

Improving TTS for Shanghainese: Addressing Tone Sandhi via Word Segmentation

Yuanhao Chen

yuanhao.chen.25@dartmouth.edu

Abstract

Tone is a crucial component of the prosody of Shanghainese, a Wu Chinese variety spoken primarily in urban Shanghai. Tone sandhi, which applies to all multi-syllabic words in Shanghainese, then, is key to natural-sounding speech. Unfortunately, recent work on Shanghainese TTS (text-to-speech) such as Apple’s VoiceOver has shown poor performance with tone sandhi, especially LD (left-dominant sandhi). Here I show that word segmentation during text preprocessing can improve the quality of tone sandhi production in TTS models. Syllables within the same word are annotated with a special symbol, which serves as a prosodic annotation for the domain of LD. Contrary to the common practice of using prosodic annotation mainly for static pauses, this paper demonstrates that prosodic annotation can also be applied to dynamic tonal phenomena. I anticipate this project to be a starting point for bringing formal linguistic accounts of Shanghainese into computational projects. Too long have we been using the Mandarin models to approximate Shanghainese, but it is a different language with its own linguistic features, and its digitisation and revitalisation should be treated as such.

1 Introduction

Shanghainese is a variety of Wu Chinese spoken primarily in urban Shanghai and globally by the Shanghainese diaspora.

Despite its formerly prominent status as a lingua franca in the Yangtze River Delta region, Shanghainese is now a minority language in Shanghai, with Putonghua (Standard Mandarin) being the dominant language in the city. Furthermore, the situation is only exacerbated by the general sentiment among the younger generation that Shanghainese is a “low-status” language, and by their adoption of the linguistic model that one nation should use only one language (Gilliland, 2006). Education plays a crucial role in this process, as this

sentiment was mainly observed among college students. As a result, most young people in Shanghai, whether native or an immigrant, are unable to speak Shanghainese fluently (Weng, 2023).

With Putonghua being the perceived authentic and superior language in many aspects of life, crucially including education, it is direly important to preserve the linguistic variety in Shanghai by promoting the use of Shanghainese. Digitisation of a substratum is an effective way to promote the language in teaching, learning, and various other dimensions of cultural life (Villa, 2002).

In this project, I aim to build a TTS (text-to-speech) system for Shanghainese, which is a crucial component in the digitisation of a language, serving as a bridge between the digitised written spoken forms of the language. This is not to say that there is no existing work on Shanghainese TTS. Notably, Apple Inc. (2017) added Shanghainese to the list of languages supported by VoiceOver, the screen reader built into Apple’s operating systems. However, the quality of the synthesised speech is not satisfactory, and definitely not on par with the quality of the synthesised speech for other Sinitic languages such as Putonghua. The main problem with Shanghainese VoiceOver is its occasional poor performance with tone sandhi, especially LD (left-dominant sandhi), a suprasegmental phonological process involving a specific bounding domain (Roberts, 2020). For example, the word $/[zā^2 he^334]_{LD\ domain}/$ ‘Shanghai’ has to be pronounced with LD as $[zā^2.he^4]$ (the left syllable’s rising contour is spread over to the right one).

This paper will explore the possibility of improving tone sandhi in Shanghainese TTS by putting focus on annotating the bounding domain of LD during preprocessing of input texts. Specifically, I will segment the input text into lexical words. My results confirm that this approach is effective in improving the quality of synthesised speech in terms of tone sandhi.

2 Methodology

2.1 Overview

The key to improving LD in Shanghai Chinese TTS is to annotate the bounding domain of LD. Instead of training a model for this task, which is difficult due to lack of resources, I will perform word segmentation, because lexical words highly correlates with the domains for LD (Kuang and Tian, 2019); formally, LD domains can be formed by the left edges of lexical words, with a few exceptions (Roberts, 2020). Thus, this prosodic annotation can be transformed into word segmentation, giving us the overall pipeline of this paper as shown in Fig. 1.

2.2 Datasets

The data basis of TTS models is a list of corresponding audio files and transcriptions. I am using a dataset of an ASR project (Cosmos-Break, 2023), which contains 2,012 audio files and corresponding transcriptions in Chinese characters, totalling 5,607 seconds of speech of a single Shanghai Chinese speaker. The types of speech in the dataset range from single words to phrases and sentences. The audio is resampled to 16kHz for training.

