Đánh giá mô hình bằng thang đo MOS

Để đánh giá được chất lượng mô hình tổng hợp giọng nói của chúng tôi bằng Tacotron2 và WaveGlow, chúng tôi tiến hành chọn ra 25 đoạn text để thử nghiệm. Chúng tôi sử dụng điểm MOS trung bình của 2 người nghe để đưa ra được mức độ chất lượng dựa trên cảm quan bằng tai người.

Điểm ý kiến trung bình (Mean Opinion Score / MOS) là thước đo được dùng để kiểm tra chất lượng trải nghiệm của mô hình tổng thể, rất phổ biến để đánh giá chất lượng video, âm thanh, thiết bị nghe nhìn. Có thể hiểu đây là giá trị trung bình cộng dựa trên tất cả các ý kiến của một chủ thể trên một thang đo xác định trước để đánh giá chất lượng của hệ thống. Điểm số MOS của chúng tôi sẽ được đánh giá dựa trên thang đo từ 1 đến 5, với 1 là chất lượng cảm nhận thấp nhất và 5 là chất lượng cảm nhận cao nhất. Đây là bảng điểm MOS được ánh xạ với các nhãn chất lượng từ "Rất tệ" đến "Xuất sắc".

Điểm	1	2	3	4	5
Nhãn	Rất tệ	Tệ	Trung bình	Tốt	Xuất sắc

Bảng 3: Bảng xác định thang đo đánh giá MOS

Để có được kết quả MOS, chúng tôi tính bằng công thức sau:

- $MOS = \frac{\sum_{n=1}^{N} Rn}{N}$, *với N là số đoạn text, R là điểm cá nhân cho mỗi N text.*

Như đã giới thiệu về tập thử nghiệm, chúng tôi sử dụng 25 đoạn text bất kỳ để đánh giá chất lượng hoạt động của mô hình Tacotron2 và WaveGlow, do đó *N* sẽ là 25.

Sau khi tính được MOS của từng người, do chúng tôi đánh giá trên 2 người nghe, do đó để tính được MOS trung bình của mô hình này, chúng tôi gọi là *MOS (a)*, được tính bằng:

- $MOS\ (a) = \frac{MOS(1)+MOS(2)}{2}$