Low-Resource Text-to-Speech Synthesis

Through Language Agnostic Meta Learning and Articulatory Features

Florian Lux & Ngoc Thang Vu

1: Introduction

Neural Text-to-Speech has made great advances, however it still requires a lot of data, which is not feasible for many languages.

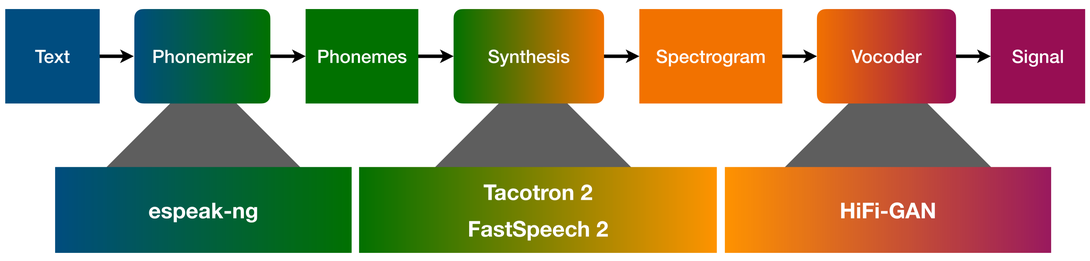
Mismatch in the Input Space

We want to **share knowledge** between higher and lower resourced languages. But languages have different phoneme sets.
→ **Meaningful featurevectors** allow to zero-shot infer phonemes! [1]

Data Need

A shared input space alone allows finetuning to a new language with less data. But it's still infeasible for many languages.
→ **Model Agnostic Meta Learning** [2] (MAML) reduces data need to learn unseen tasks. We formulate our training procedure such that different languages are the tasks in a MAML like procedure. This allows us to learn new languages with just 30 minutes of speech in an unseen language.

Basic Architecture



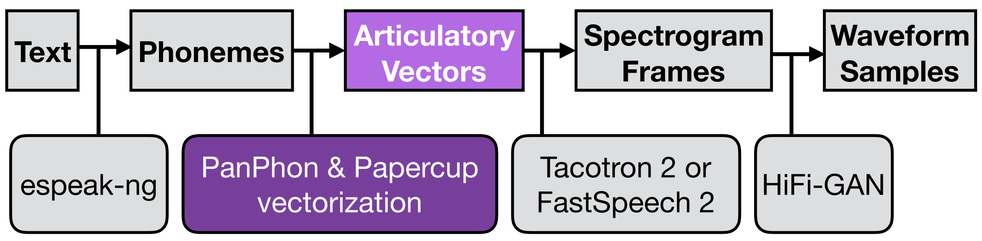
We use **FastSpeech 2** [3], **Tacotron 2** [4] and **HiFi-GAN** [5] based on ESPnet [6] implementations. We use single speaker modeling only to get the most general speech production knowledge possible in the pretraining.

Code and Models are Public

IMS Speech Synthesis Toolkit
Toucan

2: Proposed Approach

Articulatory Features

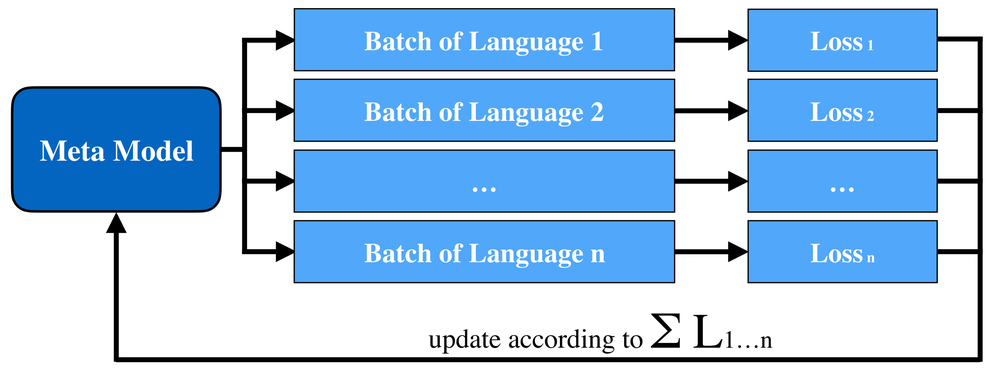


Contains features such as the type of the symbol (vowel, consonant, other), the position of articulation (e.g. palatal, alveolar, glottal), whether it's voiced or unvoiced, if there is friction noise, the position of the tongue, tongue tenseness, whether the mouth is rounded and some more.