For word segmentation and phonemisation, we are going to need a phonemically annotated lexicon of Shanghai Chinese. I am using one containing more than 125,000 lexical entries, 51,000 of which have corresponding romanisations (Chen, 2022).

2.3 Word Segmentation

The `jieba` library is the most popular open-source Chinese word segmentation library, which comes with a Mandarin dictionary out of the box (Junyi, 2023). It implements two models, namely a trie of a deterministic finite automaton (DFA) pre-built from the dictionary, and a hidden Markov model (HMM) with Viterbi algorithm as a backup for unknown words. We will have to rely on the Mandarin dictionary because word-segmentation algorithms require word frequency information, which is not available in the Shanghai Chinese lexicon, but we can patch the model by adding weights to Shanghai Chinese-specific words on top of Mandarin weights. This works well, as written Mandarin and Shanghai Chinese are very similar.

2.4 Phonemisation

The aforementioned dictionary (Chen, 2022) is used to romanise segmented words in *Yahwe Wu Chinese Romanisation* (吳語協會式拼音). As

the dictionary only contains traditional forms, before romanising, the words are converted to Traditional Chinese using Kuo (2023). Then, Qieyun (Mikazuki, 2022) is used to add any necessary tone numbers to the romanisation. The romanisation is then converted to broad IPA transcription largely following the paradigm of Qian (2007). Because the goal is not to accurately transcribe phonetically but to effectively feed the TTS model with phonemic contrasts, some notational techniques are employed to reduce the number of ambiguous digraphs, such as using ⟨c⟩ for /tʃ/ and ⟨j⟩ for /dʒ/.

2.5 Training the TTS Model

The TTS model presented in this paper is trained using the VITS end-to-end TTS model (Kim et al., 2021). Compared to previous popular TTS models such as Glow-TTS (Kim et al., 2020), VITS employs a variational autoencoder (VAE) to produce a latent model of the input text and a stochastic duration predictor, which allows the model to express the natural one-to-many relationship in which a text input can be spoken in multiple ways with different pitches and rhythms.

This is perfect for the task that this paper aims to accomplish: The model can effectively learn that the same string in the input may be pronounced with a different pitch depending on whether tone sandhi applies, i.e., depending on the prosodic environment.

The pre-processed text and audio are fed into the VITS model, which is trained for 50K steps with a batch size of 32.

3 Experiments

I conducted an $n \times 3 \times 5 \times 4$ experiment, where $n = 11$ is the number of native Shanghai Chinese participants, 3 is the number of speakers generating the audio samples, 5 is the number of sentences, and 4 is the number of metrics. A subjective human evaluation (MOS, mean opinion score) is conducted under each condition on a 1–5 scale, with 5 being the best (see details in Fig. 8).

The three speakers are

- (1) The model presented by this paper.
- (2) Shanghai Chinese VoiceOver (Apple Inc., 2017).
- (3) This paper’s author, a native speaker.

The four metrics are as proposed by Cardoso et al. (2015):

- (1) Comprehensibility: How well can you understand the meaning of the audio?

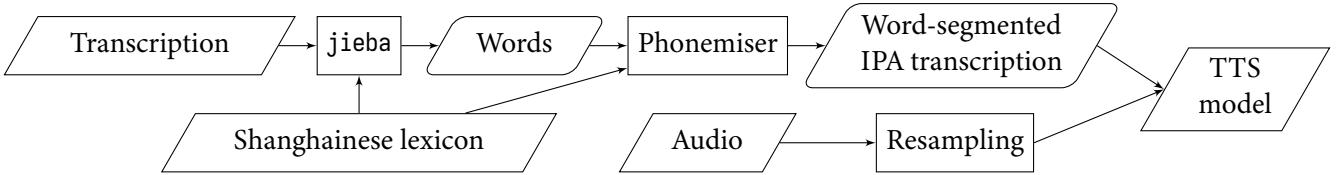


Figure 1: Overview of the pipeline in training the TTS model, with emphasis on the text preprocessing steps.

- 181 (2) Naturalness: How natural does the audio
 182 sound?
 183 (3) Accuracy: How well does the audio match how
 184 a native speaker like you would pronounce it?
 185 (4) Intelligibility: How much effort does it take to
 186 make sense of the audio?

