# Avaliação de Sistemas de Reconhecimento de Fala Robustos a Ruídos Provenientes de Maquinário da Indústria do Petróleo

**Vinicius de Souza Nunes**[1], **Julio Cesar Duarte**[1]

[1]Instituto Militar de Engenharia (IME)
Praça Gen. Tibúrcio, 80 – 22.290-270 – Urca, Rio de Janeiro – RJ – Brasil

{viniciusnunes, duarte}@ime.eb.br

***Abstract.*** *In recent years, automatic speech recognition systems have evolved from rigid rules to probabilistic models, due to advances in deep neural networks. However, speakers can be exposed to noise, which often impairs speech-to-text transcription. This work aims to evaluate the application of neural networks in noise-robust automatic speech recognition, more specifically, those noises from the machinery of the oil industry. To achieve this objective, a base of representative noises was built, and, in our experiments, the result of the average CER for the developed models with the same type of noise was 0.421984, when the SNR of the training is the same as the test, and 0.522851, when such ratios are different.*

***Resumo.*** *Nos últimos anos, sistemas de reconhecimento automático de fala evoluíram de regras rígidas para modelos probabilísticos, graças ao avanço em redes neurais. No entanto, locutores podem estar expostos a ruídos, o que prejudica a transcrição da fala. Assim, este trabalho tem por objetivo avaliar a aplicação de redes neurais no reconhecimento automático de fala robusto a ruídos, mais especificamente, àqueles provenientes do maquinário da indústria do petróleo. Para atingir o objetivo, foi construída uma base de ruídos representativos, e, nos experimentos, o resultado da média do CER para os modelos desenvolvidos com o mesmo tipo de ruído foi de 0,421984, sendo a SNR do treino a mesma do teste, contra 0,522851 quando tais relações são diferentes.*

## 1. Introdução

O Reconhecimento Automático de Fala (ASR, do Inglês, *Automatic Speech Recognition*) tem sido um assunto de grande interesse para um conjunto cada vez maior de pessoas desde a sua popularização em grandes filmes de sucesso, como foi o caso em "2001 – Uma Odisseia no Espaço", em 1968, onde um computador inteligente era capaz de falar e interpretar comandos de voz [Juang and Rabiner 2005]. No entanto, não foi somente a popularização desse tópico, a responsável pelo avanço nas tecnologias de ASR ocorridas nas últimas décadas. A grande revolução tecnológica ocorreu a partir do momento em que as pesquisas deixaram de se concentrar na codificação de regras rígidas, para focar em modelos estatísticos de tomada de decisões de forma suave e probabilística, principalmente com a introdução dos modelos ocultos de Markov (HMM, do Inglês, *Hidden Markov Models*) e das Redes Neurais.

Atualmente, muitas das arquiteturas dos ASRs são compostas pela combinação de modelos acústicos, léxicos e de linguagens, tendo início com a extração de características

acústicas da fala, e finalizando com a construção de um modelo estatístico para determinar a combinação mais provável para a sequência de palavras. Além disso, diferentes arquiteturas vêm sendo estudadas com o intuito de melhorar o desempenho das tecnologias ASR na presença de condições adversas. [Mattys et al. 2012] classificam condições adversas em três grupos principais. O primeiro é o da degradação do sinal pelo locutor, como sotaque, falta de fluência ou desordem neurogênica. Já o segundo é o da degradação devido ao ambiente e a transmissão. Por último, temos o da limitação do receptor. Nesse trabalho, nosso interesse é no segundo grupo, mas especificamente na degradação do sinal da fala devido ao ruído ambiente de uma plataforma petrolífera. Diferente de outras abordagens de conversão de voz em texto que tratam da remoção do ruído, este trabalho opta pelo treinamento do modelo robusto a um tipo de ruído específico.

Dessa maneira, trabalhadores inseridos em um ambiente muito ruidoso, como uma plataforma de petróleo, poderiam se comunicar de forma efetiva mesmo na presença do ruído industrial, sem colocar em risco sua saúde ou segurança. Esse estudo consiste na base para a implementação dessa ideia, onde, por exemplo, trabalhadores utilizariam abafadores inteligentes que conectados entre si seriam capazes de interpretar comando de voz, transcrever a voz para texto, e no destinatário sintetizar a voz no abafador destino, assim acabando com qualquer interferência do ruído.

