "stretch" TTS resulted in a slight reduction in WER over the baseline for tense vowels, yet not as much as for only emphasizing the target word. This suggests that a difference between the target word and its context was important for determining vowel perception [19].

Surprisingly, despite the objective improvements seen in the WER, the participants perceived a "target stretch" significantly ($\alpha = 0.05$) more intelligible for lax vowels (Tables 2 and 3, Figs. 3, 4). Moreover, "target stretch" also required less listening effort (although not significantly). Furthermore, although the "full stretch" TTS shows improvements in WER over the baseline for tense vowels, it was found to be significantly less respectful and encouraging in all cases. We also saw that the "full stretch" TTS was considered to be less natural and had worse prosody than the other TTS styles. This confirms our expectation that maintaining a natural (for L1 speakers) speech rate outside of difficult words can make the L2 listeners feel less patronized by the voice. Interestingly, the "target stretch" lax was rated as more natural and as having better prosody than "base" lax vowel, once again, despite have lower objective comprehension scores. These findings are particularly interesting as they suggest that L2 speakers may be overconfident in their ability to use the formant cues when the vowel is lengthened.

3.3.2. *Double word*

Once again, we computed WER as the proportion of correct target words for each TTS style. We also assessed differences between tense (tWER) and lax (lWER) vowels.

We observed the lowest WER with our proposed "clarity" TTS, both overall and for lax vowels, while vastly improving the performance for tense vowels over the baseline (relative 50% improvement) (Table 4, Fig. 5). Again, the "base" TTS showed a bias towards lax vowels (tWER: 30.00%, lWER: 18.66%), that we were able to overcome with our "clarity" mode. Additionally, we see that simply emphasizing all difficult words that had a tense/lax minimal pair decreases performance in lax vowels. Lastly, we observed "stretch" having improved performance over both "emphasis" and "base" for tense vowels. Yet, as in the "emphasis" condition, we saw participants

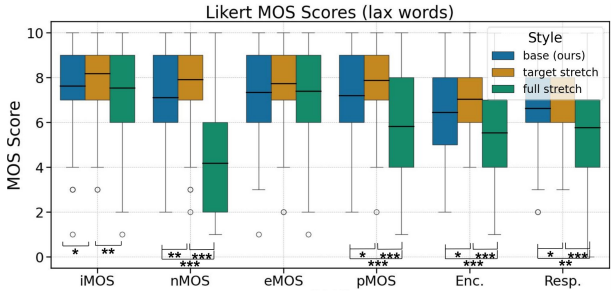

Figure 3: *Single target word results: Likert MOS scores of lax vowel containing words for intelligibility (iMOS), naturalness (nMOS), effort (eMOS), prosody (pMOS), encouragement (Enc) and respect (Resp.) for base (ours), target stretch, and full stretch. Stretching the target word has significantly higher perceived intelligibility than both full stretch as the baseline, despite having lower objective comprehension performance. Both the baseline and target word stretch has significantly higher naturalness, prosody, encouragement and respectfulness over full stretch, and target word stretch over the baseline.*

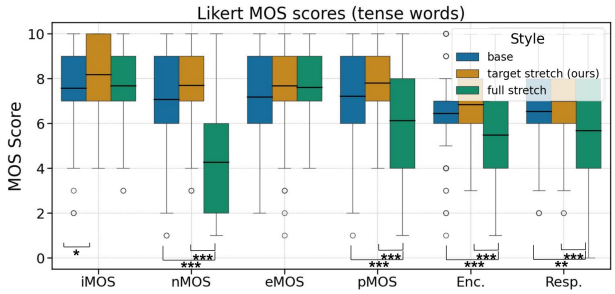

Figure 4: *Single target word results: Likert MOS scores of tense vowel containing words for intelligibility (iMOS), naturalness (nMOS), effort (eMOS), prosody (pMOS), encouragement (Enc) and respect (Resp.) for base, target stretch (ours) and full stretch. Stretching the target word has significantly higher perceived intelligibility over the baseline. Both the baseline and stretching the target word have significantly higher naturalness, prosody, encouragement and respectfulness than the full stretch.*

