# XPhoneBERT: A Pre-trained Multilingual Model for Phoneme Representations for Text-to-Speech

**Dat Quoc Nguyen**
Head of NLP department, VinAI
https://datquocnguyen.github.io

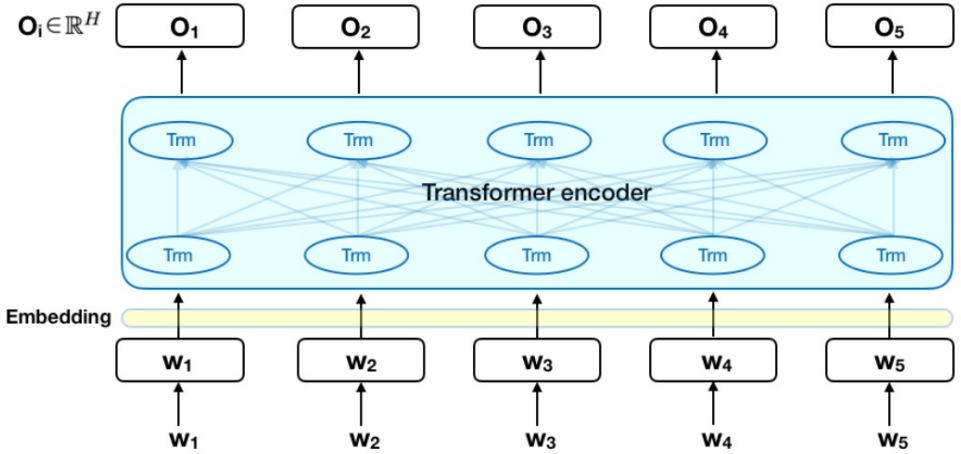
Joint work with
**Linh The Nguyen & Thinh Pham**

21/04/2023

# Motivation



- Language model BERT [1]—Bidirectional Encoder Representations from Transformers [2]
  - BERT and its variants, pre-trained on large-scale corpora, help improve the state-of-the-art performances of various NLP research and application tasks
  - Represent words by embedding vectors which encode the contexts where the words appear, i.e. contextualized word embeddings




https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270

# Motivation

- Several Text-to-Speech (TTS) works incorporate contextualized word embeddings generated by the pre-trained BERT into their standard encoder [3, 4, 5]
  - An input phoneme sequence is fed into the standard encoder to produce phoneme representations
  - The corresponding input raw text is fed into BERT to obtain contextualized word embeddings
  - Concatenate the representations of the input phonemes with the corresponding BERT-based word embeddings to construct the input vectors of the TTS decoder (for decoding)

Phoneme sequence: ˈeɪ ˌmałtiˈłɪŋwəł ˈmɑdəł

Raw text: a multilingual model

# Motivation



- Several Text-to-Speech (TTS) works incorporate contextualized word embeddings generated by the pre-trained BERT into their standard encoder [3, 4, 5]
  - Provide extra contextual information for phoneme representation

- Pre-trained models for phoneme representations, including PnG BERT [6], Mixed-Phoneme BERT [7] and Phoneme-level BERT [8], help improve advanced TTS systems
  - PnG BERT takes both phonemes and graphemes (i.e. subword tokens) as the input
  - Mixed-Phoneme BERT takes both phonemes and sup-phoneme tokens as the input
  - Phoneme-level BERT only takes phonemes as the input

# Motivation



- Pre-trained models for phoneme representations, including PnG BERT [6], Mixed-Phoneme BERT [7] and Phoneme-level BERT [8], help improve advanced TTS systems
  - *Limited to English only*
  - *Not publicly available*

➔ **A need of developing pre-trained models for phoneme representations in languages other than English**

# Motivation



- **Our contributions:**
  - Present the first large-scale pre-trained multilingual model for phoneme representations, named XPhoneBERT
  - XPhoneBERT helps significantly improve the performance of the strong TTS baseline VITS [9]
  - We will publicly release XPhoneBERT