Esse trabalho tem como objetivo analisar o impacto do treinamento de ASRs com ruído comparando-o com versões treinadas sem ruído ou ruídos distintos. O ruído aqui considerado é o emitido pelos maquinários que compõem uma plataforma de petróleo, no processo de produção de óleo e gás natural. Ademais, o ruído do maquinário presente na proa, no meio e na popa da plataforma é sobreposto ao corpus de fala utilizado.

Assim, a partir da análise da taxa de erro de caracteres (CER, do Inglês *Character Error Rate*) resultante nos testes, é possível melhor compreender o comportamento de ASRs quando expostos a esse tipo de ruído.

Este artigo está dividido nas seguintes partes. A seção 2 descreve a literatura relevante ao tema de ASRs. Já na seção 3, contextualizamos as ferramentas utilizadas na metodologia que é apresentada na seção 4. A seção 5 analisa os resultados dos experimentos propostos. Finalmente, concluímos o trabalho apresentando suas possibilidade futuras na seção 6.

## 2. Revisão da Literatura

Nos últimos anos, devido principalmente aos avanços tecnológicos dos hardwares, como as GPUs que permitem cálculos matriciais com ótimo desempenho, vem crescendo a pesquisa com foco em redes neurais artificiais, normalmente empregadas no treinamento de ASRs. Contudo, o abundante material disponível se torna escasso quando o reconhecimento automático de fala é limitado ao idioma português, conjuntos de dados robustos e na presença de ruído - principalmente quando esse último for o industrial. De qualquer forma, trabalhos com características similares podem servir como base de comparação.

[Prodeus and Kukharicheva 2016] compara as técnicas de treino *Fully Matched Traininig* (FMT) e *Spectrum Matched Training* (SMT) em ambientes ruidosos com o modelo ASR treinado com falas sem interferências. Foram utilizados no experimento 14 tipos de ruído, tais como: doméstico, computadores, rua, transporte, sala de aula e

*lobbies*. Para isso, os áudios receberam uma mistura aditiva do tipo desejado, variando o SNR entre 0 e 45 dB. O resultado mostra uma melhora na precisão do modelo quando o SNR do treino se aproxima do mesmo valor de SNR do teste.

Já [Saon et al. 2019] comprova que a injeção de ruídos de sequência nas entradas melhora a capacidade de generalização do modelo ponta a ponta, mais do que adicionar ruído gaussiano simples ou *frames* da fala sem a informação de sequência.

Por outro lado, [Kinoshita et al. 2020] avalia a melhora de desempenho *front-end* no reconhecimento de fala para ASR de canal único. Para isso, investigam se o uso de redes neurais *time-domain* na eliminação de ruído é capaz de melhorar o desempenho do ASR quando comparado a um ASR *back-end*.

[Pervaiz et al. 2020] apura o comportamento do treino de modelos ASR quando utilizada a estratégia de aumentação de dados, através da adição de ruído ao áudio original. Além disso, foram aplicados quatro tipos de arquiteturas do modelo acústico, GMM-HMM, DNN-HMM, LSTM-HMM, e CNN, sendo os testes realizados tanto com dados limpos quanto ruidosos.

Finalmente, [Quintanilha et al. 2020] monta um corpus de 158 horas de áudios de fala em português, a partir da junção de 4 bases de dados, sendo então o modelo acústico treinado com esses áudios, na ferramenta *Deep Speech*. Ademais, foi incluído um modelo de linguagem de 15-gram resultando em um CER de apenas 10,49. Tal trabalho entretanto não considera a adição de ruídos na sua avaliação.

## 3. Ferramentas

Modelos ASR eficientes dependem, em sua grande maioria, de dois pilares: uma base de áudios volumosa com suas transcrições e tecnologias para treinar redes neurais, a fim de detectar padrões e correlações. Para essas finalidades, são utilizados nesse trabalho dois projetos da *Mozilla*, o corpus de fala *Common Voice* e o *framework DeepSpeech*.