Table 4: *Double target word results: total word error rate, tense word error rate, and lax word error rate.*

TTS Style	WER ↓	tWER ↓	lWER ↓
Base	24.30%	30.00%	18.66%
Stretch	19.82%	17.99%	21.65%
Emphasis	24.44%	20.06%	28.82%
Clarity (Ours)	**15.15%**	**14.38%**	**15.92%**

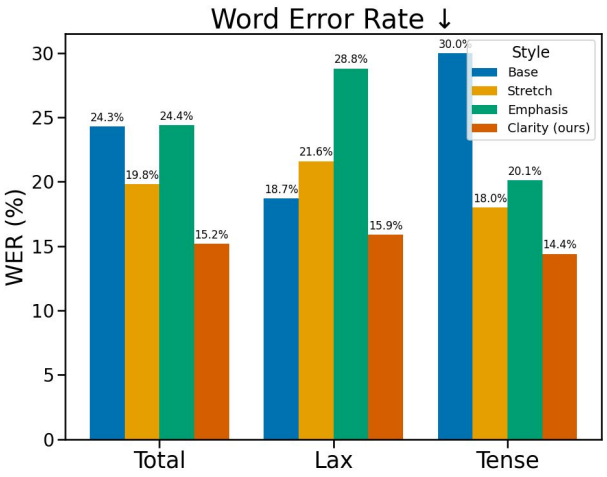

Figure 5: *Double target word results: our clarity mode has the lowest total WER (left), lax WER (center) and tense WER (right).*

struggle to understand the short, lax vowels. Although the tense vowel words between "emphasis" and "clarity" and lax vowel words between "clarity" and base had the same duration, "clarity" mode still had a lower WER when specifically comparing these words. This likely resulted from the phrases containing both a tense and a lax vowel, where the L2 participants could use duration differences between the two words to more easily differentiate the words. A more in-depth exploration of these differences remains a topic for future work.

We once again observed the fascinating result that the participants rate the "emphasis" TTS the highest in all categories (although in this case, the scores are not significantly higher than those for the "clarity" TTS) (Tables 5 and 6, and Fig. 6), despite the WER being higher for this TTS than both "clarity" and "stretch" TTS styles. We observed that "stretch" is less natural and has poorer prosody than all other TTS styles, despite showing objective improvements in WER over "emphasis" and "base". Lastly, we found that L2 participants rated both speaking too fast ("base") and too slow ("stretch") as being less respectful and less encouraging.

3.3.3. Whisper ASR

Speech synthesis studies often use human MOS scores but rely on ASR for transcription accuracy, as it correlates well with L1 intelligibility [30]. In this section, we explore how ASR relates to L2 performance and whether it uses the same duration cues as L2 participants.

We used Whisper ASR [31][1] with 72 phrases (generated in the same manner as for the human experiments) and calculated overall WER (WERt) and WER on only the target words. The phrases included those from the human study, and the additional phrases were constructed as in Sec. 3.1. We also included the

[1] v20231117, medium multilingual

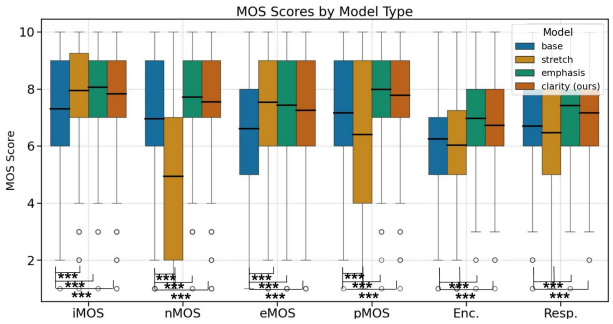

Figure 6: *Double target word results: Likert MOS scores of words containing tense vowels for intelligibility (iMOS), naturalness (nMOS), effort (eMOS), prosody (pMOS), encouragement (Enc) and respect (Resp.). The baseline is perceived as significantly less intelligible and requiring more listening effort than all of stretch, emphasis and clarity. Baseline and stretch are perceived as significantly less natural with worse prosody than both emphasis and clarity, with stretch lower than baseline. Both baseline and stretch (speaking too fast or too slow) are perceived as significantly less encouraging and respectful.*

percentage of errors in the target word resulting from minimal tense/lax pair substitution (sub) and what percentage of these substitutions were lax substituted for a tense (t-sub) and tense substituted for a lax (l-sub) vowel. Additionally, we ensured that homophones with the target words were accepted as correct transcriptions.