# Pre-training XPhoneBERT



- Multilingual pre-training data
  - *Raw dataset collection*: Collect texts for the 90+ languages and locales supported by CharsiuG2P [10], from multilingual datasets **`wiki40b`** and **`wikipedia`**
    - Perform word and sentence segmentation as well as duplicate removal and text normalization
  - *Text-to-phoneme conversion*
  - *Phoneme segmentation*

# Pre-training XPhoneBERT



- Multilingual pre-training data
  - *Raw dataset collection*
  - *Text-to-phoneme conversion*: Employ the grapheme-to-phoneme conversion toolkit CharsiuG2P to convert sentences into their phonemic description
  - *Phoneme segmentation*: Employ the **`segments`** toolkit for phoneme segmentation for a better map between phonemes and speech
  - Demonstration example:

a multilingual model ➔ CharsiuG2P ➔ ˈeɪ ˌmɑltiˈlɪŋwəl ˈmɑdəl ➔ **`segments`** ➔ ˈe ɪ _ ˌm ɑ l t i ˈl ɪ ŋ w ə l _ ˈm ɑ d ə l

# Pre-training XPhoneBERT

- Multilingual pre-training data
  - *A pre-training corpus of **330M** phoneme-level sentences across **94** languages and locales*

| LCode | #s (K) | LCode | #s (K) | LCode | #s (K) |
|---|---|---|---|---|---|
| ady | 2 | glg | 3793 | ron | 1816 |
| afr | 1793 | grc | 947 | rus | 15923 |
| amh | 73 | gre | 947 | san | 114 |
| ara | 2820 | grn | 60 | slo | 1143 |
| arg | 383 | guj | 211 | slv | 1167 |
| arm-e | 2989 | hbs-cyrl | 2007 | sme | 27 |
| arm-w | 175 | hbs-latn | 2007 | snd | 215 |
| aze | 3139 | hin | 287 | spa | 3936 |
| bak | 1272 | hun | 4372 | spa-latin | 3936 |
| bel | 2750 | ice | 776 | spa-me | 3936 |
| ben | 1785 | ido | 224 | sqi | 1373 |
| bos | 1464 | ina | 100 | srp | 2449 |
| bul | 1919 | ind | 2196 | swa | 537 |
| bur | 393 | ita | 12335 | swe | 5226 |
| cat | 4017 | jam | 8 | tam | 2289 |
| cze | 4542 | jpn | 12197 | tat | 984 |
| dan | 1714 | kaz | 1850 | tgl | 628 |
| dut | 7683 | khm | 93 | tha | 567 |
| egy | 3093 | kor | 2384 | tts | 567 |
| eng-uk | 33515 | kur | 335 | tuk | 105 |
| eng-us | 33515 | lat-clas | 597 | tur | 2148 |
| epo | 4333 | lat-eccl | 597 | ukr | 6967 |
| est | 1558 | lit | 1087 | vie-c | 2519 |
| eus | 3429 | ltz | 817 | vie-n | 2519 |
| fas | 1957 | mac | 2597 | vie-s | 2519 |
| fin | 4100 | min | 377 | wel-nw | 714 |
| fra | 11255 | mlt | 180 | wel-sw | 714 |
| fra-qu | 11255 | ori | 158 | yue | 908 |
| geo | 1211 | pap | 27 | zho-s | 6934 |
| ger | 33845 | pol | 7045 | zho-t | 6955 |
| gla | 121 | por-bz | 3437 | _ | _ |
| gle | 488 | por-po | 3437 | _ | _ |

# Pre-training XPhoneBERT



- Pre-training approach
  - Employ the BERT-base model architecture
  - Use a whitespace tokenizer, thus resulting in a vocabulary of 1960 phoneme types, and a model of 87.6M parameters
  - Use the masked language modeling objective and follow the RoBERTa pre-training approach [11]
    - Use a dynamic masking strategy and without the next sentence prediction objective
  - Train for 20 epochs in about 18 days

# Downstream TTS evaluation



- Evaluate the effectiveness of XPhoneBERT on the TTS task for English and Vietnamese
  - English: Training, validation and test sets of 12,500, 100 and 500 audio clips
  - Vietnamese: Training, validation and test sets of 12,000, 100 and 200 clips