O *Common Voice* é um conjunto de dados de fala disponível para vários idiomas. Nele, colaboradores contribuem com a leitura de sentenças e validação dos áudios, sendo o resultado final diversos arquivos de áudio de locutores distintos e suas transcrições para texto. Ademais, os dados de fala são liberados sob uma licença *Creative Commons* (*CCO*), resultando no maior corpus de domínio público para fins de reconhecimento de voz, tanto em número de horas quanto de idiomas [Ardila et al. 2019].

Os experimentos nesse trabalho utilizaram o *Common Voice Corpus 6.1*, de 11 de dezembro de 2020, versão para o idioma Português. A base de dados é formada por áudios no formato *mp3*, totalizando 2 GB e 50 horas validadas.

Já o *Deep Speech* é uma tecnologia para desenvolvimento de ASRs de código aberto, desenvolvida pela *Mozilla*, e utiliza como base o *framework TensorFlow* da *Google*. O *Deep Speech* possibilita definir os caracteres que compõem o alfabeto desejado e a inclusão de um modelo de linguagem. Diversos hiper-parâmetros podem ser facilmente definidos para os treinos, tais como: largura das camadas, taxa de *dropout* em cada camada, taxa de aprendizado do otimizador *Adam*, e o tamanho do *batch* para o treinamento.

A arquitetura é fortemente motivada no trabalho apresentado por [Hannun et al. 2014] e utiliza DNNs no treinamento dos modelos, assim sendo

desnecessário um dicionário de fonemas. Porém, o modelo DNN do *DeepSpeech* possui suas particularidades sendo composto por 5 camadas escondidas. As 3 primeiras e a quinta são não-recorrentes e utilizam uma função de ativação *clipped ReLU*. Já a quarta camada é recorrente composta por células LSTM. Por fim, a camada de saída corresponde à probabilidade de cada caractere do alfabeto utilizado.

## 4. Utilizando o Ruído no treinamento de ASRs

A metodologia proposta para o trabalho consiste primeiramente nas etapas de preparação do conjunto de dados de treino e teste para os modelos dos ASR (seções 4.1 e 4.2). Na sequência, são definidas as estratégias de treino e testes dos modelos, variando a relação sinal-ruído e até mesmo o próprio ruído (seções 4.3, 4.4 e 4.5).

### 4.1. Aplicação do Ruído

O ruído utilizado nos experimentos foi captado próximo ao maquinário responsável pelo tratamento primário na produção de óleo e gás natural e gravado em uma plataforma da Petrobras em funcionamento. O resultado da gravação compôs 3 arquivos de 10 minutos, e cada um desses áudios foi obtido em um local distinto na plataforma, mais especificamente na proa, no meio e na popa. A importância da gravação dos áudios em locais distintos é justificada pelo diferente conjunto de máquinas em cada setor. A base de áudios de ruído, construída durante o trabalho, está disponibilizada no repositório do projeto[1].

A aplicação do ruído aos áudios de fala, disponíveis no projeto *Common Voice 6.1* – português, foi realizada por intermédio do recurso de sobreposição disponibilizado no *Deep Speech*, sendo o ruído necessário para a sobreposição selecionado em pontos aleatórios do arquivo do ruído. Além disso, foram aplicadas as relações sinal-ruído (SNR) de 10 e 30. A relação SNR representa a razão entre a potência do sinal e a potência do ruído contida no sinal [Johnson 2006]. Com isso, a base de dados original de 51.714 arquivos de áudio em português foi expandida para 361.998 arquivos, porém, nem todos os áudios foram utilizados nos experimentos, haja vista que vários deles quando comparados com a transcrição não passaram pelos critérios de aprovação do *Common Voice*.

### 4.2. Separação do Treino e Teste

Originalmente, na versão utilizada do *Common Voice*, o teste possui 4.641 sentenças e o treino 6.514. Além destes, existe também o arquivo com todas as transcrições validadas pelos colaboradores perfazendo um total de 41.584 sentenças. No entanto, de forma a aumentar a base de dados de treino do modelo ASR, realizamos uma estratégia de separação do corpus que expande o arquivo de treino com base em todos os áudios validados. Para os dados de testes, entretanto, de forma a garantir comparações futuras com outros trabalhos, o arquivo original é mantido. Assim, um locutor existente no arquivo de treino original tem todas as suas transcrições do arquivo de áudios validados adicionadas a esse arquivo, resultando em um arquivo de treino expandido com 28.368 transcrições. Por fim, o arquivo de treino resultante é embaralhado.

