CSS10: A COLLECTION OF SINGLE SPEAKER SPEECH DATASETS FOR 10 LANGUAGES

1. INTRODUCTION

Recently there have been many neural TTS models.

- WaveNet, Tacotron, Char2Wav, DeepVoice, DCTTS, ...
- Internal vs. Public
- English vs. non-English
- Motivation: Public non-English datasets!

CONTRIBUTIONS

- Construction and release of datasets
- Validation / Evaluation

2. RELATED WORK

- En: LJ, WEB
- Ja: JSUT
- de: Pavoque

3. DATASETS

3.1. Selection of audiobooks

- LibriVox: 95 langs
- solo
- amount
- audio quality
- text availability
- de, el, es, fi, fr, hu, ja, nl, ru, zh

3. DATASETS

3.2. Audio processing

- Fragment into small audio clips
- Find split points automatically
- Audacity

3. DATASETS

3.3. Text processing

- Forced aligner such as Gentle
- Complicated
- Not correct
- English only

3. DATASETS

3.3.1. Text normalization

- Case retained
- Abbreviation expansion (Dr.-> Doctor)
- Arabic numbers are spelled out (2 -> two)

3. DATASETS

3.3.2. Phonetic transcription

- Latin, Cyrillic, Greek, Kana: phonetic
- Chinese: ideographic
- ja: MECAB + manual, romkan
- zh: Jieba + CC-CEDICT

EXAMPLE (ES)

19demarzo/19demarzo_0333.wav|Estos, lejos de amparar al que un día antes era su jefe, alborotaron el vecindario, y la misma turbamulta de la noche del 17 acudió con heroico entusiasmo a apoderarse de él.|Estos, lejos de amparar al que un día antes era su jefe, alborotaron el vecindario, y la misma turbamulta de la noche del diecisiete acudió con heroico entusiasmo a apoderarse de él.|11.69

EXAMPLE (JA)

meian_0000.wav| この前探った時は、途中に瘢痕の隆起があったので、ついそこが行きどまりだとばかり思って、ああ云ったんですが、| kono mae sagut ta toki wa 、tochu- ni hankon no ryu-ki ga at ta node 、tsui soko ga yukidomari da to bakari omot te 、a- yut ta n desu ga 、

EXAMPLE (ZH)

call_to_arms/call_to_arms_0001.wav|后来大半忘却了，但自己也并不以为可惜。所谓回忆者，虽说可以使人欢欣，有时也不免使人寂寞，|hòu lái dà bàn wàng què liào ， dàn zì jǐ yě bìng bù yǐ wéi kě xī 。suǒ wèi huí yì zhě ， suī shuō kě yǐ shǐ rén huān xīn ， yǒu shí yě bù miǎn shǐ rén jì mò ，

