UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
CÂMPUS CORNÉLIO PROCÓPIO
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

CÉZAR FUMIO YAMAMURA

CONVERSÃO DE FALA SUSSURRADA PARA FALA NORMAL USANDO MODELOS NEURAIS

DISSERTAÇÃO

CORNÉLIO PROCÓPIO
2021

CEZAR FUMIO YAMAMURA

CONVERSÃO DE FALA SUSSURRADA PARA FALA NORMAL USANDO MODELOS NEURAIS

Conversion of whispered speech to normal speech using neural network models

Dissertação apresentada como requisito para obtenção do título de Mestre em Engenharia Elétrica, do Programa de Pós-Graduação em Engenharia Elétrica, da Universidade Tecnológica Federal do Paraná (UTFPR).

Orientador: Prof. Dr. Paulo Rogério Scalassara

CORNÉLIO PROCÓPIO

2021

Ministério da Educação
Universidade Tecnológica Federal do Paraná
Campus Cornélio Procópio

CEZAR FUMIO YAMAMURA

CONVERSÃO DE FALA SUSSURRADA PARA FALA NORMAL USANDO MODELOS NEURAIS

Trabalho de pesquisa de mestrado apresentado como requisito para obtenção do título de Mestre Em Engenharia Elétrica da Universidade Tecnológica Federal do Paraná (UTFPR). Área de concentração: Sistemas Eletrônicos Industriais.

Data de aprovação: 17 de Novembro de 2021

Prof Paulo Rogerio Scalassara, Doutorado - Universidade Tecnológica Federal do Paraná

Prof Danilo Hernane Spatti, Doutorado - Usp-Universidade de São Paulo

Prof Sylvio Barbon Junior, Doutorado - Universidade Estadual de Londrina (Uel)

Documento gerado pelo Sistema Acadêmico da UTFPR a partir dos dados da Ata de Defesa em 19/11/2021.

RESUMO

YAMAMURA, Cézar Fumio.. **Conversão de fala sussurrada para fala normal usando modelos neurais**. 2021. 51 f. Dissertação – Programa de Pós-Graduação em Engenharia Elétrica, Universidade Tecnológica Federal do Paraná. Cornélio Procópio, 2021.

O sussurro é um mecanismo de fala comum e secundário para se comunicar, porém, em alguns casos, pode ser a principal forma de diálogo, como nos casos de pessoas com patologias da laringe ou que sofreram algum tipo de dano nas pregas vocais. As diferenças de características entre a voz normal e a sussurrada têm levantado discussões na área de reconhecimento de fala, pela dificuldade de realizar a conversão de fala sussurrada para fala vozeada. Este trabalho apresenta o estudo das características da fala normal e sussurrada, e a implementação do sistema de conversão de voz normal para sussurrada utilizando redes perceptron multicamadas e redes generativas adversárias. Os dados utilizados foram provenientes pela parceria com Faculdade de Porto, que são sinais de vogais usados no português europeu. Para a validação do estudo, foram analisado três métricas: *Mel-Cepstrum Distortion*, raiz quadrada do erro médio do $log(F_0)$ e acurácia do classificador de vogais.

Palavras-chave: Reconhecimento de fala, Redes perceptron multicamadas, Redes generativas adversárias

ABSTRACT

YAMAMURA, Cézar Fumio. **Conversion of whispered speech to normal speech using neural network models.** 2021. 51 f. Master Thesis – Electrical Engineering Graduate Program, Federal University of Technology - Paraná. Cornélio Procópio, 2021.

Whispering is a common and secondary speech mechanism for communicating, however, in some cases, it can also be the main form to communicate, such as cases of people with pathologies of the larynx or who have suffered some type of damage to the vocal folds. Differences in characteristics between normal voice and whispered discussions have raised in speech recognition area, due to the difficulty of converting whispered to normal speech. This work presents the study of the characteristics of normal and whispered speech, and the implementation of the whispered to normal voice conversion system using multilayer perceptron networks and generative adversarial networks. The database used came from the partnership with Faculdade de Porto, which are vowel signs used in European Portuguese. To validate the study, three metrics were analyzed: Mel-Cepstrum Distortion, square root of the mean error of $log(F_0)$ and accuracy of the vowel classifier.

Keywords: Speech recognition, Multilayer perceptron networks, Generative adversarial network

LISTA DE FIGURAS

FIGURA 1 – Modelo comunicacão humana. 11
FIGURA 2 – O aparelho fonador humano. 12
FIGURA 3 – Escala de percepção Mel x escala em frequFência. 15
FIGURA 4 – Modelo da produção de fala. 16
FIGURA 5 – Comparação do espectro da vogal /a/ entre fala normal e a fala sussurrada. 17
FIGURA 6 – Ilustração do modelo perceptron. 19
FIGURA 7 – Ilustração do modelo perceptron. 19
FIGURA 8 – Função de ativação ReLU. 20
FIGURA 9 – Ilustração da rede PMC. 20
FIGURA 10 – Arquitetura de uma rede GAN. 21
FIGURA 11 – rede GAN para reconstrução de voz sussurada. 22
FIGURA 12 – Arquitetura da rede DiscoGAN 24
FIGURA 13 – Diagrama de blocos da conversão de voz normal para voz sussurrada. . . 26
FIGURA 14 – Formas de onda de um sinal de fala normal e sussurrada. 27
FIGURA 15 – Mel Espectrograma e frequência fundamental de um sinal de fala normal e sussurrada. 28
FIGURA 16 – Diagrama de blocos da rede (a) Conversão MFCCs para MFCCc e (b) Conversão MFCCc para F_0c. 29
FIGURA 17 – Diagrama de blocos do classificador de vogais da voz sussurrada. 31
FIGURA 18 – Frequêcia Fundamental entre homem e mulher. 33
FIGURA 19 – (a) $log(F_0)$ da voz normal correspondente e estimativa do $log(F_0)$ usando o (b) MLP, (c) GAN, (d) DiscoGAN, dos sinal masculino e feminino. . . 35
FIGURA 20 – Matriz de confusão utilizando banco de dados 1 - Vogal /a/ e silêncio . . 36
FIGURA 21 – Matriz de confusão para agrupamento 4 sem silêncio usando GAN. . . . 37
FIGURA 22 – Matriz de confusão utilizando agrupamento 1 com silêncio. 43
FIGURA 23 – Matriz de confusão utilizando agrupamento 2 com silêncio. 44
FIGURA 24 – Matriz de confusão utilizando agrupamento 3 com silêncio. 45
FIGURA 25 – Matriz de confusão utilizando agrupamento 4 com silêncio. 46
FIGURA 26 – Matriz de confusão utilizando agrupamento 2 sem silêncio. 47
FIGURA 27 – Matriz de confusão utilizando agrupamento 3 sem silêncio. 48
FIGURA 28 – Matriz de confusão utilizando agrupamento 4 sem silêncio. 49

LISTA DE TABELAS

TABELA 1 – As nove vogais orais do Português Europeu padrão disponíveis na base de dados DyNaVoiceR, no modo sustentado. 25
TABELA 2 – Topologia usada nas redes MLP e GAN. 29
TABELA 3 – Divisão do banco de dados. 32
TABELA 4 – MCD do banco de dados com silêncio. 33
TABELA 5 – MCD do banco de dados sem silêncio. 33
TABELA 6 – RMSE do $log(F_0)$ do banco de dados com silêncio. 34
TABELA 7 – RMSE do banco de dados sem silêncio. 34
TABELA 8 – Acurácia do banco de dados com silêncio. 37
TABELA 9 – Acurácia do banco de dados sem silêncio. 37

SUMÁRIO

1 INTRODUÇÃO 9
1.1 OBJETIVOS 10
1.1.1 Objetivos Específicos 10
1.1.2 Organização do Texto 10
2 FUNDAMENTAÇÃO TEÓRICA 11
2.1 O PROCESSO DE PRODUÇÃO E PERCEPÇÃO DA FALA 11
2.2 ANATOMIA DA FALA 11
2.3 ANÁLISE DE SINAIS DE FALA 13
2.3.1 Formantes 13
2.3.2 Coeficientes cepstrais em escala de frequências Mel 14
2.3.3 Codificação Preditiva Linear 15
2.4 FALA SUSSURRADA 16
2.5 TÉCNICAS DE RECONHECIMENTO DE FALA 17
2.5.1 Gaussian Mixed Models 18
2.5.2 Hidden Markov Models 18
2.5.3 Redes Neurais Artificiais 18
2.5.4 Redes Perceptron Multicamadas 20
2.5.5 Redes Generativas Adversárias 21
3 MATERIAIS E MÉTODOS 25
3.1 BANCO DE DADOS 25
3.2 METODOLOGIA 26
3.2.1 Pré-processamento 26
3.2.2 Rede de conversão 27
3.2.3 Aplicações de métricas de desempenho 30
4 RESULTADOS E DISCUSSÕES 32
4.1 ANÁLISE DO BANCO DE DADOS 32
4.2 ANÁLISE MCD 32
4.3 ANÁLISE RMSE 33
4.4 CLASSIFICADOR DE VOGAIS 35
5 CONCLUSÃO 38
REFERÊNCIAS 39
ANEXO A – REPRODUÇÃO DAS TAREFAS DE GRAVAÇÃO (DYNAVOICER) 51

LISTA DE ABREVIATURAS E SIGLAS

ASR	Reconhecimento Automático de Fala
MLP	Redes Perceptron Multicamadas
GAN	*Generative adversarial network*
F_0	Frequência Fundamental
MFCC	*Mel-Frequency Cepstral Coefficients*
FFT	*Fast Fourier Transform*
DCT	*Discrete Cosine Transform*
LPC	*Linear Predictive Coding*
MELP	*Mixed-Excited Linear Prediction*
CELP	*Code Excited Linear Prediction*
GMM	*Gaussian Mixed Models*
HMM	*Hidden Markov Models*
RNA	Redes Neurais Artificiais
SEGAN	*Speech Enhancement GAN*
DyNaVoiceR	*Dysphonic to Natural Voice Reconstruction*
LPSA	Laboratório de Processamento de sinais e Aplicações
FEUP	Faculdade de Engenharia da Universidade de Porto
MCD	*Mel-Cepstrum Distortion*
RMSE	Raiz Quadrada do Erro Médio
DTW	*Dynamic Time Warping*

1 INTRODUÇÃO

A voz humana é um instrumento muito importante para a comunicação, capaz de transmitir nossos desejos, informações, ideias entre outros, o que nos faz sentir parte da sociedade. O sussurro é uma forma alternativa de comunicação, geralmente utilizado a fim de manter a privacidade ou conversar em locais silenciosos como bibliotecas, hospitais ou sala de reunião.

Infelizmente, em algumas situações, não se trata de uma alternativa, mas sim a única forma de se dialogar, como no caso de pessoas com patologias da laringe ou que sofreram algum tipo de dano nas pregas vocais (SHARIFZADEH et al., 2017). Consequentemente, a fala sussurrada despertou interesse em pesquisadores, principalmente em aprimoramento do sistema de reconhecimento automático de fala (ASR - *automatic speech recognition*), pois as características acústicas da fala sussurrada, como falta de excitação periódica das pregas vocais e menor energia nas baixas frequências em relação a vozes normais, dificulta o reconhecimento pelo sistema ASR (GROZDIĆ; JOVIČIĆ, 2017; ITOH; TAKEDA; ITAKURA, 2001; JOVIČIĆ; ŠARIĆ, 2008).

Existem duas abordagens básicas para garantir que ASR possa operar com sussurros: a primeira é modificar o sistema para trabalhar diretamente com a entrada de sinais de voz sussurrada (ASR dedicado a voz sussurrada). A segunda é converter o sussurro em um sinal semelhante a fala normal antes de processar normalmente no ASR. A ultima abordagem é comumente estudada, pois o usuário geralmente se adapta a sua forma de comunicação conforme a necessidade dele (LI; MCLOUGHLIN; SONG, 2014).

Diversos métodos são usados para tentar reconstruir a voz nesses casos, entre eles, pode-se citar a busca por parâmetros do modelo de produção vocal fonte-filtro para adicionar o vozeamento perdido pelo sussurro, como apresentado por Ferreira (2016). Outros trabalhos utilizam estimativas de modelo preditivo linear dos sinais, como Morris e Clements (2002), ou em conjunto com coeficientes mel-cepstrais e treinamento de modelos Gaussianos mistos (SHARIFZADEH et al., 2017).

Porém, nos últimos anos, métodos de aprendizagem de máquina também vêm sendo usados com muito êxito, como é o caso das redes perceptron multicamadas (MLP) (HINTON et al., 2012; GHAFFARZADEGAN; BORIL; HANSEN, 2017). Recentemente, dentro das diversas topologias utilizando a MLP, as redes generativas adversárias (GAN) estão ganhando popularidade na área de reconstrução de voz, trazendo uma melhora significativa em desempenho e qualidade em aplicações como conversão de voz (KANEKO et al., 2017) e em aprimoramento de voz (PASCUAL; SERRA; BONAFONTE, 2019).

As GAN, proposta por Goodfellow et al. (2014), são caracterizados por duas redes neurais: geradoras e discriminadora, as quais competem entre si para melhorar suas técnicas. No contexto deste trabalho, a aplicação da rede GAN tem como objetivo o mapeamento das características da voz para converter o sussurro em um sinal semelhante a fala normal.

A fim de implementar o sistema de conversão de fala sussurrada para normal, foi

utilizado banco de dados construído no projeto *Dysphonic to Natural Voice Reconstruction* (DyNaVoicer) pela Faculdade de Engenharia da Universidade do Porto (FEUP), no qual foi realizado vários estudos como a modelização de filtro de trato vocal para reconstrução de voz disfónica (OLIVEIRA, 2020) e segmentação de fonética adaptativa da voz sussurrada (COSTA, 2021).

1.1 OBJETIVOS

O objetivo deste trabalho é analisar as características entre os sinais de voz normal e sussurrada, a fim de reconstituir o sinal utilizando o GAN. Para isso, será utilizado o banco de dados em português europeu com nove vogais orais e comparar os resultados do GAN com o DNN.

