Desenvolvimento de um modelo de reconhecimento de voz para o Português Brasileiro com poucos dados utilizando o Wav2vec 2.0

Lucas Rafael Stefanel Gris[1]*, Edresson Casanova[2], Frederico Santos de Oliveira[3], Anderson da Silva Soares[4], Arnaldo Candido Junior[1]

[1]Universidade Tecnológica Federal do Paraná, Medianeira, Brasil

[2]Universidade de São Paulo, São Carlos, Brasil

[3]Universidade Federal do Mato Grosso, Cuiabá, Mato Grosso

[4]Universidade Federal de Goiás, Goiânia, Brasil

Abstract. *Deep learning techniques have been shown to be efficient in various tasks, especially in the development of speech recognition systems. Despite the progress in the area, its development can still be considered a difficult task, especially when there is a lack of open data available, as in Brazilian Portuguese. Considering this limitation, the Wav2vec 2.0, a deep neural network architecture that does not need a lot of labelled data, can be an interesting alternative. In this sense, this work proposes to validate the development of an automatic speech recognition system using a few freely available data in a fine-tuning of Wav2vec 2.0 model pre-trained in many languages for the Brazilian Portuguese. This work shows that it is possible to develop an automatic speech recognition system using only 1h of transcribed speech. The fine-tuned model presents an WER of only 34% against the Common Voice dataset.*

Resumo. *Técnicas de aprendizado profundo têm se mostrado muito eficientes nas mais diversas tarefas, em especial, no desenvolvimento de sistemas de reconhecimento de voz. Apesar do avanço na área, seu desenvolvimento ainda pode ser considerado uma tarefa difícil, especialmente em idiomas que apresentam poucos dados abertos disponíveis, como o Português Brasileiro. Considerando essa limitação, o Wav2vec 2.0, uma arquitetura que dispensa a necessidade de uma grande quantidade de áudios rotulados, pode ser uma alternativa interessante. Nesse sentido, este trabalho apresenta como objetivo avaliar o desenvolvimento de um reconhecedor de voz utilizando poucos dados disponíveis gratuitamente a partir do ajuste do modelo Wav2vec 2.0 pré-treinado em muitas línguas. Este trabalho mostra que é possível construir um sistema de reconhecimento de voz utilizando apenas 1h de fala transcrita para o Português Brasileiro. O modelo ajustado apresenta um WER de somente 34% contra o dataset da Common Voice.*

1. Introdução

A fala é o meio mais natural de comunicação humana. Segundo [Goodfellow et al. 2016], a tarefa de reconhecer a fala automaticamente pode ser definida como o mapeamento

* Autor correspondente: gris at alunos utfpr dot edu dot br

de um sinal acústico contendo uma sentença falada para uma sequência correspondente de palavras escritas pretendida pelo locutor. Apesar dos grandes avanços na área, os modelos estado-da-arte de ASR necessitam de uma grande quantidade de dados para que seja possível alcançar um desempenho aceitável [Amodei et al. 2016, Quintanilha et al. 2020]. Isso pode ser um grande problema em idiomas como o Português Brasileiro, que apresenta poucos recursos abertos disponíveis [Neto et al. 2011, Neto et al. 2008]. Apesar desse problema, novas técnicas de aprendizado de máquina têm surgido e melhorado significativamente o desenvolvimento de modelos com poucas quantidades de dados rotulados [Baevski and Mohamed 2020, Conneau et al. 2020, Conneau and Lample 2019, Yi et al. 2020].

Uma das técnicas recentes que têm apresentado resultados promissores é a técnica de aprendizado auto-supervisionado. Esse tipo de técnica tem se mostrado como um ótimo paradigma no aprendizado de representações a partir de dados não rotulados e o ajuste posterior do modelo em dados rotulados [Baevski et al. 2020b]. O Wav2vec 2.0 [Baevski et al. 2020b] é uma das arquiteturas que seguem esse princípio. Os autores demonstram que o modelo ajustado é capaz de aprender a transcrever áudios mesmo com uma quantidade limitada de áudios transcritos. Nesse sentido, a utilização do Wav2vec 2.0 como base para o desenvolvimento de um ASR em português brasileiro se mostra bastante promissora.

