Attention-LSTM Autoencoder for Phonotactics Learning from Raw Audio Input

Frank Lihui Tan and Youngah Do

The University of Hong Kong

Research indicates that infants acquire phonemic awareness by the age of 6 to 8 months and begin to develop phonotactic knowledge between 8 to 10 months. By this age, they exhibit statistical learning capabilities, as they prefer sequences with higher transitional probabilities over those with lower probabilities, a phenomenon observed in both natural and artificial language learning contexts (Pelucchi et al., 2009; Saffran et al., 1996). However, it is still unclear how these capabilities are present during the early stages of phonological acquisition. This study utilizes a raw audio corpus to investigate the ability of a neural network model to acquire phonotactic knowledge. Mirroring the initial stages of infant language acquisition, the model is designed without prior knowledge of phonemes or phonotactic rules. Instead, it solely relies on raw audio sequences extracted directly from the corpus as input. Our case study investigates the aspiration alternation in English voiceless stop consonants occurring between the initial position and following the sibilant fricative /s/ (Iverson & Salmons, 1995).

We utilized a subset of the LibriSpeech corpus (Panayatov et al., 2015), including recordings from 251 speakers and totaling approximately 100 hours of audio data. We selected word-initial voiceless stops (/p/, /t/, /k/) and /s/-stop sequences. We randomly allocated 20% of the data as a validation set, with the remainder as the training set. The raw audio data underwent a transformation into Mel-spectrograms, with 25ms sample window length, 10ms shift length and 64 filters. The autoencoder compresses the input Mel-spectrogram sequence into a hidden representation of the same length, subsequently decoding it autoregressively with cross-attention (Vaswani et al., 2017). This design aims to replicate the sequential processing and memory inherent in language learners. Ten distinct models were trained, each initialized with different random seeds. Models were then evaluated on an evaluation set, from which we collected the reconstructed output, hidden representations, and attention matrices for each epoch. We analyzed the attention matrices for /s/+stop sequences and calculated the foreign attention scores for each frame to quantify the model's focus on segments other than the one to which the frame belongs.

We found increased foreign attention at the boundary position between /s/ and stop sounds, indicating that the model exhibits a sensitivity to the points where contrasts occur. Notably, except for the initial epochs (< 10) under unbalanced training conditions, all subsequent epochs exhibited significantly higher foreign attention scores on the stop side compared to the /s/ side (Figure 1). This pattern suggests that based on the training on raw audio corpus, the model has adapted to allocate more attention to the /s/ segment when reconstructing the following plosive, presumably to generate the de-aspirated allophone of the plosive. We also assessed the model's ability to differentiate in the hidden representation space between stops that follow an /s/ and those that do not. The model demonstrated a notable capacity to distinguish between these two categories, achieving an average silhouette score of approximately 0.3. This contrast with the model's performance distinguishing POA among stop categories themselves, where it showed almost no ability to differentiate (Figure 2). Our study demonstrates that based on raw audio data corpus through unsupervised training, an autoencoder model can implicitly learn phonotactic knowledge, mirroring early stages of language acquisition.

References

Beguš, G. (2020). Generative Adversarial Phonology: Modeling Unsupervised Phonetic and Phonological Learning With Neural Networks. *Frontiers in Artificial Intelligence*, *3*, 44. https://doi.org/10.3389/frai.2020.00044

Iverson, G. K., & Salmons, J. C. (1995). Aspiration and laryngeal representation in Germanic. *Phonology*, *12*(3), 369–396. https://doi.org/10.1017/S0952675700002566

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5206–5210. https://doi.org/10.1109/ICASSP.2015.7178964

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009). Statistical Learning in a Natural Language by 8-Month-Old Infants. *Child Development*, *80*(3), 674–685. https://doi.org/10.1111/j.1467-8624.2009.01290.x

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical Learning by 8-Month-Old Infants. *Science*, *274*(5294), 1926–1928. https://doi.org/10.1126/science.274.5294.1926

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need*. https://doi.org/10.48550/ARXIV.1706.03762

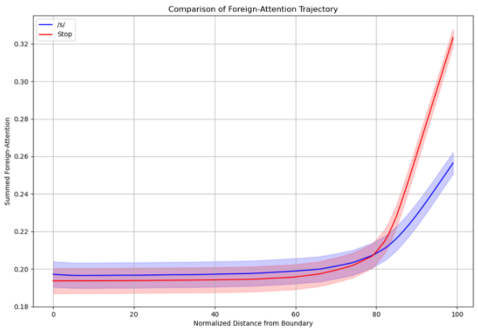

Figure 1 Foreign-attention trajectory of model trained on natural set at epoch 99.

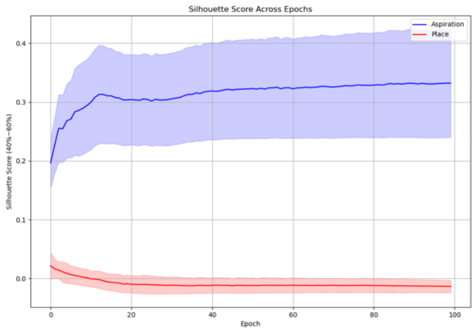

Figure 2 Silhouette score across epochs between aspiration contrasts (stops following /s/ vs at word-initial position) and POA contrasts.