### 4.3. Comportamento na Variação do SNR

O primeiro experimento consiste em geração de modelos com e sem ruído, tendo como propósito a avaliação do comportamento do *CER* conforme a variação da relação SNR.

[1]https://github.com/vsnunesrj/oil-platform-noise

Primeiramente, foram treinados os modelos com ruído da proa nas relações SNR10, SNR30 e SNR$\infty$ (sem ruído). Na sequência, foram efetuados os testes dos modelos nas mesmas proporções de sinal-ruído, porém com o ruído da proa, do meio e da popa da plataforma petrolífera. Para o treinamento dos modelos foi utilizado o *Deep Speech*, com os seus hiper-parâmetros padrão, exceto pela largura das camadas e a quantidade de épocas, que foram respectivamente de 100 e 200, seguindo recomendações de trabalhos para outros idiomas.

Para validação desses resultados, foram treinados modelos com ruído do meio da plataforma aplicando a mesma metodologia anterior no treino e no teste. O experimento também foi repetido com modelos treinados com o ruído do maquinário da popa.

### 4.4. Aplicando a Aumentação de Dados

O experimento seguinte depende da utilização da base de dados aumentada, mais especificamente com a sobreposição de ruído, descrita na seção 4.1. A aplicação da técnica de aumentação de dados costuma apresentar benefícios no aprimoramento de modelos ASR, contudo, o objetivo desse experimento é avaliar a melhora do modelo somada ao comportamento em diferentes níveis de ruído nos testes. Nesse trabalho foram consideradas as relações sinal-ruído de 10 e 30, visto que essas faixas são significativas para a compreensão dos modelos, o que pode se observado tanto em [Mošner et al. 2019] onde são treinados modelos com SNR variando de 0 a 30, quanto em [Prodeus and Kukharicheva 2016] que utiliza SNRs de 0 a 45.

Para este experimento foram utilizadas nos treinos duas abordagens. Na primeira, a base de dados foi aumenta pelo agrupamento dos áudios de fala em diversas relações sinal-ruído, isso é SNR10, SNR30 e SNR$\infty$, porém o tipo de ruído é apenas o da proa da plataforma. Na abordagem seguinte, o aumento do conjunto de dados é realizado pelo tipo de ruído, sendo utilizados os áudios de fala com ruído tanto da proa, quanto do meio e da popa da plataforma, contudo foi considerado apenas o SNR de 10. Para cada abordagem foram utilizados 113.472 áudios nos treinos.

### 4.5. Aplicando Variação na Largura das Camadas Escondidas

Por último, foi avaliado o aumento do largura das camadas nos modelos, isso é, de unidades em cada camada escondida da arquitetura do *Deep Speech*. Para essa análise foram treinados 2 modelos para as larguras de 32, 64, 128, 256, 512, 1024, 2048 e 4096. Sendo o primeiro treinado com o ruído dos equipamentos localizados na proa da plataforma e SNR10, enquanto que o outro foi treinado sem ruído. Além de se avaliar o progresso do CER de acordo com a variação na largura das camadas, também é estudado o ganho relativo entre o treino com ruído e o sem ruído.

## 5. Resultados e Discussões

Nessa seção serão apresentados e avaliados os resultados dos experimentos de geração e teste dos modelos ASR com base nas metodologias descritas nas Seções 4.3, 4.4, e 4.5.

Os testes aplicados aos 7 modelos treinados de acordo com a metodologia definida na Seção 4.3, para análise do comportamento na variação do SNR, mostraram uma tendência de melhor valor do CER quando o teste é executado na mesma relação sinal-ruído do treino, conforme esperado. Por exemplo, dos 3 modelos treinados (proa, meio,

e popa) com SNR10, quando o tipo de ruído de treino e teste são iguais, por duas vezes o CER foi melhor para o teste com o SNR10, sendo que em uma vez se apresentou melhor no SNR30. Esta tendência fica mais evidente conforme a redução do ruído no áudio. Porém, quando o ruído de treino e teste do modelo são de ambientes distintos, o melhor valor do CER varia desde o teste na mesma relação sinal-ruído até relações mais altas, com menos ruído. [Prodeus and Kukharicheva 2016] obtiveram resultados similares na utilização de outros tipos de ruído e outra arquitetura para o idioma Russo.

