MÔ HÌNH TỔNG HỢP NHANH GIỌNG NÓI TIẾNG VIỆT VỚI SPEEDYSPEECH ĐƯỢC CHUẨN HÓA

***Nguyễn Minh Trí*[1,2], *Đỗ Trí Nhân*[1,2], *Cao Xuân Nam*[2]**
[1]Chương trình tiên tiến,
[2]Khoa Công nghệ Thông tin,
Trường Đại học Khoa học Tự Nhiên, ĐHQG-HCM
{nmtri17, dtnhan}@apcs.vn, cxnam@fit.hcmus.edu.vn

Tóm tắt

Các mô hình tổng hợp giọng nói end-to-end ngày nay đã cho thấy nhiều kết quả tốt hơn so với những phương pháp truyền thống về độ thông minh và tự nhiên. Tuy nhiên những mô hình này cần nhiều dữ liệu và thời gian cho việc huấn luyện, đồng thời quá trình tổng hợp tốn kém và cần nhiều tài nguyên GPU. Với sự xuất hiện của SpeedySpeech, thời gian tổng hợp giọng nói được rút ngắn và quá trình tổng hợp có khả năng chạy được trên CPU. Tác giả ứng dụng mô hình này cho tiếng Việt bằng việc chuẩn hóa các kí tự đầu vào, thay bằng các âm vị phù hợp và kết quả cho thấy Speedyspeech có khả năng tổng hợp tiệm cận mô hình Tacotron 2 với thời gian huấn luyện cải thiện đáng kể. Việc huấn luyện SpeedySpeech trên Colab chỉ tốn 19 tiếng so với 240 giờ của Tacoton2 và độ đo Mel cepstral distortion (MCD) thu được đạt 990,69 trên 200 mẫu tập tin âm thanh.

Từ khóa: SpeedySpeech, Tổng hợp tiếng nói tiếng Việt, MelGan

SPEEDYSPEECH MODEL WITH NORMALIZATION FOR FASTER VIETNAMESE SPEECH SYNTHESIS

***Nguyen Minh Tri*[1,2], *Do Tri Nhan*[1,2], *Cao Xuan Nam*[2]**
[1]Advanced Program in Computer Science
[2]Faculty of Information Technology, University of Science, VNU-HCM
{nmtri17, dtnhan}@apcs.vn, cxnam@fit.hcmus.edu.vn

Abstract

End to end speech synthesis models have shown better results than traditional methods in terms of intelligence and spontaneity in recent years. However, these network models require a lot of data and training time, and the inference process consumes a lot of GPU resources. With the innovative experiments of SpeedySpeech, the speech synthesis time is shortened and the inference process has ability to run in real-time on the CPU. Authors apply this model for Vietnamese by standardizing the input, replace it with suitable Vietnamese phonemes, improve the embedding layer and the results showed that SpeedySpeech ‘s performance is asymptotic to the Tacotron2 model with significantly shorter training time. Training SpeedySpeech on Colab only takes 19 hours compared to 240 hours of training time on Tacotron2 and the Mel cepstral distortion (MCD) measurement is 990,69 when evaluated based on 200 sample audio tests.

Key words: SpeedySpeech, Vietnamese Speech Synthesis, MelGan