1.1.1 Objetivos Específicos

Com base no objetivo geral, define-se os seguintes objetivos específicos:

- Estudar as diferenças de características entre a voz normal e a sussurrada;
- Estudar o banco de dados utilizado neste trabalho e o pré-processamento destes dados para adequar as amostras para as redes de conversão de voz sussurrada para normal;
- Implementar a rede de conversão de voz a partir das redes MLP e GAN;

1.1.2 Organização do Texto

Este trabalho está organizado em cinco capítulos, trazendo discussões sobre o problema e os métodos aplicados para a sua solução, bem como a sua conclusão.

Neste Capítulo 1 foram introduzida o tema e a motivação da escolha na realização deste trabalho. Também são apresentados os objetivos gerais e específicos deste trabalho.

No Capítulo 2, é apresentada a fundamentação teórica, abordando os principais conceitos sobre os sinais de voz normal e sussurrada e as técnicas de reconhecimento de fala.

No Capítulo 3, são descritos os materiais e métodos, apresentado por meio de um fluxograma todas as etapas do trabalho, que é iniciado pela análise de banco de dados, pré-processamento de dados e extração de características, implementação das redes de conversão de voz e por fim, as métricas para a validação da metodologia.

No Capítulo 4 são descritos os resultados obtidos e também é realizada a discussão e a análise dos mesmos. Por fim, no Capítulo 5 são apresentadas as conclusões do trabalho e propostas de ações futuras.

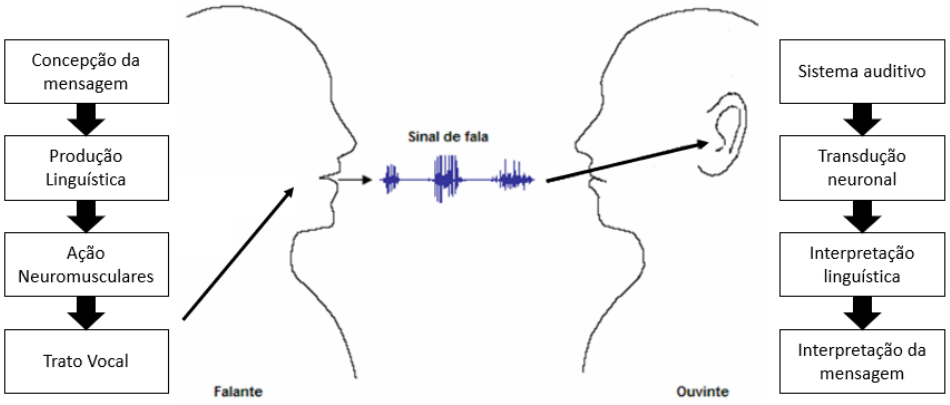
2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo serão abordados conceitos teóricos sobre os sinais de voz, desde o seu processo de produção e percepção, a anatomia até a análise dos sinais de voz normal e sussurrada.

2.1 O PROCESSO DE PRODUÇÃO E PERCEPÇÃO DA FALA

Na Figura 1 é apresentado um modelo do processo de comunicação humana através da fala. O processo de produção da fala começa quando o orador formula conceptualmente a mensagem, que pretende transmitir através de fala ao seu ouvinte. O próximo passo é o processo de conversão da mensagem em código de linguagem, no qual a mensagem é transformada num conjunto de sons e movimentos faciais que correspondem às palavras da mesma. Para produzir esses sons, é desencadeada uma série de ações neuromusculares que irão comandar diversos órgãos chamado de trato vocal (RABINER; JUANG, 1993).

Figura 1 – Modelo comunicacão humana.



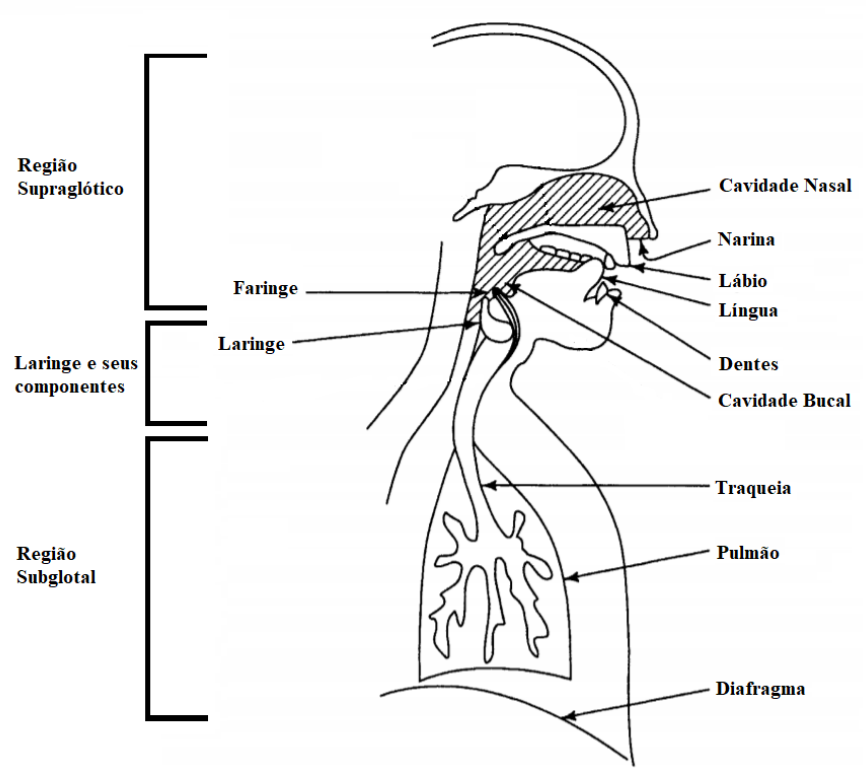
Fonte: Adaptado de Rabiner e Juang (1993)

Uma vez que o sinal de fala é gerado e propagado para o ouvinte, inicia se o processo de reconhecimento de fala. O sinal acústico é processado pela membrana basilar, no ouvido interno, que realiza uma forma de análise espectral do sinal. O espectro do sinal de vibração gerado pela membrana basilar é convertido em sinais de ativação no nervo auditivo, através de um processo de transdução neuronal. Da atividade neural, envolvendo o nervo auditivo e o córtex cerebral, surge neste um código da linguagem, que é compreendido e interpretado pelo cérebro (RABINER; JUANG, 1993).

2.2 ANATOMIA DA FALA

O sinal da fala é originado pelos movimentos voluntários de estruturas anatômicas que compõem o sistema de produção de fala humana. A Figura 2 apresenta o aparelho fonador, que é dividido em três partes: sistema subglotal (sistema abaixo da laringe), laringe e os componentes a sua volta e o sistema supraglótico (sistema acima da laringe) (STEVENS, 2000).

Figura 2 – O aparelho fonador humano.



Fonte: Adaptado de (DELLER; PROAKIS; HANSEN, 1993)

No sistema subglotal, encontram-se a traqueia, pulmões e o diafragma, responsáveis pelo suprimento da fonte de energia que gera os sons da fala. O diafragma constitui-se em uma estrutura em forma de abóbada que separa a cavidade torácica da abdominal. A traqueia é um tubo de estrutura fibrocartilaginosa que vai da cavidade torácica à laringe (SEARA; NUNES; VOLCÃO, 2011).

A Laringe é o orgão que constitui a principal fonte de conteúdo sonoro utilizado na fala. É composto em seu interior pelas pregas vocais, formadas por cartilagens aritenoide, localizadas nas paredes superiores da laringe. A glote representa a abertura entre essas duas pregas. Na respiração normal, as cartilagens estão distantes, assim o fluxo de ar passa livremente. Quando falamos, as pregas vocais vibram, abrindo e fechando rapidamente a passagem ao fluxo de ar vindo dos pulmões (STEVENS, 2000).

Dependendo do modo de excitação das pregas, geram-se tipos diferentes de sons, os quais podem ser classificados principalmente nas classes: vocálicos, fricativos, plosivos e africados. Os sons vocálicos são produzidos quando se força o ar através da glote com a tensão das pregas vocais ajustada para vibrar como uma oscilação relaxada, assim, produzem-se pulsos quase periódicos de ar no trato vocal. Como quase-periódicos, entende-se uma sequência de pulsos com variação aleatória de seus períodos de repetição (RABINER; SCHAFER, 2010).

Por outro lado, os sons fricativos são formados por uma constrição em algum ponto do trato vocal, normalmente na boca, fazendo o ar dos pulmões passar com velocidade suficiente para gerar turbulência, então, cria-se uma fonte de ruído de banda larga no trato. Já os sons

plosivos são resultantes de um fechamento do trato, geralmente na boca, criando um aumento de pressão, o qual gera o som quando liberado. Por fim, os sons africados são a combinação de uma plosiva e uma fricativa (RABINER; SCHAFER, 2010).

O sistema supraglótico age como um modulador de som, pois se comporta como um ressonador, composto pela faringe e trato vocal. O trato vocal é dividido em cavidade oral e nasal, responsável pelos sons orais e nasais. Este é o sistema responsável pela formação do timbre, formando os vogais e consoantes, através dos movimentos articulatórios. Assim, as diferentes partes do trato vocal são chamadas de articuladores e são constituídas pelos articuladores ativos, como lábios, palato mole e mandíbula, e os articuladores passivos, como dentes, crista alveolar e palato duro (SEARA; NUNES; VOLCÃO, 2011).

2.3 ANÁLISE DE SINAIS DE FALA

As ondas sonoras da fala podem ser periódicas, como no caso das vogais, aperiódicas contínuas, como fricativas consoantes, ou aperiódicas não-contínuas, como no consoantes plosivas. Além disso, os sinais de voz são geralmente classificados em sons vozeados e não-vozeados (OLIVEIRA, 2015).

- Sons vozeados: são caracterizados pela frequência fundamental (F_0), que é a menor frequência, e seus componentes harmônicos produzidos pelas pregas vocais. Elas geram ondas sonoras periódicas com valores de frequência fundamental entre 50 e $500Hz$. Além disso, o sinal é caracterizado por suas frequências de formantes e antiformantes, causados pela modificação do sinal de excitação pelo trato vocal. Cada frequência de formante tem também uma amplitude e largura de banda que deve ser considerada.
- Sons não-vozeados: Não há uma frequência fundamental no sinal de excitação e, portanto, não há harmônicos. Além disso, o sinal de excitação não é periódico e se assemelha ao ruído branco. Alguns sons são caracterizados pela interrupção do fluxo de ar seguida por uma liberação repentina no trato vocal.

A frequência fundamental e formantes são provavelmente os conceitos mais importantes na síntese de voz, portanto, iremos abordá-los em detalhes, juntamente com outros conceitos importantes que serão utilizados ao longo do texto.

2.3.1 Formantes

Formantes representam as ressonâncias acústicas do trato vocal. Cada área do trato vocal tem sua própria frequência de ressonância, portanto, a amplitude e os harmônicos do sinal são modificados durante o passagem pelas cavidades supraglóticas. Formantes são usualmente medidos como picos de amplitude na envoltória do magnitude do espectro de frequência do som. O posicionamento das formantes é essencial para distinguir em particular as diferentes vogais,

que são produzidas sem que o trato vocal ofereça uma obstrução significativa à passagem do ar. As formantes que mais contribuem para definir e distinguir cada uma das vogais são as duas primeiras (F1 e F2) e a terceira (F3) em menor grau. As consoantes, por sua vez, caracterizam-se principalmente pela introdução de obstruções temporárias à passagem do ar, parciais ou totais, através do controle e posicionamento dos diversos articuladores, que são a língua, os dentes e os lábios (RABINER; JUANG, 1993).

2.3.2 Coeficientes cepstrais em escala de frequências Mel

Para determinar a frequência fundamental, usualmente é utilizada o análise cepstral. É um método para separar as informações do trato vocal da excitação, pois permite a conversão dos sinais obtidos por convolução (como fonte e filtro) em somas de seus cepstros, proporcionando assim, uma separação linear (RABINER; SCHAFER, 2010).

Na análise cepstral, o espectro é geralmente transformado usando a escala de Mel, que se baseia na escala do ouvido humano, resultando em um *Mel-frequency cepstrum* (MFC), cujos coeficientes são denominados *Mel-Frequency Cepstral Coefficients* (MFCCs) (TYCHTL; PSUTKA, 1999). Os MFCC são representações da parte real do cepstro de um janelamento em curto período de tempo de sinais acústicos, derivados de uma Transformada Rápida de Fourier (FFT) de um sinal. A diferença destes coeficientes com coeficientes cepstrais é a utilização da escala logarítmica, aproximando os coeficientes ao comportamento do aparelho auditivo humano (MOLLA; HIROSE, 2004). A Figura 3 ilustra a escala Mel em relação a escala de frequência. A sua obtenção pode ser feita por meio da Equação 1:

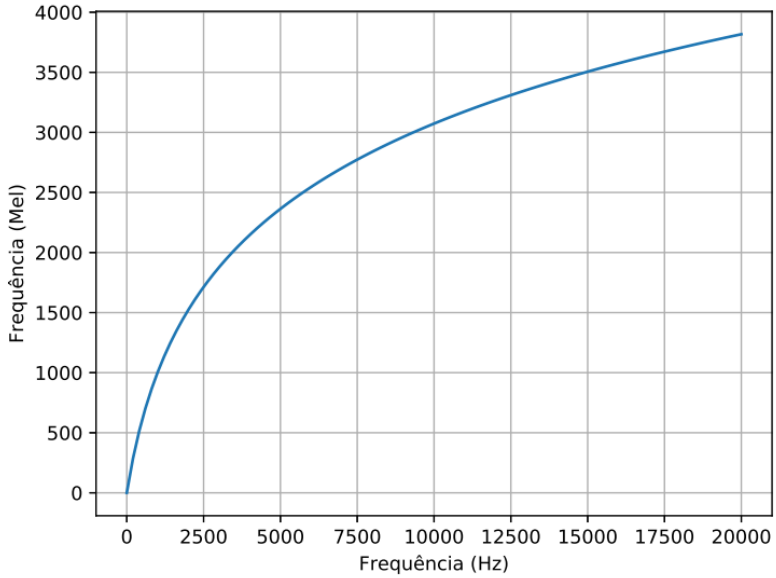
$$f_{mel} = 2595 log_{10}(1 + \frac{f}{700}) \tag{1}$$

Esta técnica, além de extrair as características úteis de um sinal de voz para posterior criação dos vetores acústicos, também reduz a quantidade de informações não essenciais para ASR, retirando partes não-vozeada, além de reduzir efeitos de reverberação. Para calcular tais coeficientes, deve-se seguir alguns passos sequenciais, dentre eles o enquadramento e janelamento do sinal, utilização dos dados no domínio da frequência por meio da FFT, aplicação da escala logarítmica (Mel) para construção de um conjunto de coeficientes cepstrais e, finalmente, utilização de uma transformada cosseno discreta (DCT) sobre os dados adquiridos no passo anterior apresentada pela Equação 2 (GOPI, 2013).

$$c_i[n] = \sum_{m=0}^{M-1} 20 log_{10}(S_i[m]) cos\left(\pi n \left(\frac{m + 0,5}{M}\right)\right) \tag{2}$$

onde M é o número de coeficientes cepstrais desejados.