Este trabalho propõe a utilização do modelo Wav2vec 2.0, uma arquitetura robusta baseada em um tipo moderno de redes neurais conhecida como Transformers [Vaswani et al. 2017], para validar o desenvolvimento de um ASR em português brasileiro utilizando um *dataset* com apenas 1 hora de fala transcrita.

O trabalho está organizado da seguinte maneira: a Seção 2 irá abordar alguns conceitos relacionados às técnicas de AM, a Seção 3 irá tratar sobre a arquitetura utilizada, o Wav2vec 2.0, e a Seção 4 irá abordar alguns trabalhos relacionados. Posteriormente, a Seção 5 irá explicar os métodos propostos. Por fim, a Seção 6 irá mostrar e discutir os resultados obtidos.

2. Redes Neurais Artificiais e Aprendizado Profundo

As Redes Neurais Artificiais (ANNs) são um tipo de algoritmo específico de AM [Goodfellow et al. 2016] e são inspiradas no funcionamento do cérebro humano [Haykin et al. 2009]. Existem diversos tipos de ANNs, dentre elas, Redes Convolucionais e Redes Recorrentes, e mais recentemente, redes do tipo Transformer, que têm apresentado grandes avanços na área.

As Redes Neurais Convolucionais (*Convolutional Neural Network*s) (CNNs) apresentam grande aplicabilidade em detecção de padrões em imagens. As CNNs são redes que empregam operações conhecidas como convoluções, um tipo especial de operação linear [Goodfellow et al. 2016]. Os principais conceitos introduzidos pelas CNNs são: campos receptivos locais, que capturam características em uma região específica; pesos compartilhados, que possibilitam a varredura em todo o espaço e otimizam o treinamento de uma maneira geral, e *Pooling*, que diminui a dimensionalidade de um tensor [Nielsen 2015].

Outro tipo de ANN é a rede recorrente. As redes recorrentes simples e com portas (como LSTMs) apresentaram avanços na área de modelos voltados para dados se-

quenciais e se destacaram como estado-da-arte em muitos problemas dessa área, como tradução de máquina e modelagem de língua [Bahdanau et al. 2015, Cho et al. 2014, Sutskever et al. 2014]. Apesar desse avanço, esses tipos de redes neurais apresentam alguns problemas, como uma maior dificuldade em lidar com sequências muito longas, dissipação do gradiente e também a dificuldade em paralelizar as entradas.

Para solucionar as limitações das redes recorrentes, [Vaswani et al. 2017] propõem um novo tipo de arquitetura de redes neurais chamado Transformer. Esse tipo de arquitetura é construído a partir de mecanismos de auto-atenção e codificação posicional, e elimina a necessidade de qualquer tipo de recorrência para a aprendizagem de dependências globais entre as entradas e as saídas do modelo. O Transformer segue uma estrutura básica de *encoder-decoder*, que é um tipo de estrutura bem sucedida em várias tarefas de transformação (ou transdução) de entradas para saídas sequenciais [Bahdanau et al. 2015, Cho et al. 2014, Sutskever et al. 2014].

3. Wav2vec 2.0

O Wav2vec 2.0 é um modelo de ponta-a-ponta inspirado nos trabalhos antecessores do Wav2vec [Schneider et al. 2019] e Vq-Wav2vec [Baevski et al. 2020a]. Assim como o Vq-Wav2vec, o modelo é pré-treinado de forma auto-supervisionada no domínio do tempo (áudio bruto) para construir representações discretas da fala. A topologia do Wav2vec 2.0 pode ser vista na Figura 1.