Similar to L2 speakers, we saw for the "base" TTS, Whisper struggled to predict the correct target words that lack context in the phrase (21.4% vs an expected 5-10% WER for this model), although the performance was slightly higher than that for the L2 speakers (Table 7, Fig 7). We did not see the same improvements in WER with the "clarity" TTS that we saw in the L2 participants. Instead, we saw an overall slowing down ("stretch") of the TTS decreases the WER in the target words. Yet, we also saw that while the WER on the target words decreased with this slowing down, the proportion of errors in the target word stemming from the minimal pair substitutions was much higher for the "base" TTS (71.42%, all other TTS $< 37\%$), while the overall difference in WER both on the whole phrase and the target words was within 3% for all TTS styles. This suggests that while slowing down slightly reduces the overall number of errors, the target words were being predicted as even further from the target, e.g. where "peel" was replaced with "pill" in the "base" TTS was replaced with "peaked" in the "stretch" TTS. Therefore, the ASR does not use the same duration mechanisms as humans when facing difficult predicting words.

4. Discussion and future work

While we improved L2 comprehension of difficult words using duration cues for tense/lax minimal pairs, our focus remains limited to vowel length. Future work should explore clarity for other vowels via spectral cues (e.g., formants), and extend to consonants through manipulations like pauses and stress to enhance attention [32]. Since we used a duration multiplier, clarity effects may depend on speech rate, which varies across speakers and could shift further with added expressivity such as emotion-laden speech. There may also be a minimum duration needed to perceive tense vowels, potentially unmet at faster rates. Understanding these limits—and how duration interacts with speech rate while preserving naturalness and respectfulness—remains

Table 5: *Double target word results: intelligibility, naturalness, effort, prosody, encouragement and respect.*

TTS Style	iMOS ↑	nMOS ↑	eMOS ↑	pMOS ↑	Enc. ↑	Resp. ↑
Base	7.30	6.95/***	6.60	7.16/***	6.25	6.70
Stretch	7.94***/	4.93	**7.53***/	6.40	6.02	6.46
Emphasis	**8.06***/	**7.71***/***	7.43***/	**7.98***/***	**6.97***/***	**7.42***/***
Clarity (Ours)	7.83***/	7.54***/***	7.25***/	7.77***/***	6.72***/***	7.16***/***

Significance values are presented as: Significance compared to base TTS/Significance compared to stretch TTS

Table 6: *Double target word results: ANOVA and Tukey test statistics on MOS Likert scores.*

TTS Style	F-Statistic (3,2504)	P-Value	Significant Tukey Results
iMOS	20.15	<.001	all/Base p<.001
nMOS	20.15	<.001	all/Stretch p<.001; Emphasis,Clarity/Base p<.001
eMOS	24.15	<.001	all/Base p<.001
pMOS	77.15	<.001	all/Stretch <.001; Emphasis,Clarity/Base p<.001
Enc.	35.74	<.001	Emphasis,Clarity/Stretch,Base p<.001
Resp.	35.87	<.001	Emphasis,Clarity/Stretch, Base p<.001

Tukey results are presented as: Higher value/Lower value

Table 7: *Whisper ASR results: overall word error rate, target word error rate, tense/lax substitutions, lax substituted for a tense, tense substituted for a lax*

TTS Style	WERt ↓	WER ↓	sub	t-sub	l-sub
Base	17.10%	21.4%	71.42%	61.9%	9.52%
Stretch	**15.98**%	**19.38**%	36.83%	31.57%	5.26%
Emphasis	16.26%	22.4%	31.81%	22.72%	9.09%
Clarity (Ours)	17.68%	24.49%	29.16%	29.16%	0%

an open area for investigation.

