Text-to-Speech for Under-Resourced Languages: Phoneme Mapping and Source Language Selection in Transfer Learning

Phat Do[1], Matt Coler[1], Jelske Dijkstra[2], Esther Klabbers[3]

[1] University of Groningen, Campus Fryslân
[2] Fryske Academy/Mercator Research Centre
[3] ReadSpeaker

SIGUL 2022 Workshop - June 24, 2022

1) Motivation & contributions

Neural text-to-speech (TTS):

+ High quality (naturalness & intelligibility)
- Large amounts of training data

  ➔ Issue for under-resourced languages (URLs)

➔ **Cross-lingual transfer learning:**

- Pre-train on source language (ample data)
- Fine-tune on target language (limited data)

1) Motivation & contributions

Challenges:

1. Input mismatch btw. source & target languages

➔ Phoneme mapping: (e.g., Chen et al. 2019, Wells & Richmond 2021)

- Complex & language-dependent

➔ Contribution **(1)**: proposed phoneme mapping method

- Simple but effective: rule-based using phonological features
- Language-independent: applicable to any language

Challenges:

2. Criterion for source language selection

- Convention in research: *language family*
  - Gutkin & Sproat (2017), Do et al. (2021) ➔ not effective

➔ Contribution **(2)**: proposed criterion for source lang. selection

  - Measures similarity btw. phoneme systems
  - Compare effectiveness with language family

2) Phoneme mapping

- **Database:** PHOIBLE (Moran & McCloy 2019)
  - Phonological inventories of 2,186 languages
  - Each phoneme:
    - **Unique** IPA symbol
    - **Unique** set of 37 binary phonological features

	tone	stress	syllabic	short	long	consonantal	sonorant	continuant	delayedRelease	approximant	tap	trill	nasal	lateral	labial	round	labiodental	coronal	anterior	distributed	strident	dorsal	high	low	front	back	tense	retractedTongueRoot	advancedTongueRoot	periodicGlottalSource	epilaryngealSource	spreadGlottis	constrictedGlottis	fortis	raisedLarynxEjective	loweredLarynxImplosive	click
m	0	-	-	-	-	+	+	-	0	-	-	-	+	-	+	-	-	-	0	0	0	-	0	0	0	0	0	0	0	+	-	-	-	-	-	-	-
n	0	-	-	-	-	+	+	-	0	-	-	-	+	-	-	0	0	+	+	-	-	-	0	0	0	0	0	0	0	+	-	-	-	-	-	-	-

2) Phoneme mapping

- **Rule:** for each phoneme (IPA symbol) in target language, if:
  - IN source language: use weight of that phoneme
  - NOT IN source language:
    - Map to phoneme with the **most similar** 37-feature set
      - Ties:
        - Compare cosine similarities (*) of phoneme frequencies of adjacent positions
    - Some diphthongs & long vowels: treat as unitary vowels

3) Phoneme similarity

- **Measure:** NLP: cosine similarity (cos_θ) to compare documents
  - Language A ➔ phoneme set P_A ➔ phoneme frequencies PF_A
  - Compare languages A & B with angular similarity (S_θ):

$$S_C(PF_A, PF_B) := \cos_\theta = \frac{PF_A \cdot PF_B}{\|PF_A\| \|PF_B\|}$$

$$S_\theta := 1 - \frac{2 \cdot \arccos(\cos_\theta)}{\pi}$$

➔ S_θ: Angular Similarity of Phoneme Frequencies (ASPF)