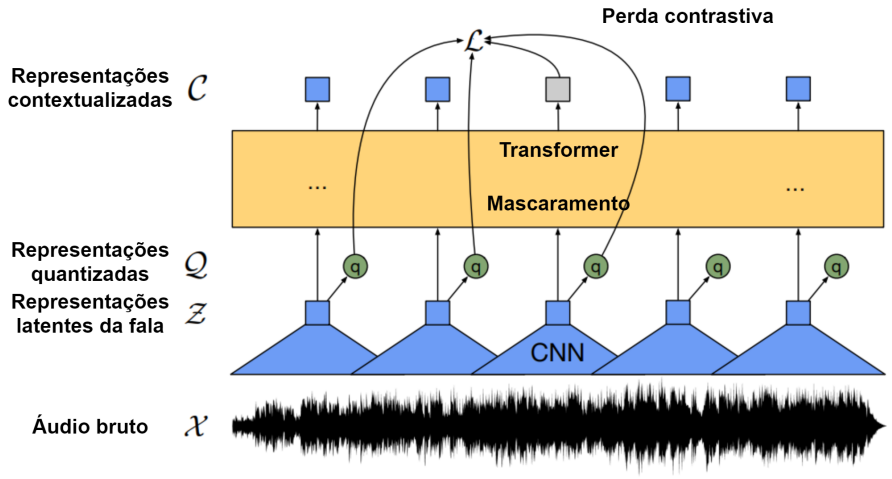

Figura 1. Arquitetura do Wav2vec 2.0 - Adaptado de [Baevski et al. 2020b].

A arquitetura do Wav2vec 2.0 é composta por um *encoder* convolucional multicamada $f : X \longmapsto Z$ que mapeia o áudio bruto X em representações latentes da fala $z_1, \ldots, z_T$ em T *time-steps*. A saída do *encoder* é então fornecida à rede de contexto $g : Z \longmapsto C$, que segue a arquitetura Transformer, e que mapeia as representações latentes Z em representações contextualizadas $c_1, \ldots, c_T$.

De forma similar ao BERT, durante a fase de treinamento, alguns *time-steps* das saídas do *encoder* são mascarados antes de serem fornecidos à rede de contexto. O mascaramento tem o objetivo de fazer com que o modelo seja capaz de predizer as partes mascaradas. As entradas do módulo de quantização não sofrem mascaramento [Baevski et al. 2020b].

Após o pré-treinamento auto-supervisionado o modelo pode ser ajustado (*fine-tuning*) em uma tarefa supervisionada, como o ASR, adicionando uma projeção linear após as saídas da rede de contexto contendo n classes e uma função de custo específica, como a CTC.

Uma das formas de se medir a qualidade de um sistema de ASR é a métrica WER (Word Error Rate)[McCowan et al. 2004]. Outras métricas populares são a CER (Character Error Rate) e Perplexidade.

[Baevski et al. 2020b] demonstram que é possível construir modelos ASR mesmo com poucos dados anotados. Os experimentos de ajuste (*fine-tuning*) com 10 minutos de dados obtiveram um WER de 4,8/8,2% nos dois conjuntos de teste do LibriSpeech (clean e other).

O Wav2vec 2.0 pode ser utilizado em outros idiomas além do inglês. [Conneau et al. 2020] disponibilizaram um modelo de Wav2vec 2.0 pré-treinado em vários idiomas, denominado XLSR-53. O modelo demonstra uma redução de 72% de WER em reconhecimento de fonemas no *benchmark* da Common Voice [Ardila et al. 2020], se comparado com os melhores resultados obtidos até então, além de uma redução de 16% de WER na comparação com o *baseline* proposto na tarefa de transcrição de caracteres. O XLSR-53 é um modelo que segue a arquitetura LARGE do Wav2vec 2.0 e é treinado em um conjunto de dados contendo 53 horas de áudios a partir de vários *datasets* contendo falas não transcritas em vários idiomas.