Figura 3 – Escala de percepção Mel x escala em frequFência.



Fonte: Autoria Própria

2.3.3 Codificação Preditiva Linear

O codificação preditiva linear (LPC) é uma ferramenta extratora de propriedades acústicas de sinais de fala que utiliza modelos probabilísticos. Esta técnica é usualmente aplicada à síntese ou reconstrução de sinais, pois a sua implementação computacional é simples e eficiente, permitindo alcançar uma boa qualidade em taxa de "bitrate"na codificação/compressão de sinais de voz (DELLER; PROAKIS; HANSEN, 1993).

O modelo da produção de fala está ilustrada na Figura 4. Na codificação LPC, o sistema de produção de fala é modelado da seguinte maneira: para sinais de fala do tipo vozeado, excitamos um filtro que representa o modelo do trato vocal com pulsos glotais espaçados por um período de "*pitch*"e multiplicamos por um ganho. Para sinais do tipo não vozeados, excitamos este mesmo filtro com ruído branco multiplicado por um ganho. Na saída deste filtro, obtém o sinal da fala. (RABINER; SCHAFER, 2010).

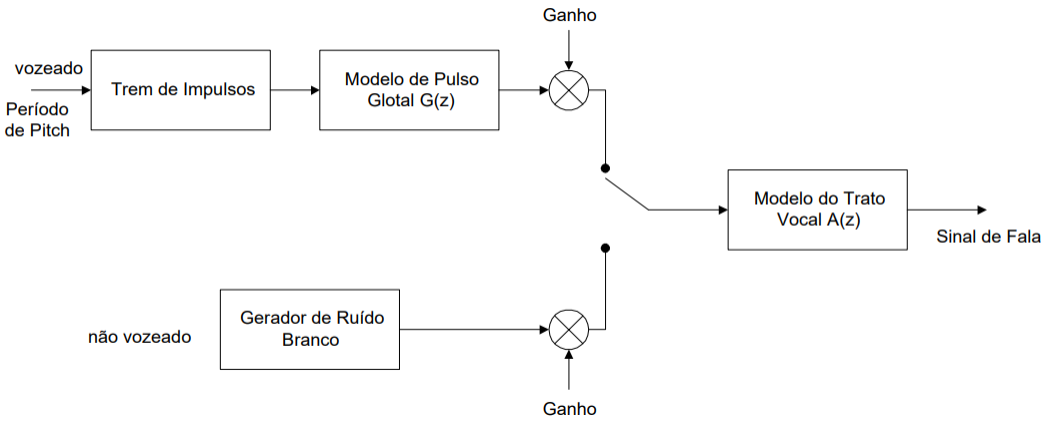
O modelo de predição linear assume que o sinal de fala é um processo auto-regressivo descrito pela Equação 3:

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) \tag{3}$$

onde $s(n)$ é o sinal de fala, a_i são os coeficientes de predição e p é a ordem do preditor.

Assim, na análise LPC, é necessário encontrar o valor dos coeficientes a_i a fim de minimizar a função de custo do erro e(n) descrito pela Equação 4. Para isso é utilizado o método da autocorrelação ou o método da covariância.

Figura 4 – Modelo da produção de fala.



Fonte: Adaptado do Rabiner e Schafer (2010)

$$e(n) = s(n) - \sum_{i=1}^{p} a_i s(n-i) \tag{4}$$

E como os sinais de fala não são estacionários, trabalha-se com segmentos do sinal considerados aproximadamente estacionários, e atualiza-se os coeficientes do modelo do trato vocal para cada um desses segmentos.

2.4 FALA SUSSURRADA

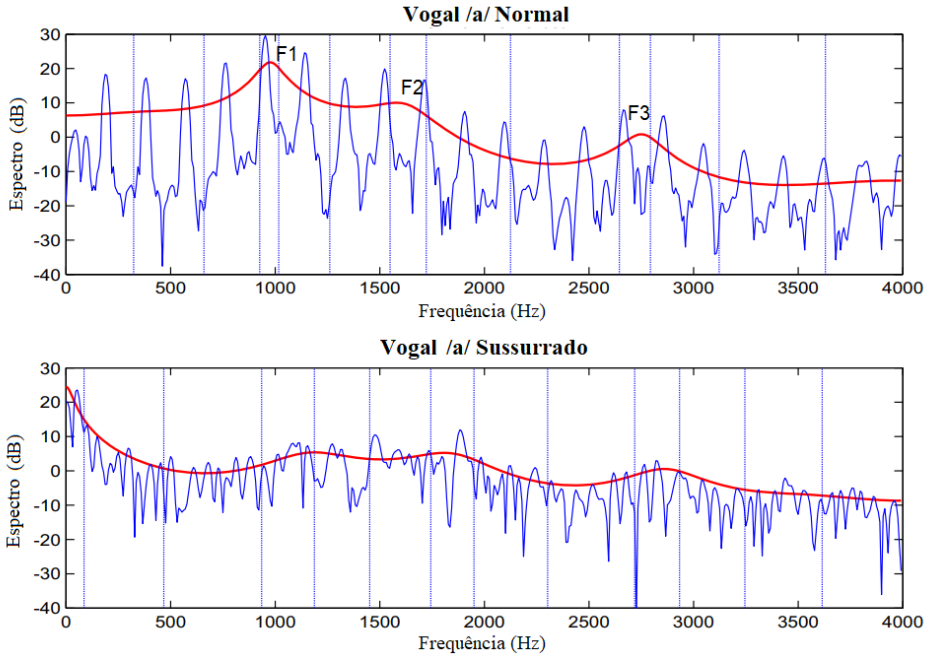
Fisiologicamente falando, a principal diferença entre a fala normal e a fala sussurrada é a falta de vibração das pregas vocais, o que causa ausência da frequência fundamental e suas harmônicas, apesar de continuar existindo a sensação de tonalidade (MORRIS; CLEMENTS, 2002).

Geralmente, a fala sussurrada possui menos de $20dB$ de potência espectral em relação a fala normal. Além disso, os formantes possuem uma frequência maior, principalmente no primeiro formante (MORRIS; CLEMENTS, 2002). A Figura 5 exemplifica essa diferença entre os harmônicos das duas falas. O sinal em azul é a representação de um segmento de sinal de voz no domínio das frequências e em vermelho é uma estimativa da envolvente espectral, sendo possível identificar as principais frequências formantes.

No espectro das vogais da fala sussurrada, observam-se os seguintes aspectos: há maior deslocamento entre os formantes comparado com a fala normal, quanto menor a frequência dos formantes, maior será o valor do deslocamento; não há uma estrutura periódica clara, mas a estrutura do formante ainda existe e difícil de identificar as principais frequências formantes. (ITOH; TAKEDA; ITAKURA, 2001).

Devido a essas diferenças, técnicas de conversão de voz sussurro para fala normal utilizando o LPC como modelo MELP (*Mixed-Excited Linear Prediction*) e CELP (*Code Excited Linear Prediction*), trabalhos desenvolvidos por, respectivamante, Morris e Clements (2002) e Sharifzadeh, McLoughlin e Ahmadi (2010).

Figura 5 – Comparação do espectro da vogal /a/ entre fala normal e a fala sussurrada.



Fonte: Adaptado de (SHARIFZADEH; MCLOUGHLIN; RUSSELL, 2012)

O MELP é um codificador que permite manipular separadamente o sinal de excitação da fonte e o modelo de filtro, dado por um LPC de 10ª ordem. Neste caso, a injeção de uma onda periódica no sinal de excitação da fonte permite adicionar vozeamento enquanto a implementação de um filtro de fase mínima, para efeitos de compensação das diferenças espectrais entre o discurso vozeado e não vozeado, modifica o filtro LPC supramencionado. Ainda, segundo os autores, para permitir o controle do "*pitch*"em tempo-real por parte do utilizador, o período deste é controlado via intensidade sonora. Porém, a ausência de segmentação é uma desvantagem deste codificador, o que significa que a banda de 0-3 kHz é permanente e forçosamente vozeada enquanto que a banda de 3-4 kHz é tratada como não vozeada (OLIVEIRA, 2020).

Já o CELP é um codificador que implementa a modificação das formantes e a modulação glótica artificial sem requerer informação a priori que dependa do orador. Isso quer dizer que dada a natureza ruidosa do sussurro, o codificador estima primeiramente a frequência e amplitude de até quatro formantes, de forma robusta, evitando formantes não genuínas. Estas formantes assim estimadas são modificadas tendo em consideração as diferenças características entre as correspondentes versões vozeadas e não vozeadas, de forma análoga ao indicado na abordagem anterior. Porém solução baseada no CELP é inadequada para fala contínua, pois exige um esforço para gerar frases completas. Na próxima seção serão apresentadas outras técnicas de conversão de voz.

2.5 TÉCNICAS DE RECONHECIMENTO DE FALA

Nesta seção serão apresentadas as principais técnicas utilizadas para reconhecimento de fala, como *Gaussian mixed models*, *Hidden Markov models* e Redes Neurais Artificiais.

2.5.1 Gaussian Mixed Models

Gaussian Mixed Models (GMM) são modelos estatísticos que assumem que uma população de dados pertence a uma distribuição em forma de mistura, ou seja, uma distribuição formada por combinações lineares de diversas funções de probabilidade (GOPI, 2013).

GMM são largamente utilizados para estimar funções de densidade de probabilidade desconhecidas, onde o objetivo é estimar os pesos de cada gaussiana e suas respectivas médias e matrizes de covariância a partir dos dados observados (GOPI, 2013).

Em trabalhos como Toda, Nakagiri e Shikano (2012) e Sharifzadeh et al. (2017), utilizou-se a conversão de voz sussurrada para normal baseado em GMM, no qual obteve resultados melhores em termos de qualidade sonora comparado ao sistema baseado em CELP. Porém, segundo Liu et al. (2018), como o GMM usa os parâmetros estatísticos segmentados do espectro da fala para identificá-lo, é difícil obter bons resultados quando utilizados com frases curtas.

2.5.2 Hidden Markov Models

Hidden Markov Models (HMM) é um processo duplamente estocástico, com um processo estocástico subjacente que não é diretamente observável, ou seja, é escondido o qual somente pode ser observado através de outro processo estocástico que produz a sequência de observações (RABINER; SCHAFER, 2010).

Aplicada em reconhecimento de padrões, eles combinam informações acústicas e determinam quais frequências estão presentes em qual instante de tempo para calcular qual a palavra mais provável que a pessoa está falando. Eles também levam em conta outros aspectos, como informações linguísticas e sintaxe para determinar quais sequências de palavras são as mais prováveis.

Em Itoh, Takeda e Itakura (2001), para comparar os espectros de um fonema, foi utilizado o modelo HMM, treinado separadamente com os dados de fala normal e fala sussurrada, na qual obteve uma precisão geral do fonema dada pelo modelo de fala normal de 52,2%, enquanto os modelos treinados com dados de fala sussurrada produziram uma precisão de 75,9%.

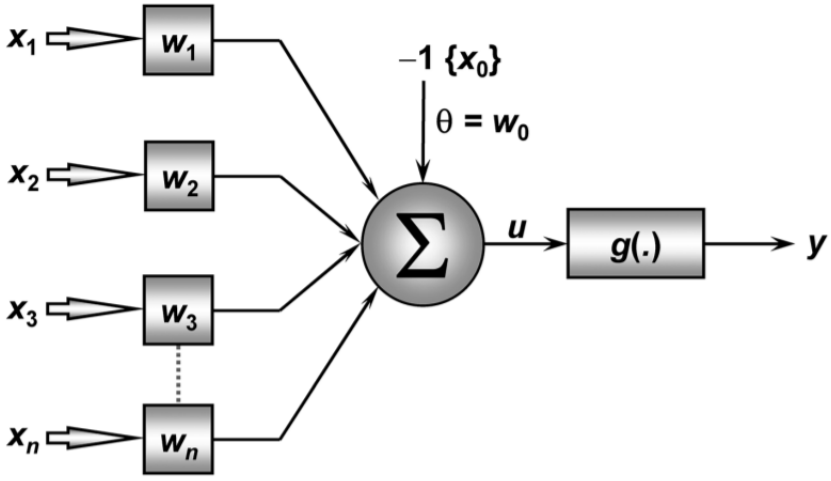
2.5.3 Redes Neurais Artificiais

A introdução das redes neurais artificiais (RNA) em reconhecimento de fala foi um grande avanço na pesquisa nas últimas quatro décadas, no qual predominava apenas os métodos GMM e HMM (HINTON et al., 2012). As RNA foi inspirada nos neurônios biológicos e como esses estão estruturados no cérebro, tendo em vista que o grande volume aumenta a capacidade de processamento. Essa ideia iniciou com um modelo computacional para representar um neurônio, proposto por McCulloch e Pitts (1943). Depois Hebb (1949) propôs um modelo de aprendizado e por último veio o modelo perceptron, proposto por Rosemblatt (1957), que a partir de uma entrada com um determinado peso, uma regra de propagação e uma função de ativação obtêm-se

uma saída. Com esses 3 estudos foi possível simular, mesmo que basicamente, o funcionamento de um neurônio (SILVA; SPATTI; FLAUZINO, 2016).