- $0 \leq ASPF \leq 1$

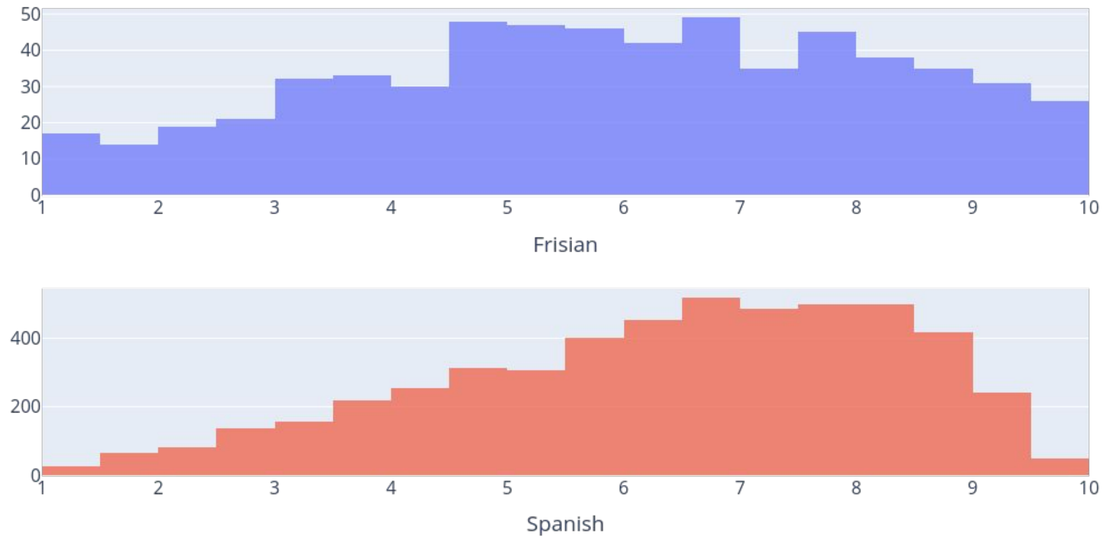
4) Experiment & results

- **Target language:**
  - Frisian ("Frysk") in Friesland province, north of the Netherlands
  - Data set:
    - Single-speaker, from a Frisian audiobook
    - Audio duration: 1 - 10 secs
    - Total duration: **30 minutes** (316 utterances)

4) Experiment & results

- **Source languages:**
  - Source data set: CSS10 (Park & Mulc 2019)
  - Selected: Dutch, Finnish, French, Japanese, Spanish
    - Balance: availability (audio duration) & language family
    - Duration: 1 - 10 secs
    - Total duration (each):

      ~ 9 hours



4) Experiment & results

- **Phonemization:**
  - Followed CMUDict (CMU 2014), except:
    - Used **IPA symbols** (from PHOIBLE)
    - Only included **primary stress**
  - Out-of-vocabulary words:
    - Grapheme-to-phoneme model using OpenNMT (Klein et al. 2017)

4) Experiment & results

- **Model architecture:**
  - Acoustic model: **FastSpeech 2** (Ren et al. 2020), open-source implementation by Chien et al. (2021)
  - Vocoder: **universal Hifi-GAN V1** (Kong et al. 2020)
- **Source language pre-training:**
  - **One** separate model for each source language
  - 100K parameter updates, batch size 16, Adam optimizer
  - 20 test sentences (CSS10) (phat-do.github.io/sigul22)

- **Target language fine-tuning:**
  - From **each** source language model: 2 scenarios
    - Without phoneme mapping (*separate*)
    - With phoneme mapping (*mapped*)