4. Trabalhos correlatos

Apesar das pesquisas na área do ASR serem bastante extensas, há poucos trabalhos relacionados ao reconhecimento de voz aberto para o Português Brasileiro. Uma possível explicação para este fato é a limitação de dados anotados disponíveis e a ausência de modelos capazes de generalizarem bem quando há poucos dados rotulados para treinamento. Entretanto, com novas pesquisas surgindo, como os trabalhos de [Schneider et al. 2019], o desenvolvimento de sistemas ASR abertos para o português brasileiro com desempenho aceitável se tornam mais próximos da realidade.

Na área de ASRs abertos para o Português, os trabalhos de [Quintanilha 2017] e [Quintanilha et al. 2020] podem ser destacados como avanços recentes importantes. [Quintanilha 2017] propôs a elaboração de um *dataset* em Português Brasileiro composto pela junção de vários outros conjuntos de dados disponíveis. O modelo de ponta-a-ponta proposto pelo autor é baseado em uma arquitetura básica contendo camadas recorrentes bidirecionais (BiLSTMs). Os áudios são pré-processados e as características são extraídas para serem fornecidas como entrada ao modelo. O autor obteve um WER de 25,13% no conjunto de teste proposto, 11% maior que os sistemas comerciais comparados pelo trabalho. Mais recentemente, [Quintanilha et al. 2020] propuseram uma versão melhorada do dataset utilizado em [Quintanilha 2017] ao adicionar o *dataset* CETUC [Alencar and Alcaim 2008], contendo aproximadamente 145 horas de fala, ao conjunto de dados do trabalho anterior, para o treinamento de uma topologia baseada no DeepSpeech 2. Os autores alcançaram um WER de 25,45% no conjunto de teste proposto.

5. Métodos

Para o ajuste do modelo de ASR propõe-se a realização do ajuste do modelo XLSR-53 no *dataset* LapsBM[1], que contém aproximadamente 1 hora de fala transcrita, e a realização do teste do modelo final obtido contra o conjunto de teste do *dataset* da Common Voice. Diversos experimentos foram realizados testando diferentes configurações de taxa de aprendizado e a quantidade máxima de iterações. O melhor experimento foi utilizado para a realização do teste contra o conjunto de teste proposto.

O conjunto de validação da Common Voice também foi selecionado para a realização dos experimentos, mais especificamente, como conjunto de validação no *fine-tuning*. Os conjuntos de validação e teste da Common Voice foram escolhidos para validação e teste nos experimentos propostos porque são conjuntos validados e que correspondem a utilização em ambientes reais, já que são áudios gravados por voluntários a partir de seus equipamentos pessoais. Além disso, os voluntários da Common Voice estão diversificados em várias regiões do país, o que garante uma melhor avaliação de desempenho em diferentes situações e sotaques. A escolha dos conjuntos propostos também viabiliza futuras comparações e pesquisas com os resultados obtidos neste trabalho. Todos os áudios foram reamostrados para 16kHz antes de serem utilizados.

O conjunto de treinamento da Common Voice não foi utilizado neste trabalho, que procurou validar a utilização do modelo XLSR-53 em situações com poucos dados rotulados. Entretanto, é provável que o ajuste do modelo com mais dados, similar a [Quintanilha et al. 2020], produza resultados melhores. Os *datasets* utilizados são detalhados na Seção 5.1.

Para o treinamento dos modelos o *framework* Fairseq[2] (Facebook AI Research Sequence-to-Sequence Toolkit) foi utilizado. Os modelos foram treinados em uma máquina contendo as seguintes especificações: processador Intel Core i7-8700, 16GB de RAM e GPU NVIDIA GTX Titan V com 12GB de RAM. Essas configurações podem ser consideradas modestas, principalmente se comparadas às configurações utilizadas em [Baevski et al. 2020b], e considerando também o tamanho da topologia Wav2vec 2.0 LARGE, que apresenta mais de 300 milhões de parâmetros treináveis na fase de *fine-tuning*.