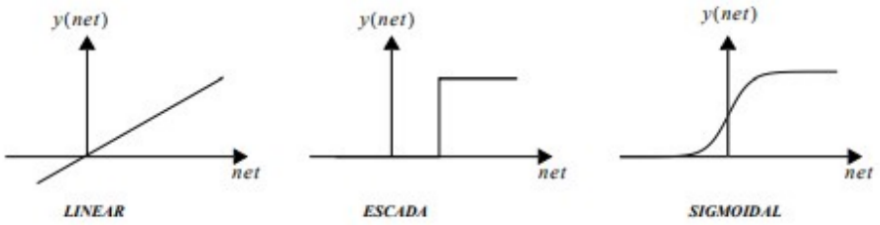
O modelo de neurônio está representada na Figura 6. E a função de ativação dita o comportamento do neurônio é exemplificada na Figura 7.

Figura 6 – Ilustração do modelo perceptron.



Fonte: Adaptado de (SILVA; SPATTI; FLAUZINO, 2016)

Figura 7 – Ilustração do modelo perceptron.



Fonte: Adaptado de (SILVA; SPATTI; FLAUZINO, 2016)

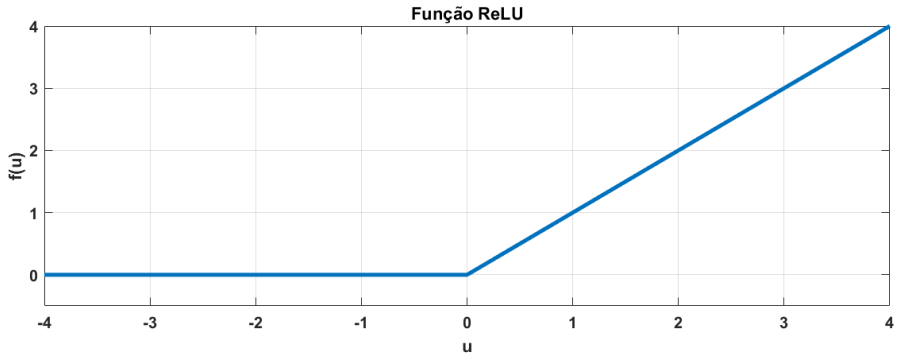
Além dessas funções apresentadas acima, uma função importante e que merece destaque é a ReLu. Segundo Krizhevsky, Sutskever e Hinton (2017) a função de ativação ReLU apresenta um desempenho melhor nas redes profundas devido a sua simplicidade computacional, por não ser necessário o cálculo do valor exponencial, e por ter uma derivada constante, facilitando o treinamento do algoritmo. A Equação (5) apresenta a função ReLU, cujo gráfico está ilustrado na Figura 8,

$$\begin{cases} f(u) = u, & se \quad > 0 \\ f(u) = 0, & se \ \ u \leq 0 \end{cases} \tag{5}$$

onde $f(u)$ é a função de ativação e u é o limiar de ativação.

Neste estudo serão usadas duas arquiteturas de redes neurais: redes perceptron multicamadas e GAN.

Figura 8 – Função de ativação ReLU.



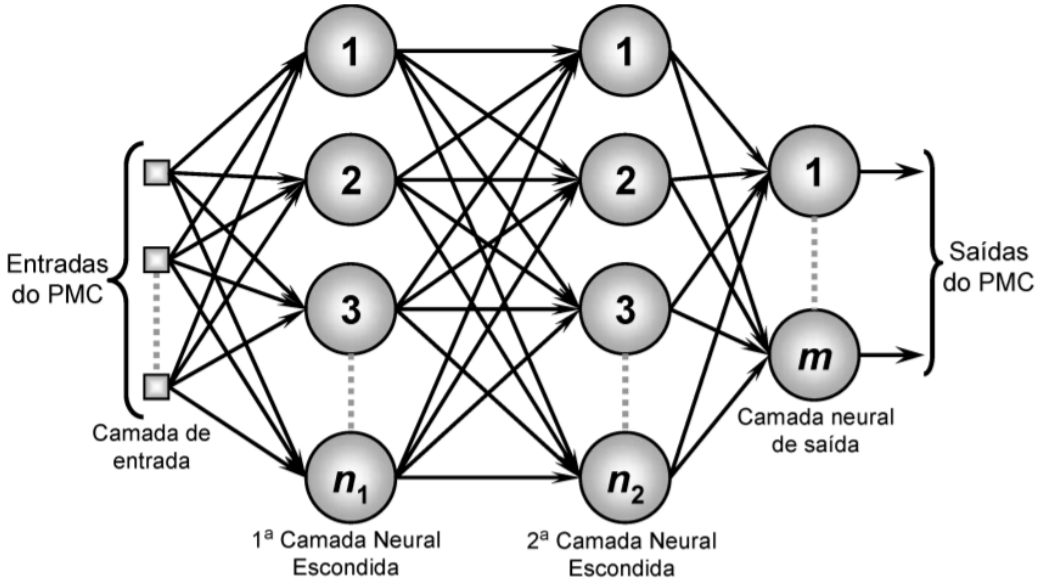
Fonte: Autoria própria

2.5.4 Redes Perceptron Multicamadas

As MLP são caracterizadas pela presença de pelo menos uma camada intermediária de neurônios, situada entre a camada de entrada e a respectiva camada neural de saída, como ilustrada na Figura 9. O fluxo básico do algoritmo consiste em duas fases, a primeira é conhecida como processamento direto, do inglês *forward pass*, onde a entrada é propagada pela rede até o última camada.

Nessa primeira fase, pode-se ver que o fluxo de ida da rede não se altera, mas como os pesos serão atualizados para a próxima passagem. Já a segunda fase é conhecida como processamento reverso, do inglês *backward pass*, para esse passo é calculado o gradiente da função de perda na última camada, ou seja, para cada previsão existirá um erro atrelado com base no valor esperado, conhecido como processo supervisionado, com isso a fase de processamento reverso garante que um sinal será propagado a cada camada da rede onde os valores serão atualizados para a próxima iteração do algoritmo. Esse processo iterativo permite minimizar o erro em função do valor esperado (SILVA; SPATTI; FLAUZINO, 2016).

Figura 9 – Ilustração da rede PMC.



Fonte: Adaptado de (SILVA; SPATTI; FLAUZINO, 2016)

As redes MLP são ainda caracterizadas pelas elevadas possibilidade de aplicações em diversos tipos de problemas relacionados com as mais diferentes áreas do conhecimento, sendo também consideradas uma das arquiteturas mais versáteis quanto à aplicabilidade. Entre essas potencias áreas, têm-se os seguintes destaques:

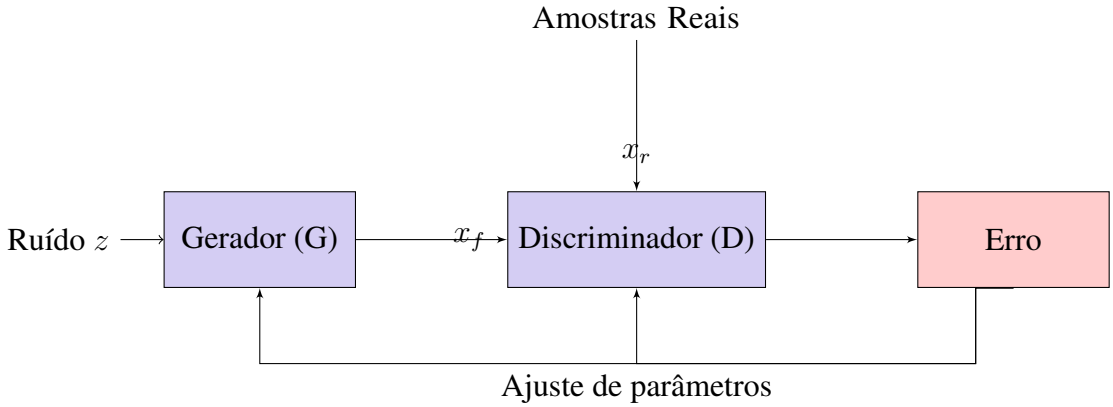
- Aproximação universal de funções;
- Reconhecimento de padrões;
- Identificação e controle de processos;
- Previsão de séries temporais;
- Otimização de sistemas;

Neste trabalho foi utilizado o MLP para aproximação universal de funções, para estimar os MFCC e F_0, e reconhecimento de padrões, para classificar os sinais de voz nos vogais corretos.

2.5.5 Redes Generativas Adversárias

As GAN foram propostas por Goodfellow et al. (2014). O diferencial das GAN em relação a outros modelos generativos, como máquinas de Boltzmann, redes Bayesianas e *autoencoders* variacionais, é o uso do treino adversário. O esquemático genérico de uma GAN é apresentado na Figura 10 (PRAGER; HARRISON; FALLSIDE, 1986; DEVIREN, 2002; GROZDIĆ; JOVIČIĆ, 2017).

Figura 10 – Arquitetura de uma rede GAN.



Fonte: Autoria própria

A entrada da rede G é um ruído derivado de uma distribuição Gaussiana que é modificada pela rede G para parecer uma amostra real. Um conjunto de amostras falsas (x_f) é criado e é associado com um conjunto de amostras reais (x_r). Esse novo conjunto formado com os dois tipos de amostras é alimentado para a rede D que avaliará a probabilidade de cada amostra pertencer ao conjunto real ou ao conjunto falso. O resultado desse processo é utilizado para

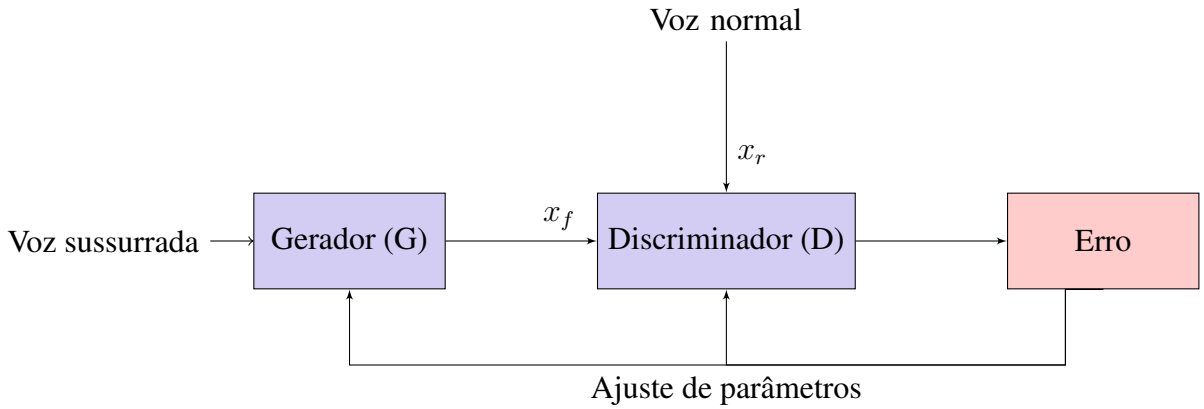
corrigir eventuais falhas de ambos os modelos, que refazem o procedimento considerando essas novas informações.

Em Goodfellow et al. (2014) faz-se analogia desse treinamento adversário com os falsificadores de dinheiro e a polícia. As GAN possuem duas redes, dos falsificadores, conhecido como rede gerador (G), que criam amostras falsas com o objetivo de passá-las como amostras autênticas, e dos policiais, conhecido como rede discriminador (D), que recebe as amostras reais e falsas para distingui-las. Ambas as redes são treinadas simultaneamente e esse ambiente de competição faz com que os dois melhorem suas técnicas.

A aplicação da arquitetura GAN está sendo amplamente utilizada em diversas pesquisas, principalmente em processamento de imagens (CRESWELL et al., 2018). Na área de reconhecimento de fala, as pesquisas realizada por Pascual, Serra e Bonafonte (2019) e Shah et al. (2018) utilizam GAN para mapear as características da fala normal e sussurrada.

Neste trabalho, a GAN é utilizada conforme mostrado na Figura 11. O sinal de ruído z como sinal de voz sussurrada, que passará pela rede geradora, tornando uma amostra falsa de voz normal x_f. A rede discriminador irá distinguir entre os sinais de voz normal x_r e x_f, e a sua resposta servirá para ajuste de parâmetros dessas duas redes. Assim, a rede G irá gerar uma voz sintetizada mais próxima do real.

Figura 11 – rede GAN para reconstrução de voz sussurada.



Fonte: Autoria própria

Embora GAN tenham apresentado grande sucesso em várias aplicações, o treinamento de uma GAN é um processo complicado. Este processo é conhecido por ser lento e instável. Alguns problemas são listado abaixo (ZHAO; XIA; TOGNERI, 2019):

- Treinamento instável: No processo de treinamento da GAN, dois modelos são treinados simultaneamente. No entanto, a função custo desses dois modelos são independentes e a atualização do gradiente de ambos os modelos ao mesmo tempo não garante a convergência. Além disso, se o desempenho do discriminador D for perfeito, a função custo cairá para zero, assim não terá nenhum gradiente para atualizar o gerador G e este aprenderá nada. Por isso, é um dilema no qual o discriminador não poderá ser muito bom ou muito ruim.

- *Mode Collapse*: Durante o treinamento, os pesquisadores tem percebido que algumas vezes as imagens geradas por G estão com o mesmo padrão. Em outras palavras, o gerador carece de diversidade e faz com que as entradas "falsas"criadas pareçam todas iguais. Este problema é conhecido como *mode collapse*, no qual o gerador, mesmo recebendo diferentes tipos de sinais de entrada, gera amostras parecidas.