4. EXPERIMENTS

4.1. Models

- Tacotron
- DCTTS

4. EXPERIMENTS

4.2. Training

- 400k steps
- T: 10 days, D: 3 days

4. EXPERIMENTS

4.3. Evaluation

- 20 Test sentences from Tatoeba
- MOS from MTurk

4. EXPERIMENTS

4.4. Results

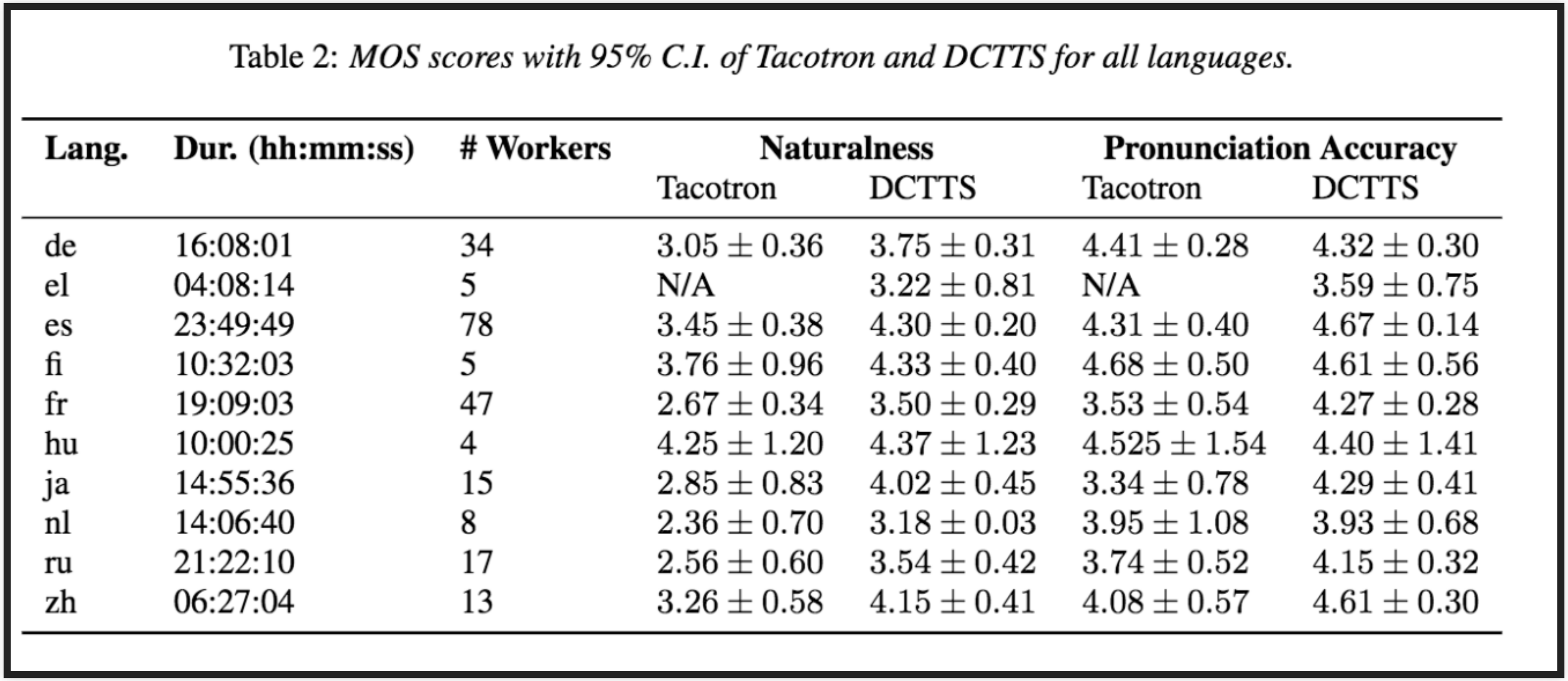
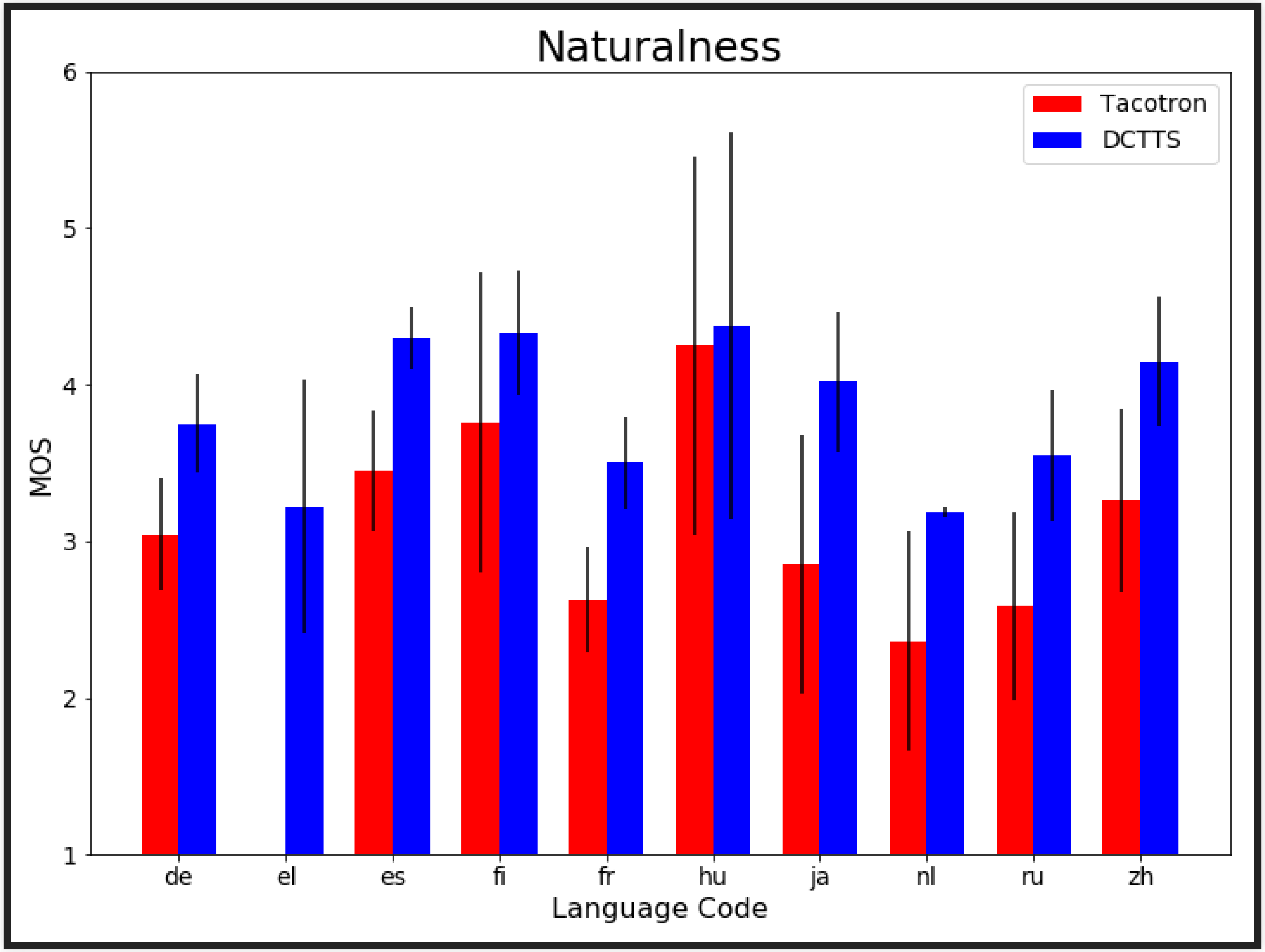
- 