- Employ the strong TTS model VITS [9] as the baseline
  - VITS is an end-to-end model that contains a Transformer encoder to encode the input phoneme sequence
- Extend VITS with XPhoneBERT by replacing the VITS's encoder with XPhoneBERT

# Downstream TTS evaluation



- Two training settings using **100%** and **5%** of the TTS training data

- Evaluation metrics
  - Subjective evaluation: For each language, following [9], we randomly select 50 ground truth test audios and their text transcription to measure the Mean Opinion Score (MOS)
  - Objective evaluation: We compute two metrics of the mel-cesptrum distance (MCD) and the F0 root mean square error (RMSE; cent)

# Downstream TTS evaluation

- XPhoneBERT helps improve the performance of VITS on all three evaluation metrics for both English and Vietnamese in both experimental settings
  - 100% of the training set for training: MOS significantly increases from 4.00 to 4.14 (+0.14) for English and from 3.74 to 3.89 (+0.15) for Vietnamese
  - 5% of the training set for training: MOS increases from 2.88 to 3.22 (+0.34) for English and especially from 1.59 to 3.35 (+1.76) for Vietnamese

Obtained results on the English test set

| | Model | MOS (↑) | MCD (↓) | $\text{RMSE}_{F_0}$ (↓) |
|---|---|---|---|---|
| | Ground truth | 4.39 ± 0.08 | 0.00 | 0.00 |
| 100% | Baseline VITS | 4.00 ± 0.08 | 7.04 | 377 |
| 100% | VITS w/ XPB | **4.14** ± 0.07 | **6.63** | **348** |
| 5% | Baseline VITS | 2.88 ± 0.11 | 7.40 | 407 |
| 5% | VITS w/ XPB | **3.22** ± 0.11 | **7.15** | **383** |

Obtained results on the Vietnamese test set

| | Model | MOS (↑) | MCD (↓) | $\text{RMSE}_{F_0}$ (↓) |
|---|---|---|---|---|
| | Ground truth | 4.26 ± 0.06 | 0.00 | 0.00 |
| 100% | Baseline VITS | 3.74 ± 0.08 | 5.41 | 249 |
| 100% | VITS w/ XPB | **3.89** ± 0.08 | **5.12** | **234** |
| 5% | Baseline VITS | 1.59 ± 0.05 | 6.20 | 291 |
| 5% | VITS w/ XPB | **3.35** ± 0.10 | **5.39** | **248** |

"XPB" abbreviates XPhoneBERT

# Downstream TTS evaluation

- MOS is not “always” correlated with MCD and RMSE
  - For Vietnamese, VITS under the first setting (100%) obtains higher MOS but slightly poorer MCD and RMSE thanVITS extended with XPhoneBERT under the second setting (5%)

- XPhoneBERT helps synthesize fairly high-quality speech with limited training data (5%)

Obtained results on the English test set

| | Model | MOS (↑) | MCD (↓) | $\text{RMSE}_{F_0}$ (↓) |
|---|---|---|---|---|
| | Ground truth | 4.39 ± 0.08 | 0.00 | 0.00 |
| 100% | Baseline VITS | 4.00 ± 0.08 | 7.04 | 377 |
| 100% | VITS w/ XPB | **4.14** ± 0.07 | **6.63** | **348** |
| 5% | Baseline VITS | 2.88 ± 0.11 | 7.40 | 407 |
| 5% | VITS w/ XPB | **3.22** ± 0.11 | **7.15** | **383** |