- Falta de uma métrica de avaliação: A função objetivo das GANs pode não dar uma indicação clara do progresso do treinamento e a relação entre as curvas de perda de treinamento e o progresso do treinamento é confusa e difícil de ser interpretada. Também não há indicadores explícitos quanto ao critério de parada do processo de treinamento.

A fim de contornar esses problemas, várias possíveis soluções foram propostos por diversos pesquisadores e serão apresentados a seguir (ZHAO; XIA; TOGNERI, 2019):

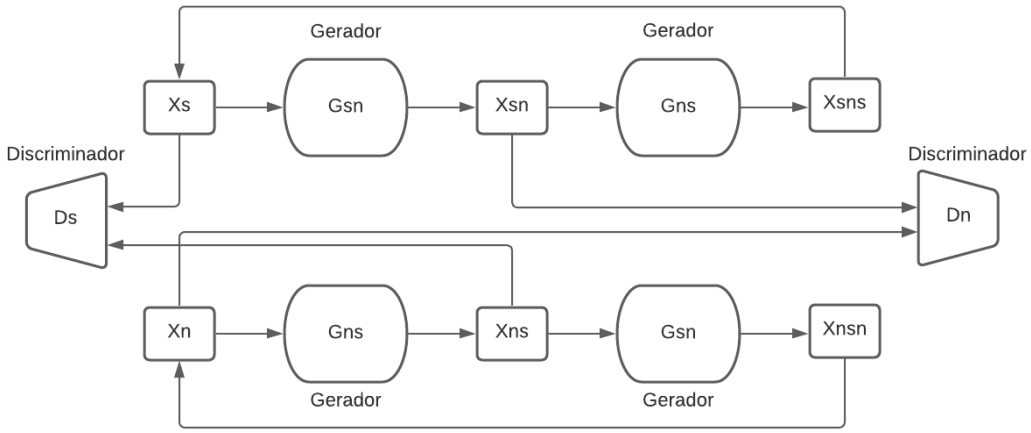
- Wassestein GAN: O método tradicional para o treinamento de GANs nada mais é do que a redução da Divergência de Jensen-Shannon entre a distribuição do gerador e a distribuição real, conforme descrito por Goodfellow et al. (2014). Divergência de Jensen-Shannon, porém apresenta um problema para quantificar a similaridade entre distribuições que não possuem sobreposição. Comparando a divergência de Jensen-Shannon com distância Wasserstein, existe o problema de medir a similaridade entre duas distribuições representando linhas paralelas e a única forma de apresentar um valor de similaridade que varia de forma suave com a distância entre as linhas é a distância Wasserstein. Devido a forma suave com a qual a distância varia com a aproximação das distribuições ela apresenta bons resultados em relação ao aumento da estabilidade no processo de treinamento de uma GAN.

- *Deep Convolutional* GAN: É uma extensão da rede GAN por utilizar as redes convolucionais nas redes geradoras e adversárias, assim, algumas configuração são ajustadas para que o treinamento da rede gerador sempre seja estável.

- DiscoGAN: A arquitetura DiscoGAN está ilustrada na Figura 12. As amostras Xn e Xs representa respectivamente a voz normal e voz sussurrada. Essa rede consiste na utilização de duas redes geradoras, Gns e Gsn, e duas redes discriminadoras Ds e Dn. Gsn converte Xs para Xsn (características convertidas a partir do voz normal) de tal modo que é semelhante as amostras reais Xn. E o Dn irá distiguir entre o Xn e Sns. As redes Gns e Ds também irá operar de forma análoga. (SHAH et al., 2018)

Para aplicações em ASR, o trabalho de Pascual, Serra e Bonafonte (2019) utilizou as redes GAN para eliminar os ruídos do ambiente, nomeando como *Speech Enhancement GAN* (SEGAN). Ainda nesse trabalho, fez uma modificação da sua rede SEGAN para conversão da fala sussurrada para normal, trazendo resultados interessantes. Em outro trabalho, Shah et al. (2018), Parmar et al. (2019) apresentaram comparações de desempenho entre arquiteturas GAN,

Figura 12 – Arquitetura da rede DiscoGAN



Fonte: Adaptado de (SHAH et al., 2018)

demonstrando as vantagens e desvantagens de cada arquitetura. No próximo capítulo serão apresentadas as metodologias utilizadas neste estudo.

3 MATERIAIS E MÉTODOS

Neste capítulo, são descritos o banco de dados utilizado neste trabalho, o método de pré-processamento de sinais, explicar como foram determinado os parâmetros para criação das topologias das redes neurais para conversão da voz normal para sussurrada e, por fim, descreve-se as métricas de validação da metodologia.

3.1 BANCO DE DADOS

Para o desenvolvimento deste estudo foi utilizado o banco de dados do projeto chamado DyNaVoiceR [1] (*Dysphonic to Natural Voice Reconstruction*) concedido pelo Prof. Dr. Aníbal Ferreira. Essa base de dados de oradores é uma ferramenta obtida pela parceria entre o LPSA (Laboratório de Processamento de sinais e Aplicações) do UTFPR com a FEUP (Faculdade de Engenharia da Universidade de Porto) em Porto, Portugal.

O banco de dados contem um conjunto de exercícios vocais realizados por 20 oradores (10 mulheres e 10 homens), gravados em arquivo de áudio no formato *.wav* amostrados a 22050 Hz, em que cada exercício possui versão normal e sussurrada. A base de dados inclui também anotação fonética manual destas gravações, na qual permite localizar e identificar cada fonema. Os exercícios contido neste banco de dados são (OLIVEIRA, 2020):

- 9 vogais orais utilizadas no Português Europeu, na forma sustentada (Tabela 1);
- 28 dissílabos, contendo várias daquelas vogais em contexto de palavra (Anexo A);
- 6 pequenas frases, contendo também alguns exemplos de vogais nasais (Anexo A).

Tabela 1 – As nove vogais orais do Português Europeu padrão disponíveis na base de dados DyNaVoiceR, no modo sustentado.

Código	Vogal	Exemplo
01	/i/	ilha
02	/ê/	peso
03	/é/	ela
04	/á/	água
05	/â/	amarelo
06	/ó/	óculos
07	/ô/	ovo
08	/u/	uva
09	/e/	sede

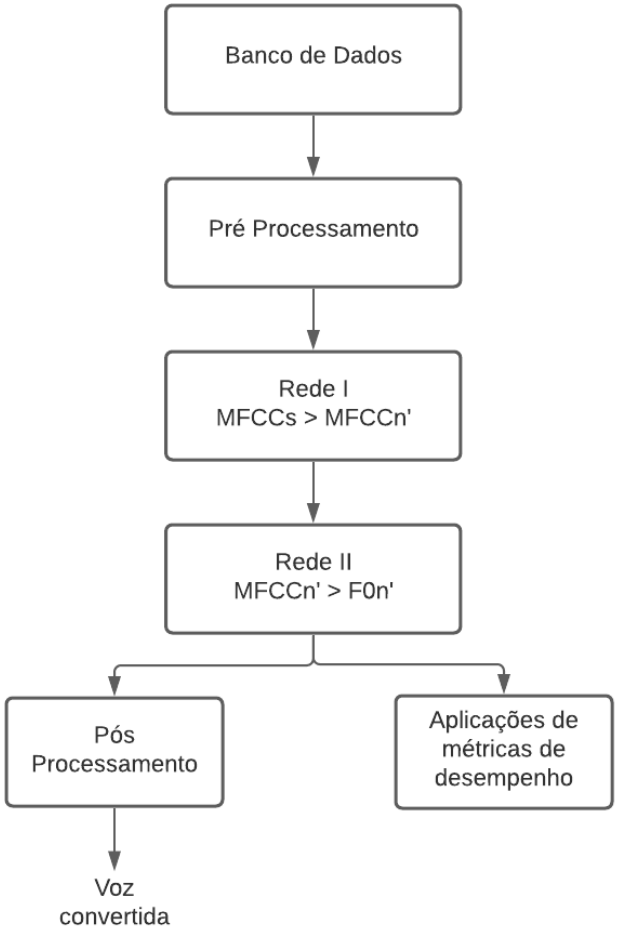
Fonte: Adaptado de (OLIVEIRA, 2020)

[1] link do projeto DyNaVoiceR : https://paginas.fe.up.pt/∼voicestudies/dynavoicer/

3.2 METODOLOGIA

A metodologia proposta nesse trabalho está representada no fluxograma na Figura 13. Inicialmente, com o detalhamento sobre o banco de dados descrito na seção anterior, em seguida, o pré-processamento de dados, a fim de adequar as amostras para a conversão de sinal realizado por duas redes neurais, e por fim, o pós processamento e a aplicação de métricas de desempenho. Cada uma das etapas será detalhada a seguir

Figura 13 – Diagrama de blocos da conversão de voz normal para voz sussurrada.



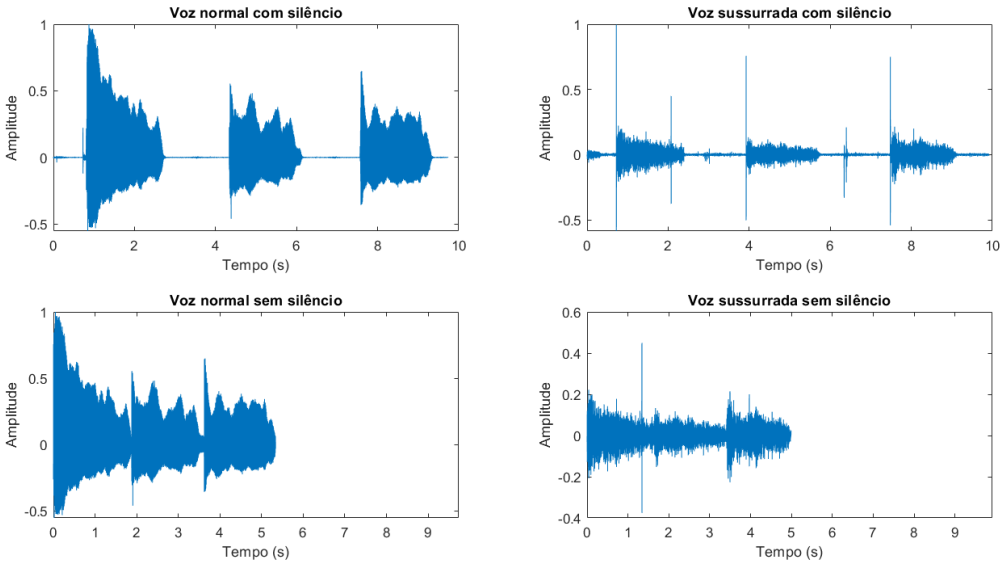
Fonte: Autoria própria

3.2.1 Pré-processamento

Inicialmente, foi retirado os intervalos sem fala dos sinais de banco de dados a fim de comparar posteriormente os resultados obtidos nas redes de conversão. Este corte foi realizado manualmente a partir das anotações fornecidas na base de dados. A forma de onda do sinal de voz normal e sussurrada de um orador falando três vezes a vogal "a"de forma sustentada está apresentada na Figura 14.

Para extração de características dos sinais, foi utilizado um vocoder (codificador de voz) chamado Ahocoder (ERRO et al., 2011). Para a utilização desse vocoder, foi necessário

Figura 14 – Formas de onda de um sinal de fala normal e sussurrada.



Fonte: Autoria própria

a redução de taxa amostral de 22050 Hz para 16000 Hz dos arquivos da base de dados. Este vocoder utiliza um janelamento de 5ms e para cada janela tem 80 amostras do sinal nas quais o Ahocoder calcula os seguintes parâmetros:

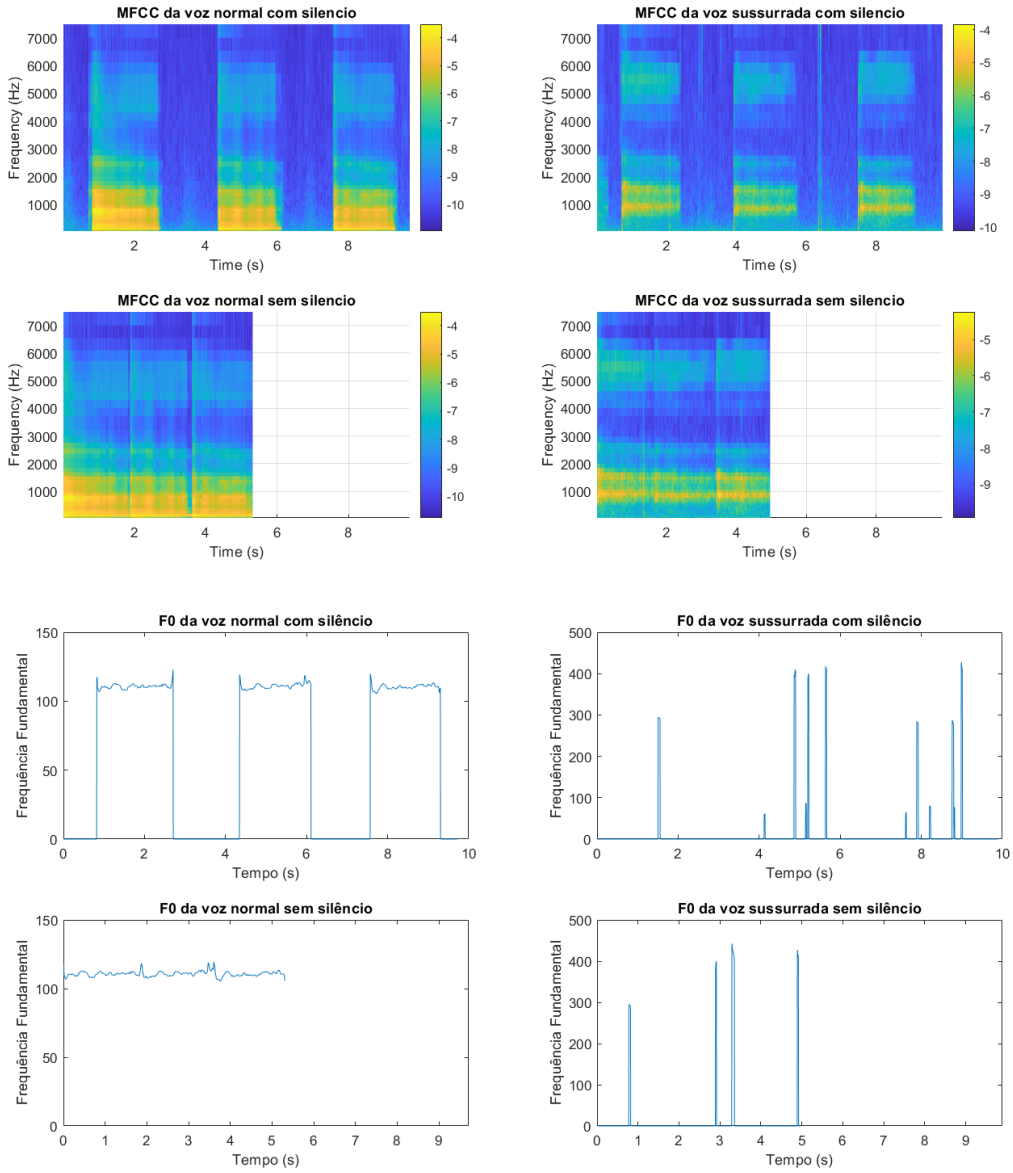
- MFCCs de ordem $p = 40$;
- O logaritmo da frequência fundamental ($log(F_0)$);

Os MFCCs e a frequência fundamental estão apresentados na Figura 15. Esta Figura são características extraídos da Figura 14, na qual observa-se também a comparação entre os sinais com o corte do silêncio. Percebe-se que a frequência fundamental na forma sussurrada possui alguns impulsos mas na maior parte do seu tempo a $F_0 = 0$, pois, como explicado no Capítulo 2, a falta de vibrações nas cordas vocais causa ausência da frequência fundamental e suas harmônicas. Após extração de características, será aplicada em sistema de conversão de voz sussurrada para normal.