187 4 Results

188 4.1 Tone Sandhi production

189 In all test sentences (containing 23 different LD
 190 domains to trigger sandhi), all speakers produce the
 191 correct tone sandhi, except for VoiceOver, which
 192 fails to produce the correct tone sandhi in sentence
 193 5, 儂弗要弗二弗三個 ‘don’t be nasty’.

194 弗二弗三個 ‘nasty’ / [və?¹².ni²³.və?¹².se⁵¹.fə?²]_{LD} /
 195 is a word forming a pentasyllabic LD
 196 domain, which should surface tonally as
 197 [və?¹.ni³.və?².se².fə?¹], but VoiceOver incor-
 198 rectly treats it as / [və?¹².ni²³]_{LD} [[və?¹².se⁵¹.fə?²]_{LD}
 199 ...]_{RD} /, splitting the word into two LD domains,
 200 and assuming that an extra RD (right-dominant
 201 sandhi) domain dominates the second LD do-
 202 main, which is doubly wrong. This produces
 203 the incorrect surface form [və?¹.ni³ və?².se¹.fə?²].
 204 Compare the pitch in Figs. 6 and 7.

205 4.2 Questionnaire

206 One out of 11 questionnaires collected is discarded
 207 for being incomplete. The valid $10 \times 3 \times 5 \times 4$ opinion
 208 scores are used for various statistical analyses.

209 The overall MOS of three speakers are shown in
 210 Table 1. Speakers 1 and 2 are not significantly differ-
 211 ent from each other ($p = 0.64$); they both have sig-
 212 nificantly lower scores than speaker 3 ($p \ll 0.001$). A
 213 breakdown of scores grouped by metrics shows
 214 similar statistical relationships between speakers
 215 (see Table 2).

216 In a breakdown of scores grouped by sentences,
 217 two statistically significant differences are found:
 218 Speaker 1 has a significantly lower score than
 219 speaker 2 in sentence 2, and speaker 2 has a sig-
 220 nificantly lower score than speaker 1 in sentence 5
 221 (both $p \ll 0.001$; see Table 3).

222 5 Discussion

223 From the comparison of overall MOS and the com-
 224 parison of MOS grouped by metrics, we can see
 225 that the human acceptability of this model is gen-
 226 erally comparable to that of VoiceOver. Statisti-
 227 cally significant differences emerge when we break
 228 down the scores by sentences, showing that both
 229 models have their own strengths and weaknesses.

230 5.1 Shortcomings of this model

231 As shown in Table 3, this model has a significantly
 232 lower score than VoiceOver in sentence 2, 上海是
 233 一座國際化大都市 ‘Shanghai is an international
 234 metropolis’.

235 There are tonal differences between what this
 236 model (Fig. 2) produces for 國際化大都市 ‘inter-
 237 national metropolis’ /ko?⁵.tɕi⁵³.ho³³⁴ da²³.tu⁵³.zz?²³ /
 238 and what VoiceOver (Fig. 3) and I (Fig. 4) pro-
 239 duce. However, this is not the reason for the
 240 low score, because this phrase indeed has two
 241 alternative prosodic segmentation possibilities:
 242 /[[ko?⁵.tɕi⁵³.ho³³⁴]_{LD} [da²³.tu⁵³.zz?²³]_{LD}]_{RD} / (what
 243 this model goes for) and /[[ko?⁵.tɕi⁵³.ho³³⁴]_{LD}]_{RD}
 244 [[da²³.tu⁵³.zz?²³]_{LD}]_{RD} / (what the other speakers go
 245 for), depending on whether ‘international’ and
 246 ‘metropolis’ are grouped together in an RD do-
 247 main.

248 The main issue with the pronunciation of this
 249 model is that the syllable /ko?/ is produced too long
 250 compared to syllables without /-?/ coda. In Shang-
 251 hainese, the glottal stop coda is often realised as
 252 the shortening of the nucleus (giving [kə]), instead
 253 of actually pronounced. However, this model pro-
 254 duces a syllable clearly longer than others, with a
 255 duration of 0.28 s, or 10% of total speech duration,
 256 whereas the native speaker produces this with a du-
 257 ration of 0.16 s, or 6% of total speech duration.