Table 2: *MOS scores with 95% C.I. of Tacotron and DCTTS for all languages.*

Lang.	**Dur. (hh:mm:ss)**	**# Workers**	**Naturalness**		**Pronunciation Accuracy**	
			Tacotron	DCTTS	Tacotron	DCTTS
de	16:08:01	34	3.05 ± 0.36	3.75 ± 0.31	4.41 ± 0.28	4.32 ± 0.30
el	04:08:14	5	N/A	3.22 ± 0.81	N/A	3.59 ± 0.75
es	23:49:49	78	3.45 ± 0.38	4.30 ± 0.20	4.31 ± 0.40	4.67 ± 0.14
fi	10:32:03	5	3.76 ± 0.96	4.33 ± 0.40	4.68 ± 0.50	4.61 ± 0.56
fr	19:09:03	47	2.67 ± 0.34	3.50 ± 0.29	3.53 ± 0.54	4.27 ± 0.28
hu	10:00:25	4	4.25 ± 1.20	4.37 ± 1.23	4.525 ± 1.54	4.40 ± 1.41
ja	14:55:36	15	2.85 ± 0.83	4.02 ± 0.45	3.34 ± 0.78	4.29 ± 0.41
nl	14:06:40	8	2.36 ± 0.70	3.18 ± 0.03	3.95 ± 1.08	3.93 ± 0.68
ru	21:22:10	17	2.56 ± 0.60	3.54 ± 0.42	3.74 ± 0.52	4.15 ± 0.32
zh	06:27:04	13	3.26 ± 0.58	4.15 ± 0.41	4.08 ± 0.57	4.61 ± 0.30