3.2.2 Rede de conversão

Para a conversão de voz sussurrada para a normal foi utilizado duas redes neurais, nas quais estão ilustradas com mais detalhes na Figura 16. A Rede I mapeia as características do MFCC normal (MFCCn) com o MFCC sussurrada (MFCCs) após passarem pelo algoritmo DTW para o alinhamento de tempo entre eles a fim da rede gerar na fase de produção a estimativa do MFCC normal (MFCCn').

Figura 15 – Mel Espectrograma e frequência fundamental de um sinal de fala normal e sussurrada.



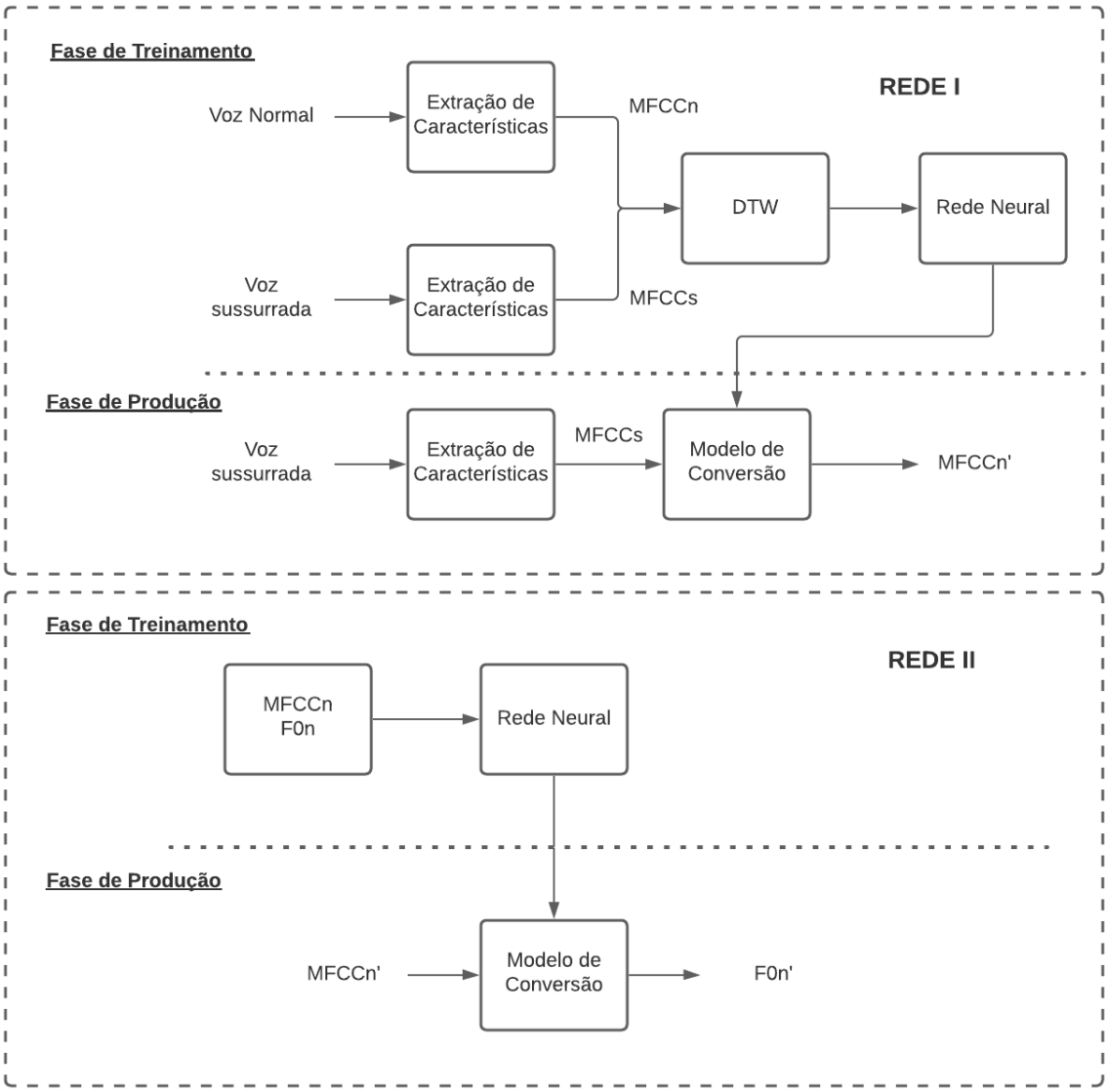
Fonte: Autoria própria

A Rede II, na sua fase de treinamento utiliza-se o MFCCn e F_0 do voz normal (F_0n) para mapear a rede, e na sua fase de produção, utiliza a saída da rede I, os MFCCn', para gerar a estimativa da F_0 (F_0n').

Para este estudo, foram utilizados três tipos de redes neurais: Redes Perceptron Multicamadas (MLP), Redes GAN e Redes DiscoGAN. A topologia das redes está descrita na Tabela 2, conforme sugerida em artigo Shah et al. (2018).

Todas as redes são compostas por três camadas escondidas de 512 neurônios com função de ativação ReLU. A camada de saída da Rede I é composta por função de ativação linear, porém, para as redes discriminadoras das GAN, são utilizados a função de ativação sigmoide, assim como toda Rede II, pois possui apenas uma saída.

Figura 16 – Diagrama de blocos da rede (a) Conversão MFCCs para MFCCc e (b) Conversão MFCCc para F_0c.



Fonte: Autoria própria

Tabela 2 – Topologia usada nas redes MLP e GAN.

Rede	Topologia Rede I	Topologia Rede II
MLP	40/512/512/512/40	40/512/512/512/1
GAN Gerador	40/512/512/512/40	40/512/512/512/1
GAN Discriminador	40/512/512/512/1	1/512/512/512/1
DiscoGAN Gerador	40/512/512/512/40	40/512/512/512/1
DiscoGAN Discriminador	40/512/512/512/1	1/512/512/512/1

Fonte: Autoria própria

As redes foram desenvolvidas em linguagem Python, e foram treinadas por 100 épocas, usando tamanho de *batches* de 1000 amostras, conforme sugerido em artigo Shah et al. (2018). Os parâmetros foram otimizados utilizando *Adam optimization*, com a taxa de aprendizagem de 0,0001.

3.2.3 Aplicações de métricas de desempenho

Para a verificação da eficácia do sistema de conversão da fala sussurrada para normal, foram utilizados três métricas: *Mel-Cepstrum Distortion* (MCD), raiz quadrada do erro médio (RMSE) do $log(F_0)$ e um classificador de vogais.

O cálculo do MCD está descrito na Equação (6), onde ${m_i}^n$ e ${m_i}^c$ são os MFCC da ordem $N = 40$ do sinal da fala normal e do sinal convertido. Como o MCD é a diferença entre os parâmetros cepstrais das falas normal e convertido da mesma pessoa e com a mesma frase, então se espera um resultado com MFCCn' parecido com o MFCCn, portanto quanto menor o MCD, mais eficiente é o sistema (PARMAR et al., 2019).

$$MCD = \frac{10}{ln(10)}\sqrt{2\sum_{i=1}^{N}({m_i}^n - {m_i}^c)^2}\ \text{[dB]} \tag{6}$$

Para estimativa do RMSE do $log(F_0)$, o sinal da fala normal e do convertido são alinhados no tempo usando o algoritmo *Dynamic Time Warping* (DTW). Esse algoritmo gera pares de sons vozeado-vozeado, vozeado-não vozeado e não vozeado-não vozeado. Para este estudo foram usados apenas os pares vozeado-vozeado para o cálculo do RMSE do $log(F_0)$, pois o F_0 é indefinido para não vozeado. A estimativa do RMSE do $log(F_0)$ é dado pela Equação (7) (PARMAR et al., 2019):

$$RMSE(log(F_0)) = \sqrt{\sum_{i=1}^{k}(log({F_{0i}}^n) - log({F_{0i}}^c))^2} \tag{7}$$

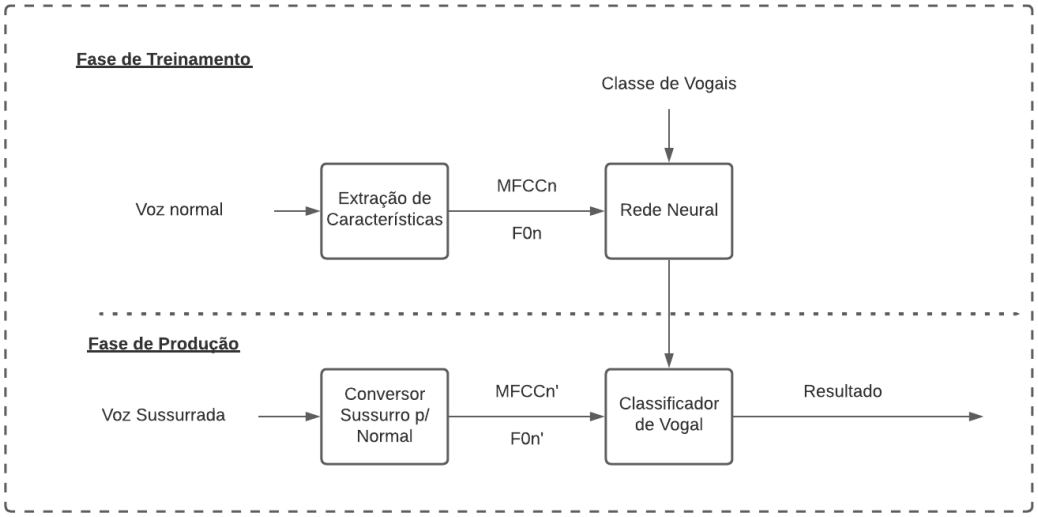
onde k é o número total de pares vozeado-vozeado depois do alinhamento, e ${F_{0i}}^n$ e ${F_{0i}}^c$ são do sinal da fala normal e convertido, respectivamente. Como novamente é necessário que as ${F_{0i}}^n$ e ${F_{0i}}^c$ sejam parecidas pois são da mesma pessoa e com a mesma frase, então quanto o menor for o valor do RMSE, mais eficiente é o sistema.

E a última métrica será a utilização do classificador de vogais, no qual está ilustrada na Figura 17. Esse classificador é uma rede perceptron multicamadas com três camadas escondidas contendo 128 neurônios com funções de ativação ReLU. Esta rede foi treinada com MFCCn e F_0n, ou seja, 41 neurônios na camada de entrada. E na camada de saída utilizou-se a função de ativação sigmoide com a quantidade de vogais que for introduzido na rede. Após o mapeamento da rede, na fase de produção é inserido o MFCCn' e F_0n', que foi obtido na rede de conversão, para que a rede classifique o vogal convertido.

A análise de resultado desta métrica é realizado a partir da mapa de confusão, que é uma tabela que demonstra a saída da rede comparando com o resultado esperado.

O próximo capítulo apresentará as análise de resultados deste trabalho, cujo objetivo principal é a validação da metologia proposta.

Figura 17 – Diagrama de blocos do classificador de vogais da voz sussurrada.



Fonte: Autoria própria

4 RESULTADOS E DISCUSSÕES

Neste capítulo, os resultados obtidos são apresentados e discutidos para a validação da proposta do trabalho. Seguindo a metodologia descrita no capítulo anterior, são analisados de forma detalhada sobre os sinais e as características extraídas do banco de dados, assim como os resultados aplicando as métricas de desempenho do sistema de conversão da fala sussurrada para o normal: MCD, RMSE do $log(F_0)$ e classificador de vogais utilizando rede perceptron multicamadas.

4.1 ANÁLISE DO BANCO DE DADOS

O banco de dados contêm gravação de diversos conjunto de exercícios vocais, mas, para este trabalho, utilizaram-se apenas as 9 vogais orais em português europeu, na forma sustentada, como descrita na Tabela 1 (Capítulo 3).

As 9 vogais orais foram separadas em quatros grupos de base de dados a fim de verificar o funcionamento e desempenho das redes de conversão da fala sussurrada para o normal. A Tabela 3 descreve esta divisão dos vogais.

