Formalismo Matemático para Utilização de TTS na Leitura Humanizada de Textos e seus Impactos na Acessibilidade e Entendimento

Thomaz Diniz Pinto de Morais

Junho 2024

1 Introdução

A pesquisa investiga a utilização de TTS (Text-to-Speech) para a leitura humanizada de textos e os impactos dessa tecnologia na acessibilidade e no entendimento dos usuários. Neste documento expressaremos 3 formalismos matemáticos distintos. Dois destes formalismos foram feitos em atividades anteriores, mas que também são relevantes para minha pesquisa e um deles foi adicionado para o caso desta repetição não ser o suficiente para se adequar a atividade da disciplina. Estes são os formalismos que são relevantes para a pesquisa que irei conduzir:

- Formalização de como o TTS que utiliza um autocoder variacional funciona (Primeira atividade de FPCC3)
- Formalização de parâmetros objetivos para definir se um TTS é melhor que outro (Segunda atividade de FPCC3)
- Formalização de parâmetros subjetivos para definir se um TTS atingiu um nível de fala humanizada (Uma terceira formalização para o caso das duas anteriores não serem o suficiente para a atividade)

2 Formalização de um TTS que utiliza um autocoder variacional

A formalização a seguir explica um autoencoder variacional. O objetivo de um autoencoder variacional é codificar um input automaticamente em uma representação comprimida que pode depois ser decodificado para que o input possa ser reconstruído a partir do que foi codificado. Isso pode ser usado para explicar o processo de reconstrução de voz a partir de um treinamento utilizando este autoencoder. Em um primeiro momento é feita a codificação em algo que é

chamado de vetor latente Z. Em um momento do artigo é feita a formalização matemática do autoencoder para explicar seu funcionamento.

X: Voz original

Y: Sequência de texto

$q(z \mid x)$: Encoder que parametriza a distribuição dos dados X para a a variável latente Z.

$p(x \mid z)$: Decoder que reconstrói o input que foi passado que no caso é o audio passado.

$p(z \mid y)$: Um segundo encoder que codifica uma sequência de texto Y em fonemas. E será passado ao Decoder posteriormente

$S : p(z \mid y) \rightarrow p(x \mid z)$: Voz Sintetizada a partir do processo de decodificação onde em um primeiro momento se obtém o latente Z a partir do texto, para depois ser feita a decodificação do spectograma que é utilizado em um segundo momento para formar a voz em si.

3 Formalização de parâmetros objetivos para definir se um TTS é melhor que outro

Dado que temos um modelo de TTS, podemos avaliar o quão bom é um modelo. Um modelo é considerado melhor se minimiza o número de palavras 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.

4 Formalização de parâmetros subjetivos para definir se um TTS atingiu um nível de fala humanizada

A leitura humanizada pode ser classificada como uma leitura que possui os áudios com entonação, ritmo e pausas naturais que imitam a fala humana. tan et al [1] define que um bom parâmetro para se ter uma leitura humana é utilizar uma avaliação por falantes nativos da lingua com avaliações do tipo MOS(mean opinion score) com a escala likert, comparando com leitores nativos. Se as médias entre os leitores nativos e o sistema TTS forem estatisticamente iguais, então o TTS é capaz de fazer uma leitura humanizada (ou bem próximo da qualidade humana). Portanto, há margem de fazer esta formalização também.

Utilizamos uma escala Likert de 1 a 5 para as avaliações, onde 1 representa a pior qualidade e 5 a melhor qualidade.

Sejam M_H a média dos escores MOS para falantes humanos nativos e M_T a média dos escores MOS para o sistema TTS. Nosso objetivo é comparar M_H e M_T.

Formulamos as hipóteses estatísticas:

- H_0: $M_H = M_T$ (As médias são estatisticamente iguais)
- H_a: $M_H \neq M_T$ (As médias são estatisticamente diferentes)

A partir destas hipóteses conduziremos nosso teste T de student para testar se as médias são ou não iguais utilizando o $\alpha = 0.05$.

$$t = \frac{M_H - M_T}{\sqrt{\frac{s_H^2}{n_H} + \frac{s_T^2}{n_T}}} \tag{8}$$

onde s_H e s_T são os desvios padrões das amostras de falantes humanos e TTS, e n_H e n_T são os tamanhos das amostras de humanos e de TTS, respectivamente. Com isso encontraremos a estatística t e com esse valor Calculamos o valor-p associado ao teste t onde $p - valor = P(|t| < T)$ que nos dará um valor maior ou menor que α. Se o p-valor for menor que α, rejeitamos a hipótese nula H_0; caso contrário, não rejeitamos H_0.

References

[1] Xu Tan, Jiawei Chen, Haohe Liu, Jian Cong, Chen Zhang, Yanqing Liu, Xi Wang, Yichong Leng, Yuanhao Yi, Lei He, Frank Soong, Tao Qin, Sheng Zhao, and Tie-Yan Liu. Naturalspeech: End-to-end text to speech synthesis with human-level quality, 2022.