Obtained results on the Vietnamese test set

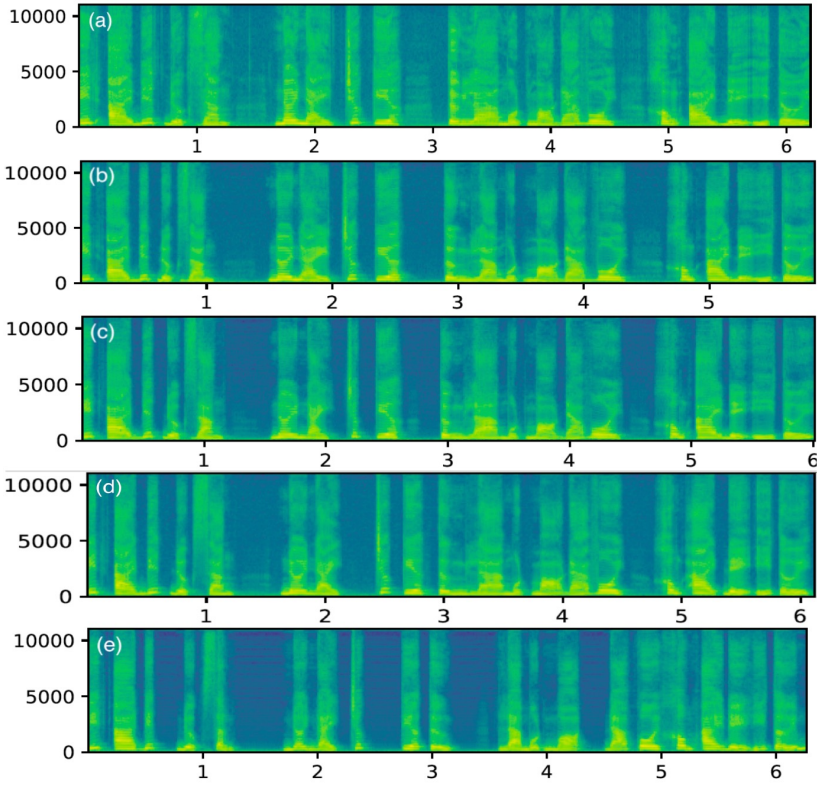
| | Model | MOS (↑) | MCD (↓) | $\text{RMSE}_{F_0}$ (↓) |
|---|---|---|---|---|
| | Ground truth | 4.26 ± 0.06 | 0.00 | 0.00 |
| 100% | Baseline VITS | 3.74 ± 0.08 | 5.41 | 249 |
| 100% | VITS w/ XPB | **3.89** ± 0.08 | **5.12** | **234** |
| 5% | Baseline VITS | 1.59 ± 0.05 | 6.20 | 291 |
| 5% | VITS w/ XPB | **3.35** ± 0.10 | **5.39** | **248** |

“XPB” abbreviates XPhoneBERT

# Downstream TTS evaluation



- Spectrograms visualization by different models, illustrating that XPhoneBERT helps improve the spectral details of the baseline VITS's output
  - "Ít ai biết được rằng nơi này trước kia từng là một mỏ đá vôi không ai để ý tới" (Little is known that this place was once a limestone quarry that no one paid any attention to)
  - (a): Ground truth
  - (b): VITS with XPhoneBERT (100%)
  - (c): VITS with XPhoneBERT (5%)
  - (d): Original VITS (100%)
  - (e): Original VITS (5%)



# Takeaways



- XPhoneBERT is the first large-scale multilingual language model pre-trained for phoneme representations
- Using XPhoneBERT as an input phoneme encoder improves the quality of the speech synthesized by a strong neural TTS baseline
  - XPhoneBERT also helps produce fairly high-quality speech when the training data is limited
- We will publicly release XPhoneBERT, which can be used with popular libraries **`fairseq`** and **`transformers`**

# Thank you!

@VinAI

https://www.vinai.io/

# References



1. BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding
2. Attention is All you Need
3. Pre-Trained Text Embeddings for Enhanced Text-to-Speech Synthesis
4. Improving the Prosody of RNN-based English Text-To-Speech Synthesis by Incorporating a BERT model
5. Improving Prosody Modelling with Cross-Utterance BERT Embeddings for End-to-end Speech Synthesis
6. PnG BERT: Augmented BERT on Phonemes and Graphemes for Neural TTS
7. Mixed-Phoneme BERT: Improving BERT with Mixed Phoneme and Sup-Phoneme Representations for Text to Speech
8. Phoneme-Level BERT for Enhanced Prosody of Text-to-Speech with Grapheme Predictions
9. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech
10. ByT5 model for massively multilingual grapheme-to-phoneme conversion
11. RoBERTa: A Robustly Optimized BERT Pretraining Approach