Melhor explicação da validação utilizando métricas de erro para o artigo Conversão Texto-Fala para o Português Brasileiro Utilizando Tacotron 2 com Vocoder Griffin-Lim

Thomaz Diniz Pinto de Morais

June 3, 2024

1 Contextualização aos revisores

O artigo aqui avaliado é o "Conversão Texto-Fala para o Português Brasileiro Utilizando Tacotron 2 com Vocoder Griffin-Lim". O contexto deste artigo é a utilização de uma técnica de deep learning chamada de TACOTRON 2 para a realização da conversão de texto para voz em português brasileiro. Artigo relevante para áreas de machine learning e acessibilidade que são as áreas que pretendo fazer pesquisa. Este artigo não possui formalização matemática alguma, minha proposta é formalizar pelo menos a maneira como os autores decidiram validar o modelo que eles utilizaram. Em tese é algo simples, mas que pode deixar mais fechadinho a maneira como a validação é feita.

Os autores notaram que ao utilizar esta técnica com um conjunto de dados obtidos por eles 2 erros podiam acontecer: Erros de pronúncia e Erros de palavras puladas (ou ignoradas). Minha proposta é fazer uma validação matemática do que é o ideal esperado por um sistema melhor (no caso que estes erros tendam a zero). Ou seja, um sistema de TTS(TEXT TO SPEECH) melhor é aquele que não possui nem erros de pronúncia nem erros de palavras ignoradas.

Minha proposta é para a criação de uma nova seção entre os resultados e conjunto de dados que discuta a validação oferecida por eles. Isso facilitaria no entendimento de que a técnica deles é realmente melhor do que a utilizada para avaliar. Ou seja, a adição de uma seção sobre o "Método de validação".

2 Método de validação

Um modelo é considerado melhor se minimiza o número de palavras puladas ou ignoradas e erros de pronúncia. Seja N o número total de palavras no texto original e I o número de palavras ignoradas pelo modelo TTS. A taxa de palavras ignoradas E_I é dada por:

$$E_I = \frac{I}{N} \tag{1}$$

Seja E o número de pronúncias errôneas detectadas na saída do modelo TTS. A taxa de pronúncias errôneas E_P é dada por:

$$E_P = \frac{E}{N} \tag{2}$$

O que nos dá um percentual de palavras pronunciadas erroneamente (E_P) e um percentual de palavras ignoradas (E_I). O ideal em um sistema de conversão texto-fala é que, para cada palavra no texto de entrada, não exista uma palavra com erro de pronúncia nem uma palavra ignorada. Formalmente, podemos expressar isso da seguinte maneira:

Seja $\mathbf{X} = \{x_1, x_2, \ldots, x_N\}$ o conjunto de palavras no texto original. O sistema TTS deve garantir que nenhuma palavra seja ignorada, ou seja,

$$I = 0 \tag{3}$$

o que implica que:

$$E_I = \frac{I}{N} = \frac{0}{N} = 0 \tag{4}$$

Bem como para toda palavra, o sistema deve garantir que nenhuma palavra seja pronunciada erroneamente, ou seja,

$$E = 0 \tag{5}$$

o que implica que:

$$E_P = \frac{E}{N} = \frac{0}{N} = 0 \tag{6}$$

Portanto, o sistema TTS ideal é aquele para o qual:

$$E_I = 0 \quad \text{e} \quad E_P = 0 \tag{7}$$

Ou, em outras palavras, é desejável que um sistema TTS tenha uma quantidade de erros de pronúncia e erros de palavras ignoradas iguais a zero. E uma maneira objetiva de avaliar se um TTS é melhor que o outro é quanto o TTS sendo avaliado minimiza estes dois tipos de erros.