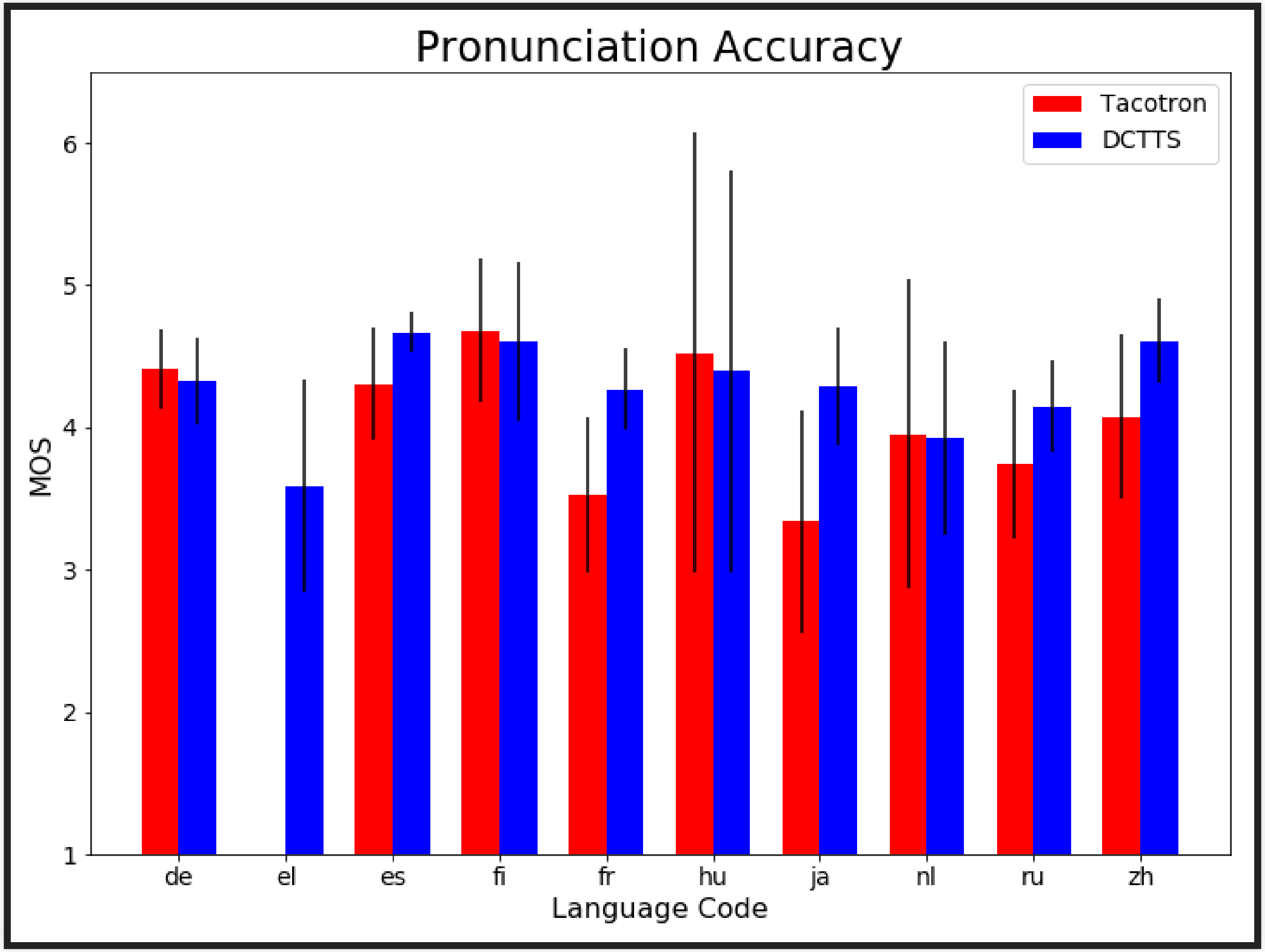
4. EXPERIMENTS

4.4. Results


Naturalness
Tacotron
DCTTS
MOS
6
5
4
3
2
1
de
el
es
fi
fr
hu
ja
nl
ru
zh
Language Code

4. EXPERIMENTS

4.4. Results


Pronunciation Accuracy
Tacotron
DCTTS
MOS
1
2
3
4
5
6
de
el
es
fi
fr
hu
ja
nl
ru
zh
Language Code

●

5. APPLICATIONS

- Cross-lingual TTS?
- Voice Conversion?
- multi-lingual speech recognition