258 This behaviour is possibly due to VITS’s
 259 makeshift treatment of blanks in speech — VITS
 260 inserts a BLANK token between every input char-
 261 acter, which supposedly enhances performance
 262 in general (Kim et al., 2021). 25% of the duration
 263 of /ko?/ is actually nearly silence at the end of the

264 syllable, a weird pause to have within a word.

265 **5.2 Advantages of this model**

266 The model presented by this paper is significantly
267 better than VoiceOver in sentence 5, because it cor-
268 rectly handles LD of $[\text{və}?^{12}.\text{ni}^{23}.\text{və}?^{12}.\text{se}^{51}.\text{fə}?]_{\text{LD}}$,
269 which VoiceOver fails (compare Figs. 5 and 6 to the
270 native speaker version in Fig. 7).

271 This difference is manifested by the mostly cor-
272 rect word segmentation of this model, which out-
273 puts $\text{və}?-\text{ni}=\text{və}?= \text{se}1 \text{ gə}?$, where hyphens connect
274 syllables in a known word and double hyphens
275 connect those in an inferred word; $\text{gə}?$ (/fə?/
276 when cliticised) is left out but can be inferred from
277 context by the TTS model because it is a com-
278 mon suffix that clings to the previous word. Ad-
279 mittedly, this result might be improved even fur-
280 ther if we train a dedicated prosodic annotation
281 model that can detect clitics, rather than just word-
282 segmentation, but the current result is already sat-
283 isfactory.

284 The output of VoiceOver is also correct under
285 a low-high pitch accent analysis of Shanghainese
286 (Roberts, 2020), which predicts a pitch contour
287 of LHLLL for this word. However, to sound
288 more tonally natural, one has to correctly formu-
289 late LD domains to get the exact pitch contour,
290 which further highlights the importance of accu-
291 rate prosodic structure in Shanghainese.

292 **5.3 Ethical considerations of this project**

293 The nature of TTS models is to mimic human
294 speech as perfectly as possible. With progress in
295 more natural-sounding TTS models such as the
296 one in this paper, it is possible to use TTS models
297 to impersonate other people’s voices, which can be
298 used for malicious purposes such as fraud. More
299 concerningly, voice conversion is easy to do with
300 TTS models with VAE such as VITS, which dras-
301 tically widens the potential scope of malicious use
302 to any person. While driving the digitisation and
303 revitalisation of a minority language like Shang-
304 hainese, we should simultaneously be aware of the
305 potential harm that technology can bring, and def-
306 initely refrain from any malicious application.

307 **6 Conclusion**

308 In this work, I have presented a TTS model for
309 Shanghainese with the novel approach of empha-
310 sising bounding domains of tone sandhi, specifi-
311 cally LD, during text preprocessing. Due to lack

312 of material to train a dedicated annotation model,
313 word segmentation is employed as a proxy for this
314 phonological information, which is shown to be ef-
315 fective in improving the tone sandhi quality of the
316 output speech compared to Apple Inc. (2017). Fur-
317 ther improvement in performance of timing and
318 pausing may be achieved by switching to a TTS
319 model that handles blanks in speech better.

320 Beyond just prosody, the significance of this
321 project should be to raise awareness of the impor-
322 tance of a formal linguistic account in every aspect
323 of the development of computational systems re-
324 garding Shanghainese. For example, in the dataset
325 used in this project is originally for an ASR project
326 (Cosmos-Break, 2023), but the transcription is scat-
327 tered with 假借 (phonetic loan characters), where a
328 character is used for its Mandarin pronunciation to
329 approximate the “dialectic” pronunciation, likely
330 because the transcriber read fluently only in Man-
331 drin. For example, 薩 (a surname, Mandarin /sa/) is
332 used for 啥 (‘what’, Shanghainese /sa/); such
333 practice is common but greatly hinders a consist-
334 ent and formal treatment of Shanghainese orthog-
335 raphy and lexicon in computational systems, as the
336 character used to approximate varies from person
337 to person, and the phenomenon itself is a manifes-
338 tation of Mandarin centralism which marginalises
339 Shanghainese.

340 Specific to the topic of this project, the lack of a
341 computational model implemented as per a formal
342 linguistic account of Shanghainese tone system is
343 a major obstacle to the improvement of tonal per-
344 formance in TTS. Even word segmentation, which
345 is a makeshift solution of prosodic annotation, is
346 carried out by the makeshift approach of using the
347 Mandarin-pre-trained jieba model.

