Unsupervised phoneme segmentation using a transformer autoencoder

Hahn Koo (hahn.koo@sjsu.edu)
Linguistics and Language Development, San José State University

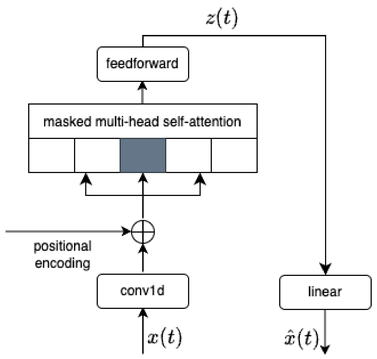
BayPhon2025
05/10/2025

Summary

- Unsupervised (aka blind) phoneme segmentation uses only the audio signal to find phoneme boundaries. It could be an efficient alternative to forced-alignment which requires transcriptions and/or pre-trained acoustic models.
- I propose that one approach is to train a deep learning autoencoder composed of a convolution module, a masked multi-head self-attention module, and a feedforward module to reconstruct the log mel spectrogram of a given audio recording and find peaks in how its latent representation changes across frames.
- Segmentation performance of the proposed approach was not far behind forced-alignment when evaluated on the TIMIT corpus.

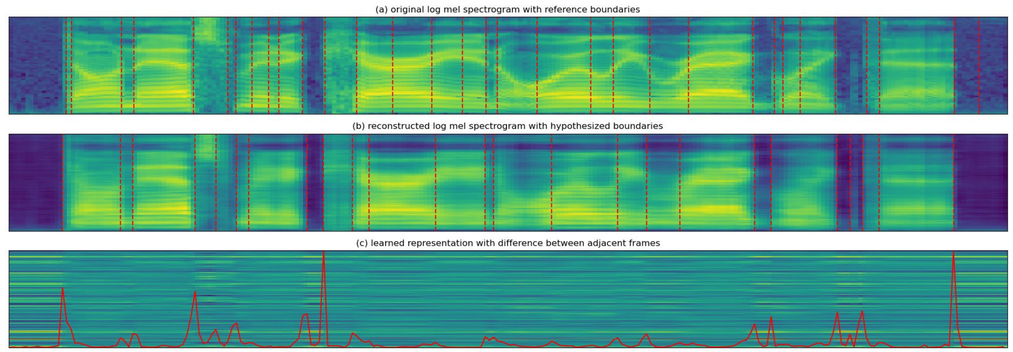
Model

The autoencoder has a similar structure to a transformer encoder (Vaswani et al., 2017):



- $x(t)$: original log mel-spectrum (80 frequency bands) for frame t (10 ms shift, 25 ms long)
- $z(t)$: latent representation of $x(t)$
- $\hat{x}(t)$: reconstructed log mel-spectrum
- conv1d: 80 input channels, 128 output channels, kernel size of 1, followed by ReLU and batch normalization
- positional encoding from sinusoidal functions
- masked multi-head self-attention: 16 attention heads with the current position masked
- feedforward: 128 output & hidden units with ReLU
- linear: 80 output units

For each audio recording, the autoencoder is initialized afresh and trained to minimize reconstruction loss, mean square error between $x(t)$ and $\hat{x}(t)$, using the Adam optimizer. After training, t is considered a phoneme boundary if $\|z(t) - z(t+1)\|_2$ is a local maximum:



Evaluation

Segmentation performance was evaluated on the standard test set of the TIMIT corpus in terms of precision, recall, F1-score, over segmentation (OS) rate, and R-value.

- TIMIT test set: 1,680 recordings, 62,465 manually identified phoneme boundaries
- Hit: a hypothesized boundary within 20 ms of a reference boundary
- OS rate: ratio of number of hypothesized boundaries to number of reference boundaries
- R-value: a measure similar to F-score but penalizes OS more (Räsänen et al., 2009)

Performance was compared with two other models: a baseline model and the Montreal Forced Aligner (MFA; McAuliffe et al., 2017).

- Baseline: Similar to Aversano et al. (2001), compare log mel spectra (80 frequency bands) between two left frames and two right frames for each frame (10 ms shift, 25 ms long) and draw a boundary if their cosine distance peaks
- MFA: Time-align a recording and its phonetic transcription using a pre-trained GMM-HMM acoustic model along with speaker adaptation
  - Phonetic transcriptions generated by applying a pronunciation lexicon to word-level transcriptions provided in TIMIT
  - Pronunciation lexicon and acoustic model from the MFA website (`english_us_arpa`)
  - 168 speakers with 10 recordings each for speaker adaptation

	Prec.	Recall	F-score	OS (%)	R-value
Baseline	0.599	0.824	0.694	37.490	0.598
MFA	0.796	0.722	0.757	-9.284	0.788
Proposed	0.722	0.722	0.722	-0.053	0.763

Error analysis revealed that boundaries involving silence and liquids can be difficult:

- Top 10 insertion errors involved five phones related to silence (`h#`, `q`, `kcl`, `tcl`, `pau`), two fricatives (`s`, `f`), and three vowels (`iy`, `ae`, `ay`)
- Top 10 deletion errors involved four phone pairs for stop closure and release (`bcl-b`, `dcl-d`, `kcl-k`, `tcl-t`), four phone pairs including a liquid (`r-iy`, `l-iy`, `g-r`, `eh-r`), one vowel-nasal pair (`ix-n`), and one pair specifying silence at the utterance beginning.

Future research

Performance may be improved by

- Applying silence detection
- Training on additional unlabeled recordings: e.g. TIMIT train set or LibriSpeech as in Kreuk et al. (2020)
- Utilizing latent representations from pre-trained models like Meta's Wav2vec2 (Baevski et al., 2020) or OpenAI's Whisper (Radford et al., 2023)

More evaluation necessary:

- Performance on other English data (e.g. Buckeye corpus) as well as other languages
- Comparison with other segmentation algorithms: e.g. CTC-based forced-alignment (Huang et al., 2024) or supervised phoneme segmentation (Franke et al., 2016)
- Practical use of phoneme segmentation in other domains: e.g. for disfluency detection such as phoneme duration for prolongations, similarity among neighboring phonemes for repetitions

References

- Aversano, G., Esposito, A., & Marinaro, M. (2001). A new text-independent method for phoneme segmentation. In *Proceedings of MWSCAS 2001* (Vol. 2, pp. 516-519).
- Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). Wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, *33*, 12449-12460.
- Franke, J., Mueller, M., Hamlaoui, F., Stueker, S., & Waibel, A. (2016). Phoneme boundary detection using deep bidirectional lstms. In *Speech Communication; 12. ITG Symposium* (pp. 1-5). VDE.
- Huang, R., Zhang, X., Ni, Z., Sun, L., Hira, M., Hwang, J., ... & Khudanpur, S. (2024). Less peaky and more accurate CTC forced alignment by label priors. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 11831-11835).
- Kreuk, F., Keshet, J., & Adi, Y. (2020). Self-supervised contrastive learning for unsupervised phoneme segmentation. In *Proceedings of INTERSPEECH 2020* (pp. 3700-3704).
- McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal forced aligner: Trainable text-speech alignment using Kaldi. In *Proceedings of INTERSPEECH 2017* (pp. 498-502).
- Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning* (pp. 28492-28518).
- Räsänen, O. J., Laine, U. K., & Altosaar, T. (2009). An improved speech segmentation quality measure: the R-value. In *Proceedings of INTERSPEECH 2009* (pp. 1851-1854).
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 1-11.

Supplementary materials