Further, through explorations of the results on the participant level in [22], we found that the duration mechanism was not clearly universal; Rather, it appears to be very strong in certain participants and weak or non-existent in others, which echoes work showing inter-individual differences in the weighting of L2 acoustic cues in speech perception [33]. This suggests avenues for future work, which could harness pronunciation data to further customize TTS to individual listeners. Moreover, since our work used force choice to limit participant responses, a topic of future work is understanding how increasing clarity in a target word affects the rest of the phrase and how robust this duration mechanism is to broader linguistic contexts. Additionally, since our participants had a relatively high level of English proficiency, future work should explore "L2 clarity TTS" with participants of lower proficiency levels, which could explore the modification of words that are less likely to be in the vocabulary of L2 speakers (e.g., "cooed" vs. "could"), using individual vocabularies as a factor predicting perception. Lastly, even more fine-grained control over a TTS [34] could aid in uncovering other possible mechanisms to aid L2 perception, specifically for how formant control and duration control can be combined.

5. Conclusions

This study resulted in multiple interesting findings. First, we provided a "clarity mode" as an open-source addition to Matcha-TTS and confirmed that, by applying a stretch to tense vowels for difficult target words, L2 speakers' transcription errors for these words were reduced. Indeed, emphasizing all target words or simply slowing the whole phrase without consideration of the linguistic properties of the vowel reduced the transcription performance of L2 speakers (from French L1 backgrounds), who primarily use duration to determine the difference between English tense and lax vowels. Moreover, we confirm prior findings that slowing down an entire phrase can be seen as less respectful and encouraging to L2 listeners.

Second, we found that our sample of L2 speakers are unaware they are using this duration mechanism. Through MOS scores, these L2 speakers indicated that they could more easily identify a stretched word, perhaps because they believed they could use the longer vowels (more stable formants) to transcribe a word more easily. This suggests that L2 listeners are limited in their ability to evaluate the effectiveness of adaptive TTS systems, and that objective metrics such as word error rate should serve as a gold standard in future work.

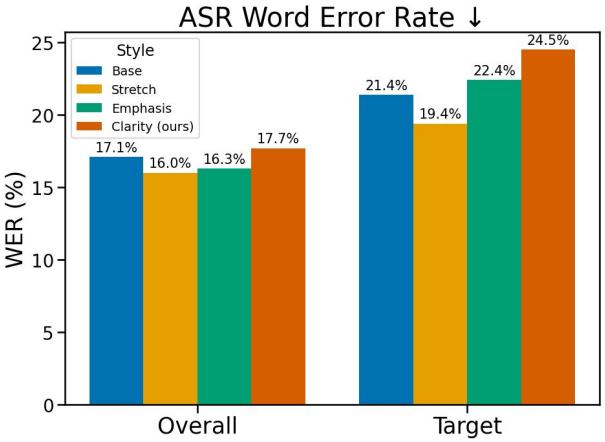

Figure 7: *Whisper ASR results: We observe that ASR does not align with L2 perception (Fig. 5). Stretch had the lowest WER both for the whole phrase (left) and the target words (right).*

Lastly, we found that Whisper ASR does not use the same duration mechanisms as L2 speakers and, therefore, does not present an adequate replacement for determining the transcription accuracy of synthesized speech for these individuals. This suggests, again, that future development of adaptive TTS systems for L2 speakers must, at the point in time, rely more on data from real human listeners.

Overall, these results have important consequences for TTS assessments, especially for L2 speakers. We cannot simply rely on the same methods used for L1 speaker TTS assessments. Neither user self-rated intelligibility assessments nor automatic systems reflect the true accuracy of difficult vowel perception in L2 speakers, even for those with high proficiency as in our sample, and improved accuracy does not necessarily reflect positive perceptions of the voice (as for "stretch"). Researchers must use both objective and subjective assessments with human participants to ensure they are building inclusive and accessible speech synthesis systems.