ʁ	symbol_type	= phoneme
	vowel_consonant	= consonant
	voicedness	= voiced
	consonant_place	= uvular
	consonant_manner	= fricative

i	symbol_type	= phoneme
	vowel_consonant	= vowel
	voicedness	= voiced
	vowel_frontness	= central
	vowel_openness	= close
	vowel_roundedness	= unrounded

Language Agnostic Meta Learning

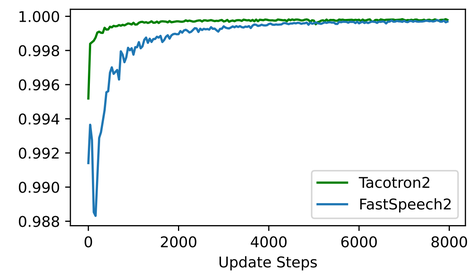


In terms of MAML: Every language is one task and the inner loop is run for only one step. This stabilizes training and generalizes to unseen languages.

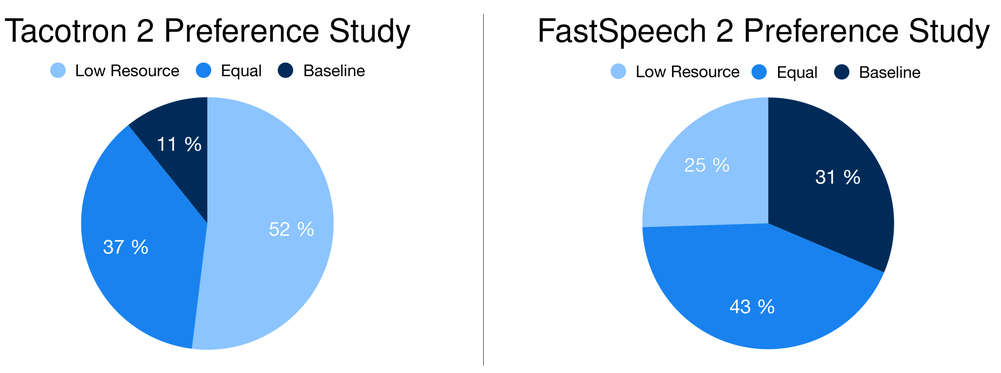
3: Evaluation

How fast does it adapt to new speakers?

Tacotron 2 reaches excellent speaker similarity after just 1000 steps, FastSpeech 2 takes roughly 5000. In both cases, speaker similarity is excellent before overfitting problems become apparent.

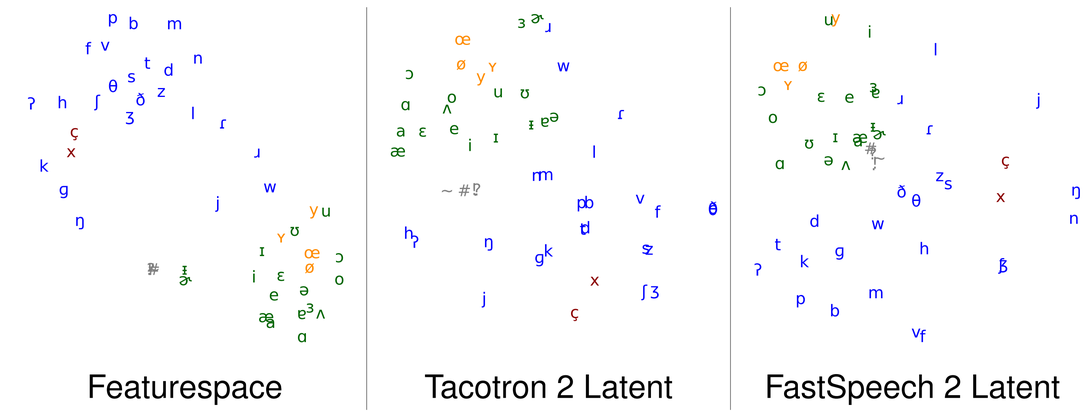


How is the quality compared to high-resource?



For both Tacotron 2 and FastSpeech 2, more than half of the 102 ratings by the 34 raters say that the low-resource model trained on 30 minutes is better or equal to one trained on 29 hours without pretraining.

Can it handle unseen phonemes?



Featurespace | Tacotron 2 Latent | FastSpeech 2 Latent

Consonants are blue, vowels are green. Unseen vowels are orange and unseen consonants are red. Unseen phonemes get placed in meaningful positions in the latent space.

Take Home: Using articulatory phoneme descriptions as input and training on multiple languages simultaneously without conditioning leads to a language agnostic TTS model initialization.

References

[1] M. Staib, T. H. Teh, A. Torresquintero, D. S. R. Mohan, L. Foglianti, R. Lenain, and J. Gao, "Phonological Features for 0-Shot Multilingual Speech Synthesis," *Interspeech*, 2020.

[2] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *ICML*, 2017.

[3] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech 2: Fast and High-Quality End-to-End Text to Speech," in *ICLR*, 2020.

[4] Y. Wang, R. Skerry-Ryan, D. Stanton *et al.*, "Tacotron: Towards End-to-End Speech Synthesis," *Interspeech*, 2017.

[5] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[6] T. Hayashi, R. Yamamoto, K. Inoue, T. Yoshimura *et al.*, "ESPnet-TTS: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit," in *ICASSP*, 2020.