5.1. *Datasets*

O LapsBM 1.4 é um *dataset* utilizado pelo grupo Fala Brasil para avaliar sistemas ASR em Português Brasileiro. Contém 35 locutores (10 mulheres), cada um pronunciando 20 falas únicas, totalizando 700 sentenças. O áudio foi gravado em 22,05 kHz sem controle do ambiente. Segundo os desenvolvedores do projeto, todas as gravações foram realizadas em computadores utilizando microfones comuns. O *dataset* é originalmente proposto para ser utilizado como teste de sistemas de ASR.

A Common Voice [Ardila et al. 2020] é um projeto iniciado pela Mozilla para criar uma base de dados gratuita para sistemas reconhecedores de fala. Nesse projeto, voluntários doam e validam vozes por meio de um *site*[3]. Existem vários idiomas dis-

[1]"Falabrasil – UFPA" (`https://github.com/falabrasil/gitlab-resources`)

[2]`https://github.com/pytorch/fairseq`

[3]`https://commonvoice.mozilla.org/pt`.

poníveis. O conjunto em Português versão 5.1 (pt_53h_2020-06-22), utilizada neste trabalho, contém aproximadamente 48 horas validadas e 744 locutores únicos. Os conjuntos de teste e validação apresentam mais de 4000 sentenças.

6. Resultados e discussões

Vários experimentos foram realizados variando a taxa de aprendizado máxima (5×10^{-5}, 10^{-4}, 5×10^{-4} e 10^{-3}) e a quantidade máxima de atualizações (13 mil e 20 mil). A taxa de aprendizado também varia durante o treinamento, pois sua configuração segue o padrão definido por [Baevski et al. 2020b], em que o valor varia em três estágios distintos de aumento, platô e decaimento. O melhor experimento foi configurado com taxa de aprendizado igual a 5×10^{-5} (a menor das testadas) e um máximo de iterações igual a 20 mil, o que possibilitou o treinamento do modelo em até aproximadamente 500 épocas. O valor de WER e custo na validação podem ser visualizados nos gráficos da Figura 2. Os experimentos com taxa de aprendizado igual a 5×10^{-4} e 10^{-3} não convergiram, o que sugere que taxas menores podem ser mais interessantes. Alguns parâmetros relacionados ao mascaramento testadas não produziram resultados promissores.

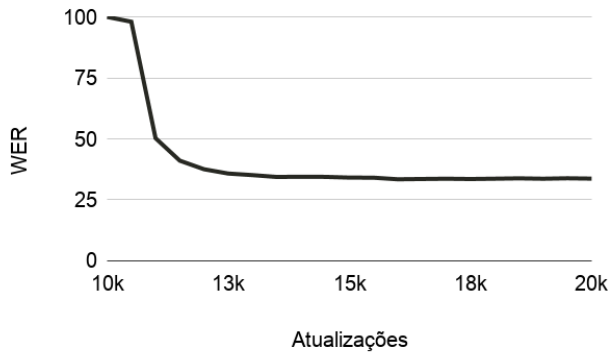

Figura 2. Gráficos de WER e custo (perda) para o melhor experimento.

O tamanho do *batch* é configurado automaticamente a partir do valor máximo de *frames* definido, portanto, devido a limitações de hardware, esse tamanho ficou em torno de 12, durante as 10 mil primeiras atualizações, em que as camadas Transformer não são ajustadas, e 8, nas atualizações restantes, em que as camadas Transformer são descongeladas. O extrator de características nunca é ajustado. É possível que melhores resultados fossem alcançados se o modelo fosse treinado em hardwares mais robustos, pois isso viabilizaria *batches* maiores.

O melhor modelo foi selecionado quando o menor WER de validação foi obtido durante afase de ajuste pelo próprio *framework* utilizado. O teste contra o conjunto de teste da Common Voice do melhor modelo obteve um WER de 34%. Esse resultado pode ser considerado promissor, principalmente considerando as limitações de hardware utilizadas no treinamento do modelo, a quantidade de áudios transcritos utilizados para o ajuste e as características dos *datasets* utilizados neste trabalho (cerca de 1 hora), isto é, áudios gravados em equipamentos comuns e que apresentam ruído de ambiente.