6. Acknowledgements

This work was supported by the Simon Fraser University FASS Breaking Barriers Interdisciplinary Incentive Grant, the Social Sciences and Humanities Research Council of Canada Grant (SSHRC Insight Grant 435–2019–1065), and the NSERC Discovery Grant (RGPIN-2024-06519). The authors thank Paul Maublanc for always being our first pilot French speaker, as well as the Rajan Family for their support.

7. References

[1] A. D. Vieira, H. Leite, and A. V. L. Volochtchuk, "The impact of voice assistant home devices on people with disabilities: A longitudinal study," *Technological Forecasting and Social Change*, vol. 184, p. 121961, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0040162522004826

[2] G. A. A. de Oliveira, O. d. F. Oliveira, S. de Abreu, R. W. de Bettio, and A. P. Freire, "Opportunities and accessibility challenges for open-source general-purpose home automation mobile applications for visually disabled users," *Multimedia Tools Appl.*, vol. 81, no. 8, p. 10695–10722, Mar. 2022. [Online]. Available: https://doi.org/10.1007/s11042-022-12074-0

[3] A. J. London, Y. S. Razin, J. Borenstein, M. Eslami, R. Perkins, and P. Robinette, "Ethical issues in near-future socially supportive smart assistants for older adults," *IEEE Transactions on Technology and Society*, vol. 4, no. 4, pp. 291–301, 2023.

[4] W. Seymour, X. Zhan, M. Coté, and J. Such, "A systematic review of ethical concerns with voice assistants," in *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 131–145. [Online]. Available: https://doi.org/10.1145/3600211.3604679

[5] E. Martinson and D. Brock, "Improving human-robot interaction through adaptation to the auditory scene," in *HRI*, 2007, p. 113–120.

[6] J. Xiao, J. Liu, D. Li, L. Zhao, and Q. Wang, "Speech intelligibility enhancement by non-parallel speech style conversion using cwt and imetricgan based cyclegan," in *MultiMedia Modeling*, 2022, pp. 544–556.

[7] S. Novitasari, S. Sakti, and S. Nakamura, "Dynamically adaptive machine speech chain inference for tts in noisy environment: Listen and speak louder," in *Interspeech*, 2021, pp. 4124–4128.

[8] M. Cohn and G. Zellou, "Perception of concatenative vs. neural text-to-speech (tts): Differences in intelligibility in noise and language attitudes," in *Interspeech 2020*, 2020, pp. 1733–1737.

[9] C. Valentini-Botinhao and J. Yamagishi, "Speech enhancement of noisy and reverberant speech for text-to-speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 8, pp. 1420–1433, 2018.

[10] C. Redmon, K. Leung, Y. Wang, B. McMurray, A. Jongman, and J. A. Sereno, "Cross-linguistic perception of clearly spoken english tense and lax vowels based on auditory, visual, and auditory-visual information," *Journal of phonetics*, vol. 81, p. 100980, 2020.

[11] N. B. Aoki and G. Zellou, "Being clear about clear speech: Intelligibility of hard-of-hearing-directed, non-native-directed, and casual speech for l1- and l2-english listeners," *Journal of Phonetics*, vol. 104, p. 101328, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0095447024000342

[12] K. Rothermich, H. L. Harris, K. Sewell, and S. C. Bobb, "Listener impressions of foreigner-directed speech: A systematic review," *Speech Communication*, vol. 112, pp. 22–29, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639319300676

[13] M. Monnot, "La prononciation du français contemporain," *The French Review*, vol. 48, no. 1, pp. 284–285, 1974.

[14] W. Strange, A. Weber, E. S. Levy, V. Shafiro, M. Hisagi, and K. Nishi, "Acoustic variability within and across german, french, and american english vowels: Phonetic context effects," *The Journal of the Acoustical Society of America*, vol. 122, no. 2, pp. 1111–1129, 2007.

[15] J. L. Miller, M. Mondini, F. Grosjean, and J.-Y. Dommergues, "Dialect effects in speech perception: The role of vowel duration in parisian french and swiss french," *Language and speech*, vol. 54, no. 4, pp. 467–485, 2011.