A Figura 1, exibe uma visão comparativa da variação do CER, tanto com os 2 modelos treinados com áudios que apresentam a interferência do maquinário da proa, quanto com o modelo treinado sem ruído. Por fim, os modelos ASR treinados e testados com áudios que apresentam o mesmo tipo de ruído ou a ausência desses apresentaram um CER médio de 0,421984, sendo o SNR do treino igual ao do teste, contra 0,522851 de quando o SNR diverge. Enquanto que, os outros modelos resultaram no CER médio de 0,508566, sendo SNR do treino igual ao do teste, e 0,543441 quando diferentes.

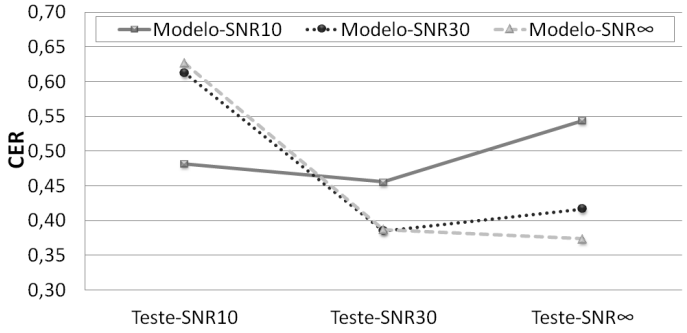


**Figura 1. Avaliação do CER dos modelos com a variação do SNR.**

O modelo treinado com aumentação de dados, devido à variedade da relação sinal-ruído, isso é, SNR10, SNR30 e SNR$\infty$, sendo o ruído o da proa, apresentou um melhor CER para todos os testes quando comparados com os treinados individualmente em cada relação sinal-ruído, conforme metodologia da Seção 4.3. A média do CER foi respectivamente de 0,422627 contra 0,497995.

No tocante ao modelo treinado com aumentação de dados devido às sobreposições dos áudios com os ruídos da proa, meio ou popa da plataforma, assim como no modelo anterior, os CERs apresentaram-se melhor na comparação com os testes equivalentes dos modelos treinados de forma individual para cada tipo de ruído. No entanto, o mesmo não ocorreu para o teste sem a presença de ruído, mas ainda assim, as médias do CER foram respectivamente de 0,446398 contra 0,497995.

Avaliando o impacto no CER com o aumento da largura das camadas, conforme metodologia detalha na Seção 4.5, no modelo treinado a partir de áudio com ruído em SNR10 ocorre a melhora do CER com o aumento do números de unidades, obtendo o melhor resultado nos testes com e sem ruído na largura de 512, sendo os valores apresentados para SNR10, SNR30 e SNR$\infty$, respectivamente, de 0,424374, 0,382383 e 0,460767.

Além disso, é possível identificar que nesse modelo o aumento na largura das camadas é mais benéfico ao reconhecimento de fala na presença de ruído do que em ambientes sem ruído, conforme Figura 2. Quanto avaliado o percentual de melhoria do CER

nos 4 primeiros incrementos da largura das camadas, isso é, de 32 até as 512 unidades. No teste com ruído em SNR30, em relação ao modelo imediatamente anterior o CER teve respectivamente a redução de 14,98%, 15,34%, 12,25%, e 7,81%, enquanto que nas falas sem ruído foi de 11,40%, 12,65%, 9,80%, e 5,53%.

Por fim, o modelo treinado sem ruído teve redução do CER em SNR30 até a largura definida com 512, em SNR10 e SNR$\infty$ o melhor resultado foi com 2048 unidades. Sendo os mínimos valores do CER de respectivamente de 0,608579, 0,318938 e 0,302572. Ademais, nesse modelo o aumento da largura na camadas escondidas apresentou-se mais benéfico para o teste sem ruído. Contudo, ambos os modelos apresentaram *overfitting* quando a largura atingiu o valor de 4096 unidades. [Xu et al. 2020] identifica o aprimoramento do modelo com o aumento da largura das camadas, contudo o experimento não identificou um limite.