7. Conclusão

Este trabalho procurou realizar experimentos para validar a construção de um sistema de ASR aberto para o português brasileiro utilizando apenas 1h de fala transcrita. O modelo ajustado obteve um WER de apenas 34% no conjunto de teste da Common Voice, um resultado promissor se comparado aos melhores modelos abertos disponíveis para o português brasileiro. De uma forma geral a principal contribuição desse trabalho foi mostrar que é possível realizar o desenvolvimento de ASRs robustos para a língua portuguesa, mesmo com poucos dados rotulados disponíveis. Espera-se que o ajuste em mais dados possa melhorar o resultado de forma significativa, contribuindo assim para o desenvolvimento de modelos de ASR abertos para o Português Brasileiro.

Agradecimentos

Agradecemos a Corporação NVIDIA pela doação da GPU usada nos experimentos apresentados. Agradecemos também à CyberLabs pelo apoio para a realização dos experimentos deste trabalho.

Referências

Alencar, V. and Alcaim, A. (2008). Lsf and lpc-derived features for large vocabulary distributed continuous speech recognition in brazilian portuguese. In *2008 42nd Asilomar Conference on Signals, Systems and Computers*, pages 1237–1241. IEEE.

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., et al. (2016). Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182. PMLR.

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., and Weber, G. (2020). Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215.

Baevski, A. and Mohamed, A. (2020). Effectiveness of self-supervised pre-training for asr. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7694–7698.

Baevski, A., Schneider, S., and Auli, M. (2020a). vq-wav2vec: Self-supervised learning of discrete speech representations. In *International Conference on Learning Representations (ICLR)*.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020b). wav2vec 2.0: A framework for self-supervised learning of speech representations. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. 3rd International Conference on Learning Representations, ICLR 2015.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder

for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, É., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Conneau, A. and Lample, G. (2019). Cross-lingual language model pretraining. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

Haykin, S. S. et al. (2009). *Neural networks and learning machines/Simon Haykin.* New York: Prentice Hall,.

McCowan, I. A., Moore, D., Dines, J., Gatica-Perez, D., Flynn, M., Wellner, P., and Bourlard, H. (2004). On the use of information retrieval measures for speech recognition evaluation. Technical report, IDIAP.

Neto, N., Patrick, C., Klautau, A., and Trancoso, I. (2011). Free tools and resources for brazilian portuguese speech recognition. *Journal of the Brazilian Computer Society*, 17(1):53–68.

Neto, N., Silva, P., Klautau, A., and Adami, A. (2008). Spoltech and ogi-22 baseline systems for speech recognition in brazilian portuguese. In *International Conference on Computational Processing of the Portuguese Language*, pages 256–259. Springer.

Nielsen, M. A. (2015). *Neural networks and deep learning*, volume 25. Determination press San Francisco, CA, USA:.

Quintanilha, I. M. (2017). End-to-end speech recognition applied to brazilian portuguese using deep learning. *MSc dissertation*.

Quintanilha, I. M., Netto, S. L., and Biscainho, L. W. P. (2020). An open-source end-to-end asr system for brazilian portuguese using dnns built from newly assembled corpora. *Journal of Communication and Information Systems*, 35(1):230–242.

Schneider, S., Baevski, A., Collobert, R., and Auli, M. (2019). wav2vec: Unsupervised pre-training for speech recognition. In *INTERSPEECH*.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA. MIT Press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Neural Information Processing Systems (NIPS)*.

Yi, C., Wang, J., Cheng, N., Zhou, S., and Xu, B. (2020). Applying wav2vec2. 0 to speech recognition in various low-resource languages. *arXiv preprint arXiv:2012.12121*.