348 In general, despite the low-resource status of
349 most substrata, it is important to be alerted that
350 over-reliance on resources of the superstratum is
351 only a makshift solution that can both under-
352 mine the authenticity of the result and take the
353 focus away from the development of the scaffolding
354 (implementation of formal linguistic accounts)
355 of substrata. Therefore, to engineer reliable and
356 minority-friendly computational systems, further
357 research should really put the development of the
358 scaffolding of substrata at the top of the agenda.

359 **References**

360 Apple Inc. 2017. VoiceOver. Apple Inc.

- 361 Walcir Cardoso, George Smith, and Cesar Gar-
362 cia Fuentes. 2015. *Evaluating text-to-speech*
363 *synthesizers*. In *Critical CALL – Proceedings of the*
364 *2015 EUROCALL Conference, Padova, Italy*, pages
365 108–113. Research-publishing.net.
- 366 Yuanhao Chen. 2022. *Rime Yahwe Zaonhe*. Zenodo.
- 367 Cosmos-Break. 2023. *Shanghainese ASR*.
- 368 Joshua Gilliland. 2006. *Language Attitudes and Ideolo-*
369 *gies in Shanghai, China*. Ph.D. thesis, The Ohio State
370 University.
- 371 Sun Junyi. 2023. *Jieba*.
- 372 Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sun-
373 groh Yoon. 2020. *Glow-TTS: A Generative Flow for*
374 *Text-to-Speech via Monotonic Alignment Search*.
- 375 Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. *Con-*
376 *ditional Variational Autoencoder with Adversarial*
377 *Learning for End-to-End Text-to-Speech*.
- 378 Jianjing Kuang and Jiapeng Tian. 2019. *Tone Represen-*
379 *tation and Tone Processing in Shanghainese*. In *Pro-*
380
381 *Sciences*, Melbourne, Australia.
- 382 Carbo Kuo. 2023. *OpenCC (Open Chinese Convert 開*
383 *放中文轉換*).
- 384 Ayaka Mikazuki. 2022. *Qieyun-python*. nk2028.
- 385 Nairong Qian. 2007. *Shanghai Fangyan (Shang-*
386 *hainese)*, first edition. Haipai Wenhua Congshu.
387 Wenhui Press, Shanghai.
- 388 Brice David Roberts. 2020. *An Autosegmental-Metrical*
389 *Model of Shanghainese Tone and Intonation*. Ph.D.
390 thesis, UCLA.
- 391 Daniel J. Villa. 2002. *Integrating technology into minor-*
392 *ity language preservation and teaching efforts: An in-*
393 *side job*. *Language Learning & Technology*, 6(2).
- 394 Shihong Weng. 2023. *The second generation of “New*
395 *Shanghainese”: Their language and identity*. *San*
396 *Diego Linguistic Papers*, 12.

A Tables

Speaker	Overall MOS
1	4.14 ± 0.12
2	4.19 ± 0.14
3	4.83 ± 0.06

Table 1: Overall MOS of three speakers. Confidence interval: 95%; same for tables below.

Speaker	Accuracy	Comprehensibility	Intelligibility	Naturalness
1	4.06 ± 0.26	4.48 ± 0.17	4.36 ± 0.22	3.66 ± 0.28
2	4.02 ± 0.36	4.58 ± 0.18	4.38 ± 0.24	3.76 ± 0.32
3	4.82 ± 0.11	4.82 ± 0.11	4.86 ± 0.10	4.82 ± 0.14

Table 2: MOS of three speakers by metrics.

Speaker	Sentence 1	Sentence 2	Sentence 3	Sentence 4	Sentence 5
1	4.35 ± 0.22	3.70 ± 0.34	3.85 ± 0.28	4.28 ± 0.25	4.53 ± 0.24
2	4.45 ± 0.24	4.70 ± 0.18	4.15 ± 0.26	4.55 ± 0.24	3.08 ± 0.41
3	4.83 ± 0.12	4.82 ± 0.12	4.88 ± 0.13	4.75 ± 0.16	4.88 ± 0.11

Table 3: MOS of three speakers by sentences.