    ➔ Total: 10 fine-tuned models
  - Each: 100K parameter updates, batch size 4

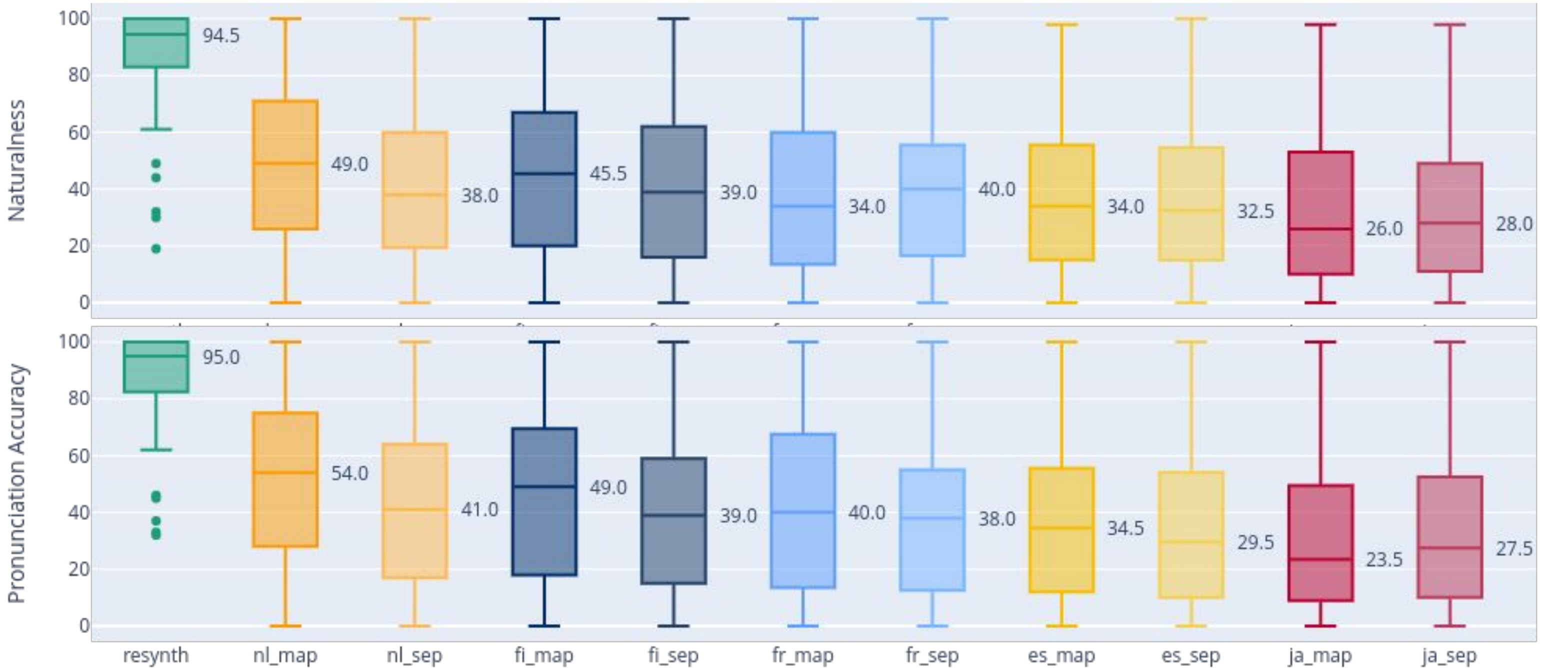
4) Experiment & results

- **Evaluation:** (stimuli available online)
  - 20 test sentences, divided into 5 sets (avg. **duration 5s**), each:
    - Contains all Frisian phonemes
    - Phoneme distribution close to Frisian data set
  - Online listening experiment (**MUSHRA**) for native speakers:
    - Each sentence: **12 stimuli** (10 models + truth + resynth)
    - Rate naturalness & pronunciation accuracy (0-100)
    - Answers from 46 participants (n = 2024)

4) Experiment & results

- Results:

Source language	Naturalness $(M_{map} - M_{sep})$	Accuracy $(M_{map} - M_{sep})$
nl (Dutch)	**11 (p <.001)**	**13 (p <.001)**
fi (Finnish)	**6.5 (p = .003)**	**10 (p <.001)**
fr (French)	-6 (p = .82)	**2 (p = .02)**
es (Spanish)	1.5 (p = .17)	5 (p = .21)
ja (Japanese)	-2 (p = .56)	-4 (p = .11)



- <u>Phoneme mapping:</u> Increased naturalness by **2.42** (± 0.85) (p = *.004*)

  Increased pron. accuracy by **3.79** (± 0.88) (p < *.001*)

➔ Effective, but depended on source language

4) Experiment & results

- **Results:**
  - **<u>Language family:</u>** (compared to Frisian)
    - **Dutch, French, Spanish** (Indo-European): same family
    - **Finnish** (Uralic), **Japanese** (Japonic): different family
    - Did **NOT** have a significant effect (p = .56 and p = .50)
  - **<u>ASPF:</u>** sentence-level, for every 10-percentage-point increase:
    - Increased naturalness by **2.93** (± 0.36) ($p < .001$)
    - Increased pron. accuracy by **3.66** (± 0.37) ($p < .001$)

5) Conclusions & future work

- **Conclusions:** 2 contributions
  - Phoneme mapping improved quality (depended on source language)
  - Source language selection: ASPF more effective than lang. family

  ➔ **Applicable for TTS for URLs** (language-independent)
- **Future work:**
  - Verify with a wider range of languages (families)
  - Try phoneme mapping without (target language) lexicon

Thank you for listening!