Tabela 3 – Divisão do banco de dados.

Agrupamentos	Vogais
1	/á/ /â/
2	/á/ /â/ /ê/ /é/
3	/á/ /â/ /ê/ /é/ /ó/ /ô/
4	/á/ /â/ /ê/ /é/ /ó/ /ô/ /i/ /u/ /ú/

Fonte: Autoria própria

Além de separar o banco de dados por vogais, também realizou-se a divisão por sexo (homens e mulheres) e por sinais com corte no silêncio como demonstrado na Figura 14 (Capítulo 3). A Figura 18 apresenta a frequência fundamental do homem e da mulher.

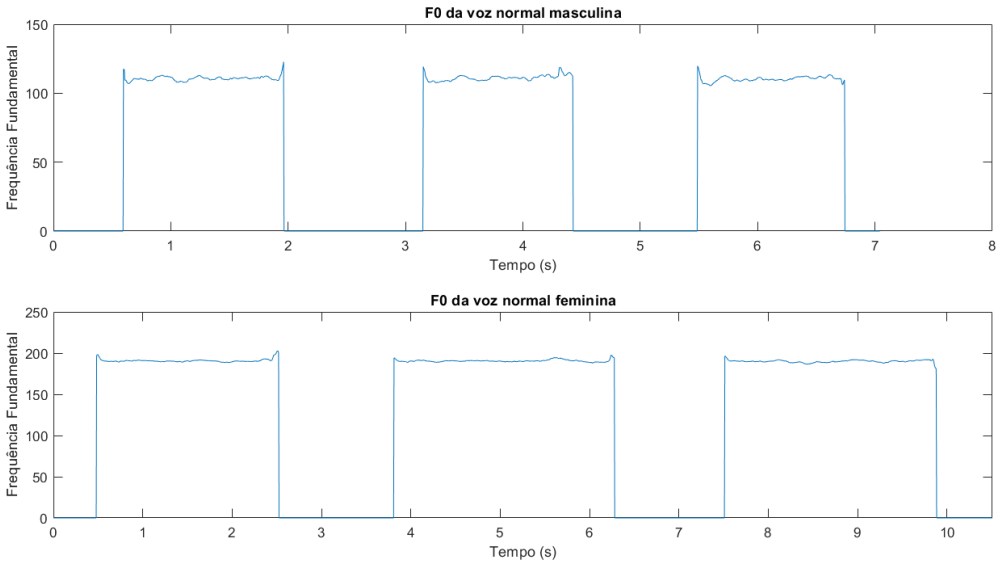
Observa-se que a frequência fundamental é diferentes entre os sexos, pois segundo Behlau (2001), o homem adulto tem pregas vocais maiores e espessas, por isso suas pregas vibram entre 80 e 150 Hz, e das mulheres possuem pregas vocais menores e menos espessas, variando sua vibração entre 150 e 250 Hz.

4.2 ANÁLISE MCD

A primeira métrica a ser discutida neste capítulo é o MCD. Esta é a diferença entre os parâmetros cepstrais da fala normal e convertido, e quanto menor o MCD, mais eficiente será o sistema. A Tabela 4 e Tabela 5 apresentam o MCD e desvio padrão do sistema utilizando banco de dados com e sem silêncio, respectivamente.

Comparando os resultados em relação as redes neurais, a média do MCD usando o MLP possui um desempenho melhor que o GAN e DiscoGAN, uma diferença de 36% e 28%,

Figura 18 – Frequêcia Fundamental entre homem e mulher.



Fonte: Autoria própria

Tabela 4 – MCD do banco de dados com silêncio.

Agrupamentos	Masculino			Feminino		
	MLP	GAN	DiscoGAN	MLP	GAN	DiscoGAN
1	**4,58 ± 0,39**	$5,60 \pm 0,38$	$5,26 \pm 0,48$	**4,83 ± 0,61**	$5,05 \pm 0,74$	$5,91 \pm 0,78$
2	**3,93 ± 0,30**	$5,31 \pm 0,83$	$5,24 \pm 0,56$	**4,03 ± 0,45**	$5,81 \pm 0,75$	$5,96 \pm 0,85$
3	**4,70 ± 0,47**	$5,58 \pm 0,60$	$6,78 \pm 0,54$	**4,87 ± 0,64**	$9,20 \pm 1,01$	$5,99 \pm 0,66$
4	**4,72 ± 0,57**	$6,52 \pm 0,56$	$5,82 \pm 0,59$	**4,99 ± 0,66**	$6,69 \pm 0,93$	$5,94 \pm 0,73$
média	**4,48 ± 0,43**	$5,75 \pm 0,59$	$5,77 \pm 0,54$	**4,68 ± 0,59**	$6,69 \pm 0,85$	$5,95 \pm 0,75$

Fonte: Autoria própria

Tabela 5 – MCD do banco de dados sem silêncio.

Agrupamentos	Masculino			Feminino		
	MLP	GAN	DiscoGAN	MLP	GAN	DiscoGAN
1	**3,93 ± 0,48**	$7,91 \pm 1,78$	$4,74 \pm 0,73$	**4,21 ± 0,66**	$6,41 \pm 0,83$	$6,04 \pm 1,01$
2	**4,43 ± 0,75**	$5,11 \pm 0,80$	$5,12 \pm 0,71$	**4,84 ± 0,61**	$6,31 \pm 0,58$	$6,76 \pm 0,78$
3	**4,58 ± 0,75**	$5,17 \pm 0,80$	$5,22 \pm 0,85$	**4,52 ± 0,70**	$5,05 \pm 0,81$	$5,16 \pm 0,72$
4	**4,71 ± 0,79**	$7,39 \pm 0,92$	$5,27 \pm 0,78$	**4,84 ± 0,71**	$12,01 \pm 2,61$	$5,33 \pm 0,77$
média	**4,41 ± 0,69**	$6,39 \pm 1,08$	$5,08 \pm 0,77$	**4,60 ± 0,67**	$7,45 \pm 1,20$	$5,82 \pm 0,82$

Fonte: Autoria própria

respectivamente. Foi possível verificar que estatisticamente, não há diferença entre voz masculina e feminina. Por fim, comparando os resultados da Tabela 4 e Tabela 5, não houve diferença no MCD em utilizar o banco de dados com corte no silêncio.

4.3 ANÁLISE RMSE

A segunda métrica é o RMSE do $log(F_0)$. Este é o cálculo do erro dos pares vozeados depois do alinhamento no tempo usando o DTW, portanto, quanto menor o RMSE, mais eficiente será o sistema. A Tabela 6 e Tabela 7 apresenta o RMSE e desvio padrão do sistema utilizando

banco de dados com e sem silêncio, respectivamente.

Tabela 6 – RMSE do $log(F_0)$ do banco de dados com silêncio.

Agrupamentos	Masculino		
	MLP	GAN	DiscoGAN
1	$13,90 \pm 5.24$	$13,68 \pm 5,97$	**10,59 ± 4,05**
2	$16,00 \pm 9,27$	**10,70 ± 7,16**	$12,95 \pm 7,41$
3	$15,59 \pm 7,31$	**17,44 ± 10,59**	$17,44 \pm 11,38$
4	$17,06 \pm 9,32$	$19,52 \pm 9,49$	**16,75 ± 10,24**
média	$15,63 \pm 7,78$	$15,36 \pm 8,30$	**14,43 ± 8,27**

Agrupamentos	Feminino		
	MLP	GAN	DiscoGAN
1	$21,33 \pm 8,10$	**17,72 ± 9,86**	$21,80 \pm 13,93$
2	$17,52 \pm 8,91$	$14,56 \pm 8,97$	**13,67 ± 8,24**
3	$22,86 \pm 10,20$	$20,98 \pm 9,92$	**20,65 ± 10,77**
4	$23,40 \pm 11,55$	**18,15 ± 8,57**	$21,61 \pm 12,53$
média	$21,27 \pm 9,69$	**17,85 ± 9,33**	$19,43 \pm 11,36$

Fonte: Autoria própria

Tabela 7 – RMSE do banco de dados sem silêncio.

Agrupamentos	Masculino		
	MLP	GAN	DiscoGAN
1	**6,43 ± 3,67**	$10,10 \pm 7,45$	$6,69 \pm 7,27$
2	**6,17 ± 2,23**	$6,59 \pm 2,42$	$8,04 \pm 3,52$
3	$6,87 \pm 3,54$	**6,72 ± 3,79**	$7,24 \pm 4,29$
4	**7,83 ± 3,61**	$9,38 \pm 5,12$	$8,13 \pm 4,39$
média	**6,82 ± 3,26**	$8,20 \pm 4,69$	$7,52 \pm 4,86$

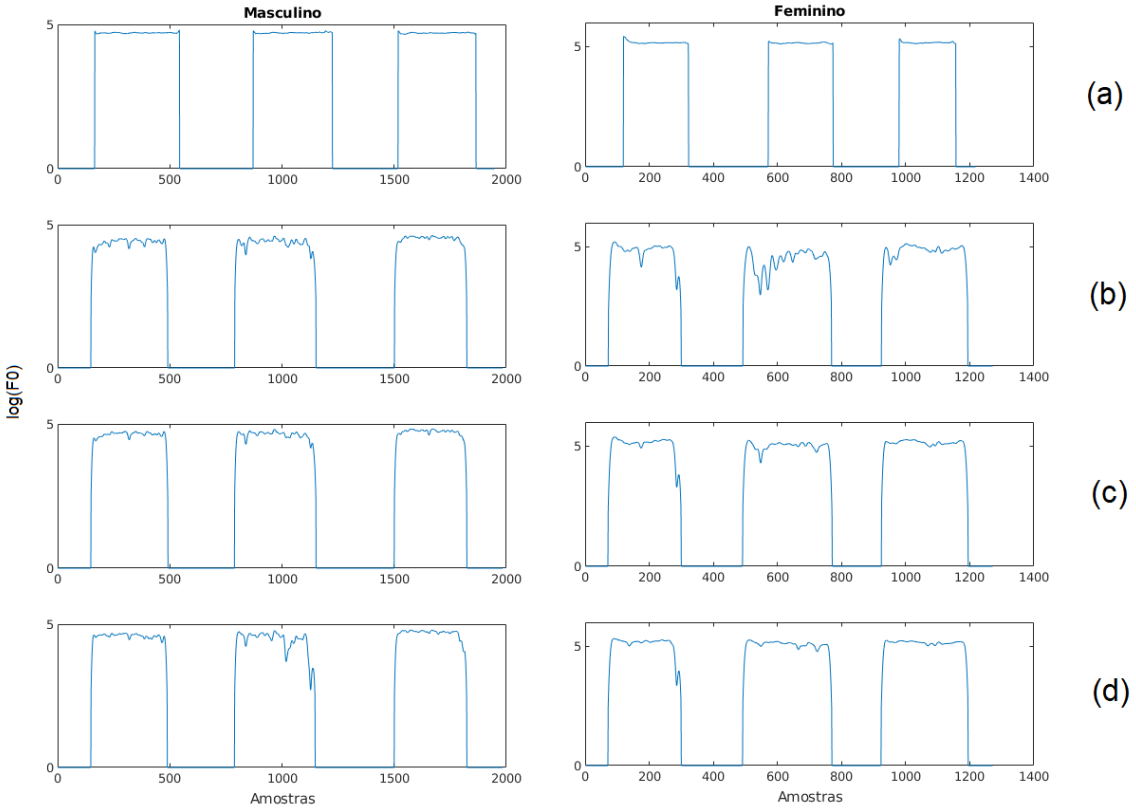
Agrupamentos	Feminino		
	MLP	GAN	DiscoGAN
1	$7,55 \pm 3,05$	**7,07 ± 3,35**	$8,70 \pm 6,79$
2	$19,56 \pm 8,30$	**16,50 ± 9,21**	$21,73 \pm 11,51$
3	$9,04 \pm 3,34$	**7,29 ± 3,48**	$7,43 \pm 3,01$
4	$10,02 \pm 3,77$	$9,38 \pm 7,19$	**7,20 ± 3,43**
média	$11,54 \pm 4,61$	**10,06 ± 5,80**	$11,26 \pm 6,18$

Fonte: Autoria própria

Observando os valores de RMSE, para banco de dados com silêncio foi obtido valores em casas decimais enquanto para banco de dados sem silêncio teve valores melhores, isso demonstra que a presença do silêncio dificulta o cálculo da estimativa do $log(F_0)$. A Figura 19 ilustra a estimativa do contorno do $log(F_0)$ de um sinal sem o corte do silêncio utilizando as três redes neurais.

Observa-se na Figura 19 que a estimativa do $log(F_0)$ usando MLP, GAN e DiscoGAN ficaram próxima à voz normal correspondente à voz sussurrada, porém com uma certa oscilação. Em relação aos sistemas, o MLP obteve um desempenho inferior que o GAN e DiscoGAN, porém, se considerarmos o desvio padrão do RMSE, os três sistemas possuem o mesmo desempenho. E a utilização do corte no silêncio teve efeito nos resultados, uma melhora de 46%, em relação ao sistema com o banco de dados com silêncio.

Figura 19 – (a) $log(F_0)$ da voz normal correspondente e estimativa do $log(F_0)$ usando o (b) MLP, (c) GAN, (d) DiscoGAN, dos sinal masculino e feminino.