B Figures

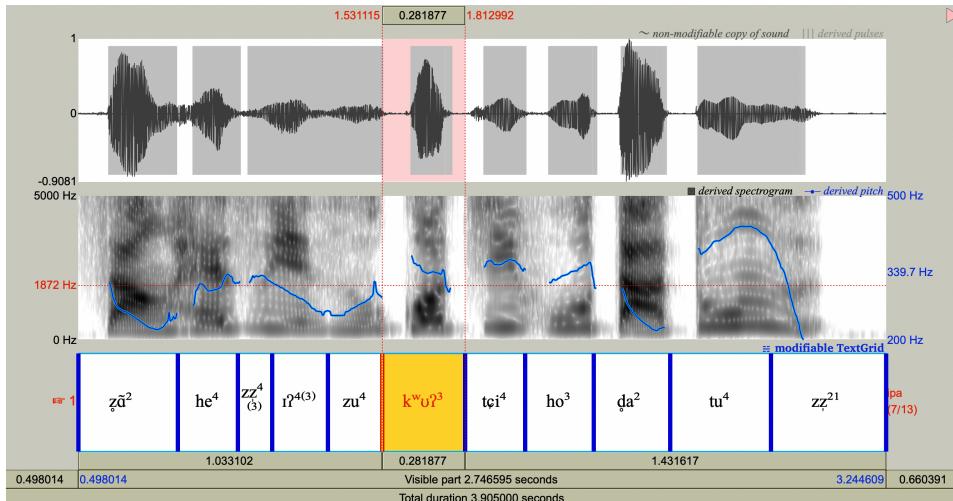


Figure 2: Sentence 2 by speaker 1 with broad phonetic annotation. Tones are represented by Chao tone letters as their phonetic realisations. Parentheses indicate uncertain tone height that is within acceptable range and not crucial to the analysis. Same for figures below.

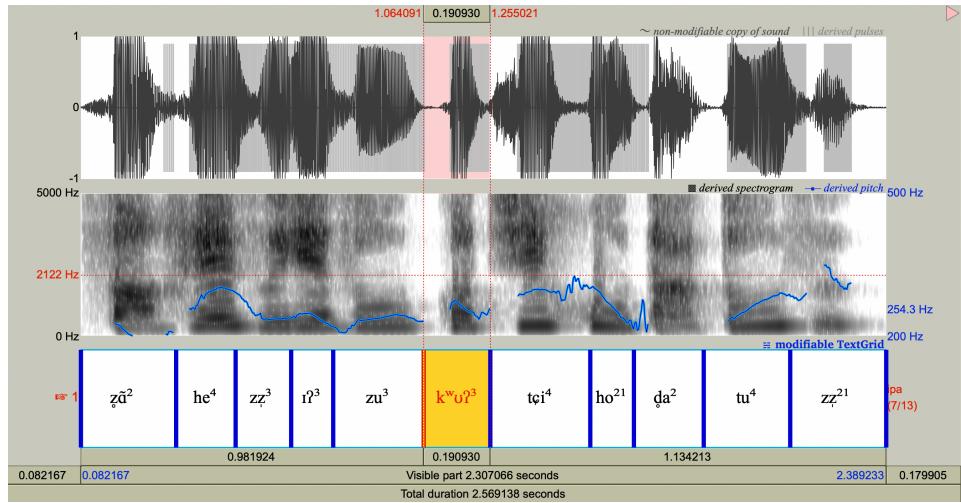


Figure 3: Sentence 2 by speaker 2.