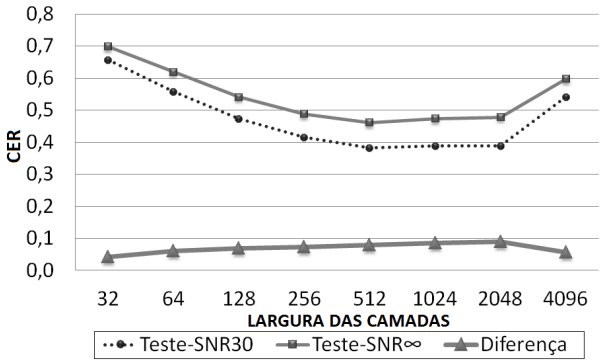


**Figura 2. Modelo SNR10 com aumento gradual da largura das camadas.**

## 6. Conclusões

O Reconhecimento Automático de Fala (ASR) tem se popularizado muito nos últimos anos, principalmente devido a todo avanço em redes neurais profundas, potencializado pela maior disponibilidade de GPUs. No entanto, o tema ainda é carente de estudos para a fala em ambientes ruidosos, principalmente, quando tais ruídos são peculiares, como é o estudo de caso desse trabalho que explora a interferência de ruídos provenientes do maquinário da indústria do petróleo.

Nesse trabalho foram avaliados diversos modelos de Reconhecimento Automático de Fala robusto ao ruído, com base no CER e seguindo três metodologias (seções 4.3, 4.4, e 4.5), tendo como objetivo o entendimento desses modelos conforme a variação na relação sinal-ruído e dos ruídos em si. Observou-se uma tendência de melhor transcrição da fala quando o nível de ruído nos testes segue a mesma relação SNR do treinamento do modelo. Os modelos treinados com aumento da base de dados apresentaram resultados superiores aos treinados apenas com uma relação SNR, enquanto que o aumento da largura das camadas foi favorável aos modelos robustos a ruídos, limitados a 512 unidades. A partir desse valor, tais modelos se apresentaram sem melhora e, por vezes, enviesados. Uma contribuição adicional do trabalho é a disponibilização da base de dados com os ruídos capturados e utilizados nos experimentos.

Uma direção para trabalhos futuros seria validar a configuração de outros hiperparâmetros do modelo no aprimoramento do reconhecimento de fala em ambientes com

ruído. Além disso, pretendemos avaliar a aplicação de filtros para redução do ruído. Primeiramente nos testes, e mais adiante, aplicá-lo também no treinamento dos modelos.

## Referências


Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., and Weber, G. (2019). Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., et al. (2014). Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.

Johnson, D. H. (2006). Signal-to-noise ratio. *Scholarpedia*, 1(12):2088.

Juang, B.-H. and Rabiner, L. R. (2005). Automatic speech recognition–a brief history of the technology development. *Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara*, 1:67.

Kinoshita, K., Ochiai, T., Delcroix, M., and Nakatani, T. (2020). Improving noise robust automatic speech recognition with single-channel time-domain enhancement network. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7009–7013. IEEE.

Mattys, S. L., Davis, M. H., Bradlow, A. R., and Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes*, 27(7-8):953–978.

Mošner, L., Wu, M., Raju, A., Parthasarathi, S. H. K., Kumatani, K., Sundaram, S., Maas, R., and Hoffmeister, B. (2019). Improving noise robustness of automatic speech recognition via parallel data and teacher-student learning. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6475–6479. IEEE.

Pervaiz, A., Hussain, F., Israr, H., Tahir, M. A., Raja, F. R., Baloch, N. K., Ishmanov, F., and Zikria, Y. B. (2020). Incorporating noise robustness in speech command recognition by noise augmentation of training data. *Sensors*, 20(8):2326.

Prodeus, A. and Kukharicheva, K. (2016). Training of automatic speech recognition system on noised speech. In *2016 4th International Conference on Methods and Systems of Navigation and Motion Control (MSNMC)*, pages 221–223. IEEE.

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.

Saon, G., Tüske, Z., Audhkhasi, K., and Kingsbury, B. (2019). Sequence noise injected training for end-to-end speech recognition. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6261–6265. IEEE.

Xu, J., Matta, K., Islam, S., and Nürnberger, A. (2020). German speech recognition system using deepspeech. In *Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval*, pages 102–106.