[16] P. Inverson, M. Pinet, and B. G. Evans, "Auditory training for experienced and inexperienced second-language learners: Native french speakers learning english vowels," *Applied psycholinguistics*, vol. 33, no. 1, pp. 145–160, 2012.

[17] Y. Kita and Y. Kita, "Japanese learners of english and japanese phonology," *Research bulletin of Naruto University of Education*, vol. 34, pp. 209–216, 2019.

[18] Y.-A. Lu and S.-I. Lee-Kim, "The effect of linguistic experience on perceived vowel duration: Evidence from taiwan mandarin speakers," *Journal of Phonetics*, vol. 86, p. 101049, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0095447021000218

[19] P. Tuttösí, H. H. Yeung, Y. Wang, F. Wang, G. Denis, J.-J. Aucouturier, and A. Lim, "Mmm whatcha say? uncovering distal and proximal context effects in first and second-language word perception using psychophysical reverse correlation," in *Interspeech 2024*, 2024, pp. 1010–1014.

[20] A. A. Jr and J. Lovell, "Stimulus features in signal detection," *The Journal of the Acoustical Society of America*, vol. 49, no. 6B, pp. 1751–1756, 1971.

[21] D. Kewley-Port, O.-S. Bohn, and K. Nishi, "The influence of different native language systems on vowel discrimination and identification," *The Journal of the Acoustical Society of America*, vol. 117, no. 4_Supplement, pp. 2399–2399, 2005.

[22] P. Tuttösí, "I know you're listening: Adaptive voice for hri," PhD thesis, Simon Fraser University, Burnaby, BC, Canada, May 2025, available at https://arxiv.org/pdf/2506.15107.

[23] C. Redmon, K. Leung, Y. Wang, B. McMurray, A. Jongman, and J. A. Sereno, "Cross-linguistic perception of clearly spoken english tense and lax vowels based on auditory, visual, and auditory-visual information," *Journal of phonetics*, vol. 81, p. 100980, 2020.

[24] S. Mehta, R. Tu, J. Beskow, É. Székely, and G. E. Henter, "Matcha-TTS: A fast TTS architecture with conditional flow matching," in *ICASSP*, 2024.

[25] A. Joly, M. Nicolis, E. Peterova, A. Lombardi, A. Abbas, A. van Korlaar, A. Hussain, P. Sharma, A. Moinet, M. Lajszczak, P. Karanasou, A. Bonafonte, T. Drugman, and E. Sokolova, "Controllable emphasis with zero data for text-to-speech," in *12th ISCA Speech Synthesis Workshop (SSW2023)*, 2023, pp. 113–119.

[26] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, "Grad-tts: A diffusion probabilistic model for text-to-speech," in *ICML*, 2021, p. 8599–8608.

[27] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-tts: A generative flow for text-to-speech via monotonic alignment search," in *Neurips*, 2020, pp. 8067–8077.

[28] G. Ma, P. Hu, J. Kang, S. Huang, and H. Huang, "Leveraging phone mask training for phonetic-reduction-robust e2e uyghur speech recognition," in *Interspeech 2021*, 2021, pp. 306–310.

[29] J. R. Lewis and IBM.Corp, "Investigating mos-x ratings of synthetic and human voices," *Voice Interaction Design*, vol. 2, no. 1, p. 22, 2018.

[30] J. Taylor and K. Richmond, "Confidence intervals for asr-based tts evaluation," in *Interspeech 2021*, 2021, pp. 2791–2795.

[31] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022.

[32] N. Nagata, "The effects of silent pauses on listening comprehension: a case of japanese learners of english as a foreign language," *Doctoral dissertation, Waseda University*, 2002.

[33] J. Schertz, T. Cho, A. Lotto, and N. Warner, "Individual differences in phonetic cue use in production and perception of a non-native sound contrast," *Journal of phonetics*, vol. 52, pp. 183–204, 2015.

[34] C. Tånnander, S. Mehta, J. Beskow, and J. Edlund, "Beyond graphemes and phonemes: continuous phonological features in neural text-to-speech synthesis," in *Interspeech 2024*, 2024, pp. 2815–2819.