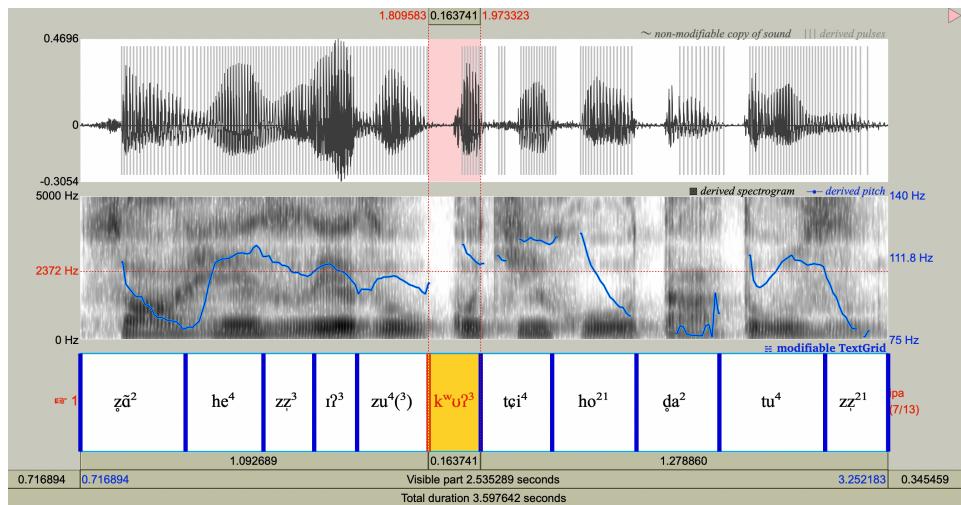


Figure 4: Sentence 2 by speaker 3.

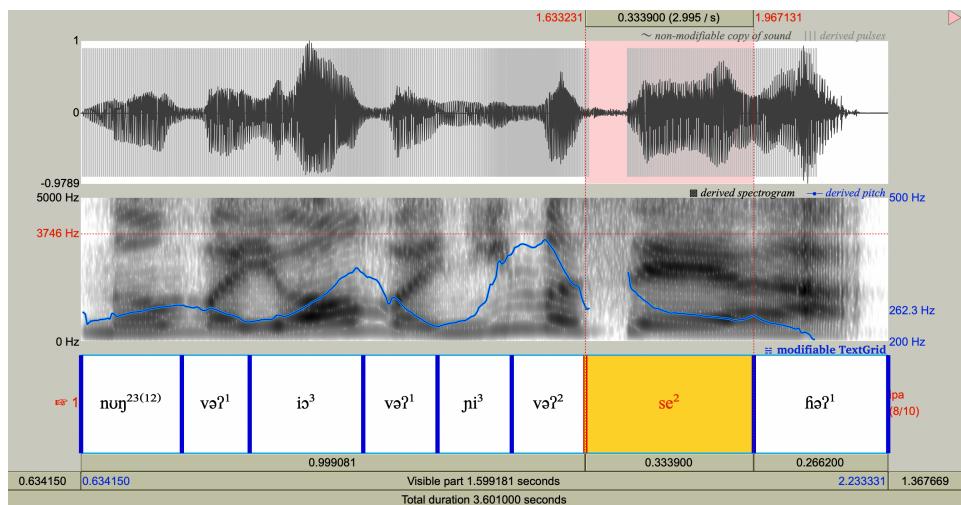


Figure 5: Sentence 5 by speaker 1.

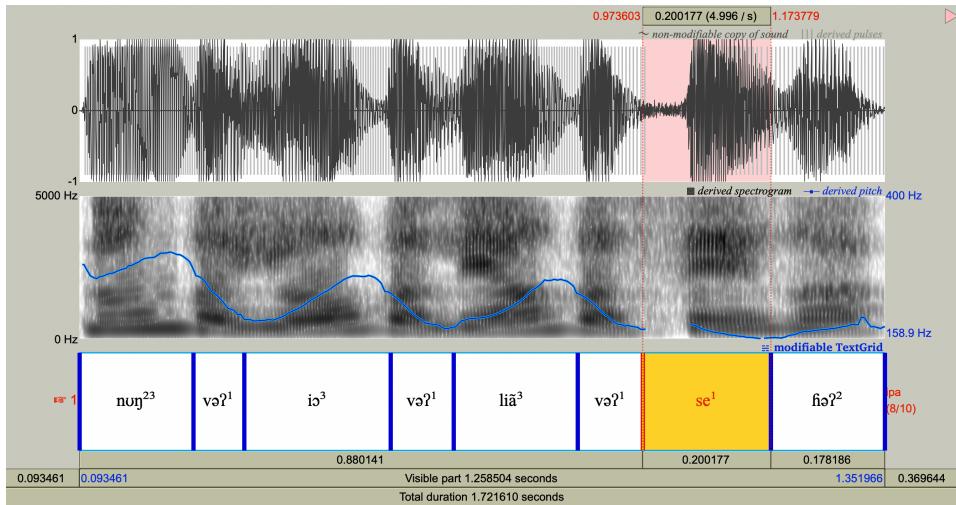


Figure 6: Sentence 5 by speaker 2.

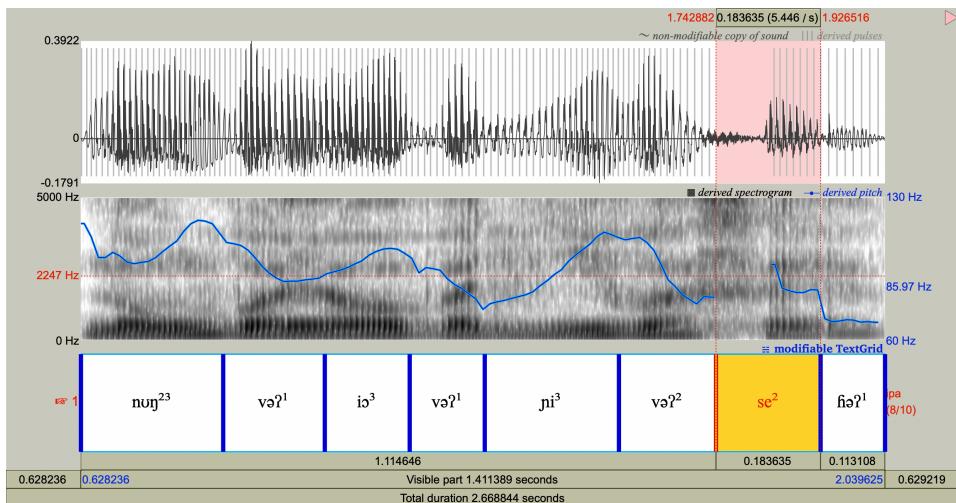


Figure 7: Sentence 5 by speaker 3.

C Questionnaire

Dartmouth Ling 48 Final Project: Improving TTS for Shanghainese
 這項茅斯語言學48期末項目：改進上海話文字轉語音
 Compare Speech Production Qualities: A qualitative questionnaire
 問卷：比較語言生成質量
 Please fill out this questionnaire only if you are a Shanghainese speaker/listener
 如若您會聽/會說上海話，請繼續填寫這個問卷。
 In this folder, you will find 15 audio files, for each of which, please rate how well-spoken you think it is based on the following criteria:
 1. Comprehensibility: How well can you understand the meaning of the audio?
 2. Naturalness: How natural does the audio sound?
 3. Accuracy: How well does the audio match how a native speaker like you would pronounce it?
 4. Intelligibility: How much effort does it take to make sense of the audio?
 在這個文件夾裏，有15個音頻文件。請對每個音頻文件進行打分：
 1. 可以理解嗎？（您可以理解這個音頻的內容嗎？）
 2. 自然嗎？
 3. 發音準確嗎？（聽起來和上海話母語者的發音有多大區別？）
 4. 容易理解嗎？（如果可以理解，那麼聽起來很費勁嗎？）
 Please rate on a 1-to-5 scale: 1 (bad), 2 (poor), 3 (fair), 4 (good), 5 (excellent)
 請用1-5的數字來打分：1 (很差), 2 (差), 3 (還可以), 4 (好), 5 (很好)

	Comprehensibility/可以理解嗎？	Naturalness/自然嗎？	Accuracy/發音準確嗎？	Intelligibility/容易理解嗎？
我老實得很，昨日夜裡腳扭着。-1.wav				
我老實得很，昨日夜裡腳扭着。-2.wav				
我老實得很，昨日夜裡腳扭着。-3.wav				
上海是一座國際化大都市。-1.wav				
上海是一座國際化大都市。-2.wav				
上海是一座國際化大都市。-3.wav				
虹桥機場分為一號航站樓搭兩號航站樓。-1.wav				
虹桥機場分為一號航站樓搭兩號航站樓。-2.wav				
虹桥機場分為一號航站樓搭兩號航站樓。-3.wav				
很好，世界！-1.wav				
很好，世界！-2.wav				
很好，世界！-3.wav				
張秀英第二三個。-1.wav				
張秀英第二三個。-2.wav				
張秀英第二三個。-3.wav				

Thank you for your time!
 感謝您參與調查！

Figure 8: Questionnaire.