Fonte: Autoria própria

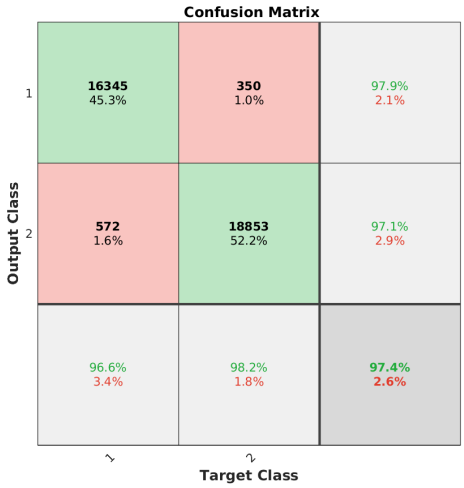
4.4 CLASSIFICADOR DE VOGAIS

A última análise é o classificador de vogais usando MLP com três camadas escondidas. Esta rede foi treinada a partir dos sinais de voz normal e usada para distinguir as vogais dos sinais convertidos. Foi utilizado a matriz de confusão para demonstrar as características dos resultados obtidos com a rede, como demonstrado na Figura 20, que ilustra o resultado da rede usando como entrada o banco de dados com vogal /a/ e /e/ masculino do MLP, GAN e DiscoGAN. Esta matriz de confusão apresenta a saída da rede classificando os vogais e também o silêncio. Todos os resultados são a média da validação cruzada K-fold com k=10, portanto foi dividido o banco de dados em 10 partes e feito revezamento do teste.

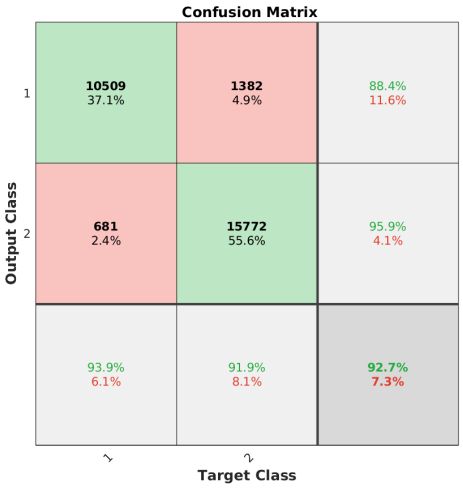
As Tabela 8 e Tabela 9 apresentam a acurácia das redes com e sem o silêncio, respectivamente. No caso do base de dados 1, que possui apenas o vogal /a/, foi utilizado para classificar entre o vogal /a/ e o silêncio, assim, na Tabela 9 não foi utilizado este base de dados. As matrizes de confusão não apresentadas neste capítulo estão no apêndice A.

Observando as Tabelas 8 e 9, verifica-se que em modo geral, o MLP possui melhor desempenho em relação a GAN e DiscoGAN. Percebe-se também que conforme vai aumentando

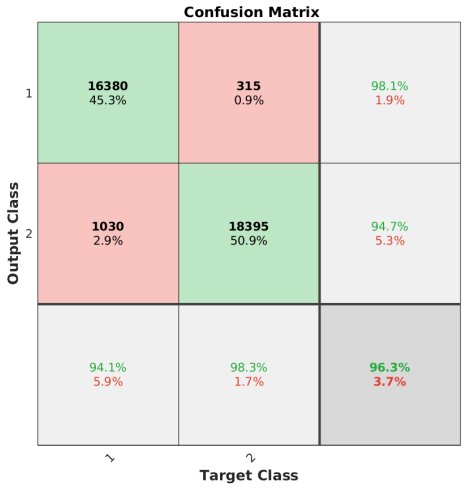
Figura 20 – Matriz de confusão utilizando banco de dados 1 - Vogal /a/ e silêncio



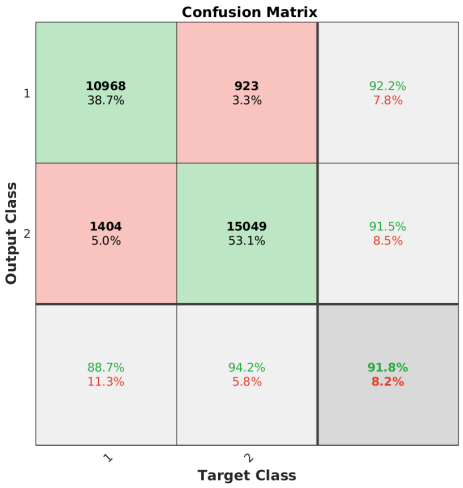
(a) MLP - Masculino



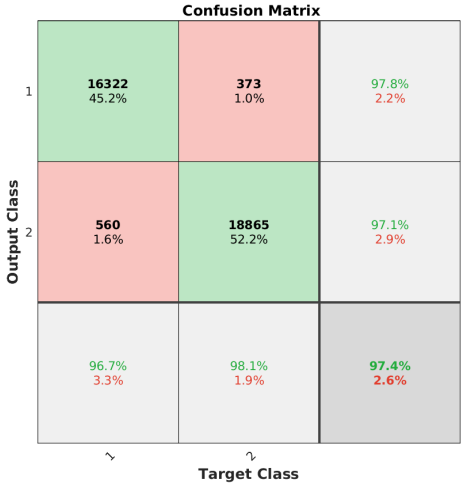
(b) MLP - Feminino



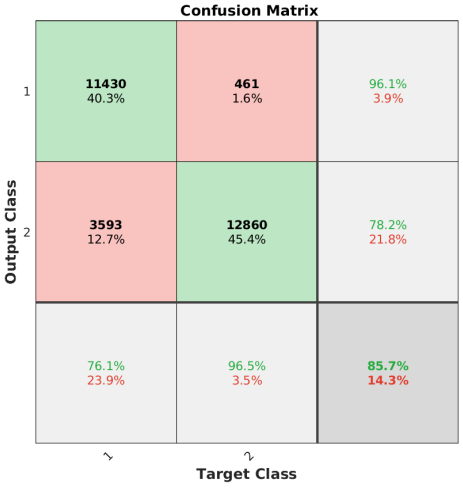
(c) GAN - Masculino



(d) GAN - Feminino



(e) DiscoGAN - Masculino



(f) DiscoGAN - Feminino

Fonte: Autoria própria

Tabela 8 – Acurácia do banco de dados com silêncio.

Base de dados	Masculino			Feminino		
	MLP	GAN	DiscoGAN	MLP	GAN	DiscoGAN
1	**97,4%**	96,3%	**97,4%**	**92,7%**	91,8%	85,7%
2	92,8%	**93,5%**	91,2%	85,4%	90,4%	**92,4%**
3	**94,8%**	93,1%	88,8%	**89,4%**	74,1%	84,8%
4	**88,3%**	83,4%	85,5%	79,8%	**81,9%**	77,0%
média	**93,3%**	91,6%	90,7%	**86,8%**	84,5%	84,9%

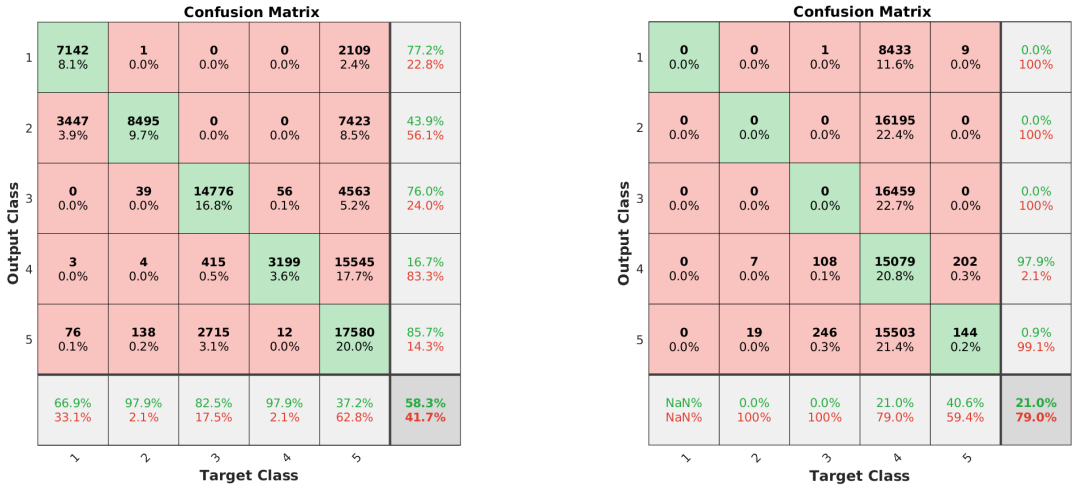
Fonte: Autoria própria

Tabela 9 – Acurácia do banco de dados sem silêncio.

Base de dados	Masculino			Feminino		
	MLP	GAN	DiscoGAN	MLP	GAN	DiscoGAN
2	**99,5%**	99,4%	99,3%	80,3%	82,6%	**83,7%**
3	**97,6%**	97,4%	94,6%	93,9%	94,5%	**95,7%**
4	**87,7%**	58,3%	77,6%	**84,7%**	21,0%	84,2%
média	**94,9%**	85,0%	90,5%	86,3%	66,0%	**87,8%**

Fonte: Autoria própria

Figura 21 – Matriz de confusão para agrupamento 4 sem silêncio usando GAN.



(a) Masculino

(b) Feminino

Fonte: Autoria própria

a quantidade de vogais na base de dados, menor a sua acurácia, isso é justificável pelo aumento de variáveis que a rede terá que classificar.

Na Tabela 9, ao usar a base de dados 4, o GAN teve o pior desempenho, de $58,3\%$ e $21,0\%$, masculino e feminino respectivamente. A matriz de confusão para agrupamento 4 usando GAN é apresentado na Figura 21. É possível analisar pela matriz de confusão que a rede não conseguiu classificar corretamente os vogais. Na voz masculina, a rede acertou $58,3\%$ e o restante a classificou como variável 5, que é a vogal /u/. E na voz feminina, a rede classificou todos os sinais como variável 4, a vogal /o/.

O próximo capítulo apresenta a conclusão desse estudo evidenciando a metodologia e os resultados discutidos e propondo novos caminhos de pesquisa.

5 CONCLUSÃO

O principal objetivo deste trabalho foi o estudo da características de voz a fim de implementar um sistema de conversão de voz normal para sussurrada utilizando a rede generativa adversárias. E a contribuição deste trabalho está na utilização do banco de dados em português europeu fornecido pela parceria entre FEUP e UTFPR em um sistema de conversão de voz.

Para a validação da metodologia, foi aplicada três redes neurais em sistema de conversão de voz: i)MLP, ii)GAN e iii)DiscoGAN. Em todas as redes foram realizadas com banco de dados separados por sexos, agrupamentos de vogais diferentes e sinais com e sem corte do silêncio nos intervalos da fala. A fim de verificar o desempenho de cada sistema, foi realizado três tipos de análise: i) MCD, ii) RMSE do $log(F_0)$ e iii) Classificador de vogal.

Na primeira análise, foi comparado o MCD em relação as redes neurais e a média do MCD usando o MLP possui um desempenho melhor que o GAN e DiscoGAN, uma diferença de 36% e 28%, respectivamente. Na segunda análise, o RMSE do $log(F_0)$ do sistema foi alto, pois a presença do silêncio dificulta o cálculo da estimativa de $log(F_0)$. Contudo, com o corte do silêncio entre os intervalos de fala, o RMSE teve uma melhora de 46% em relação ao sistema com o banco de dados com silêncio, valores de 6,82 para voz masculina e 10,06 para voz feminina.

A última análise foi o classificador de vogais, usada para identificar corretamente as vogais dos sinais convertidos. Na classificação com o banco de dados com silêncio, a rede teve uma melhor acurácia com os sinais convertido pelo sistema usando MLP, de 93,3% para voz masculina e 86,8% para voz feminina. Enquanto a acurácia do banco de dados sem o silêncio, os sinais do MLP foram melhores, 94,9% para voz masculina, e para voz feminina, o DiscoGAN foi melhor, com 87,8%.

Nas análises foram observados que os resultados utilizando o banco de dados com o corte do silêncio entre o intervalo de fala teve uma melhora no RMSE. E avaliando entre as redes MLP, GAN e DiscoGAN, o MLP teve resultados melhores para análise de MCD e classificador de vogais, e GAN foi melhor no RMSE do $log(F_0)$.

Para os trabalhos futuros associados a esta dissertação possuem a perspectiva de: aplicar a metodologia em palavras e frases completas do banco de dados DyNaVoiceR mostrado no Anexo A; utilizar outros tipos de redes neurais, como redes convolucionais, para comparar com resultados deste estudo; realizar uma pesquisa subjetiva com pessoas para ouvirem os sinais convertidos de sussurrada para normal das redes MLP, GAN e DiscoGAN.

REFERÊNCIAS

BEHLAU, Mara. **Voz, o livro do especialista**. São Paulo: Revinter, 2001. Citado na página 32.

COSTA, João Filipe Torres. Adaptive phonetic segmentation in dysphonic voice. Dissertação de Mestrado Integrado, Faculdade de Engenharia da Universidade do Porto, Porto, Portugal, 2021. Citado na página 10.

CRESWELL, Antonia et al. Generative adversarial networks: An overview. **IEEE Signal Processing Magazine**, v. 35, n. 1, p. 53–65, 2018. Citado na página 22.

DELLER, John R.; PROAKIS, John G.; HANSEN, John H. **Discrete Time Processing of Speech Signals**. 1st. ed. USA: Prentice Hall PTR, 1993. ISBN 0023283017. Citado 2 vezes nas páginas 12 e 15.

DEVIREN, Murat. Dynamic Bayesian Networks for Automatic Speech Recognition. In: **Eighteenth National Conference on Artificial Intelligence**. Edmonton, Alberta, Canada: American Association for Artificial Intelligence, 2002. p. 981. ISBN 0262511290. Citado na página 21.

ERRO, Daniel et al. Improved HNM-based vocoder for statistical synthesizers. In: **INTERSPEECH**. Florence, Italy: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2011. p. 1809–1812. Citado na página 26.

FERREIRA, Aníbal. Implantation of voicing on whispered speech using frequency-domain parametric modelling of source and filter information. In: . Tunis, Tunisia: 2016 International Symposium on Signal, Image, Video and Communications (ISIVC), 2016. p. 159–166. Citado na página 9.

GHAFFARZADEGAN, Shabnam; BORIL, Hynek; HANSEN, John. Deep neural network training for whispered speech recognition using small databases and generative model sampling. **International Journal of Speech Technology**, v. 20, 2017. Citado na página 9.

GOODFELLOW, Ian et al. Generative Adversarial Networks. **Advances in Neural Information Processing Systems**, v. 3, 2014. Citado 4 vezes nas páginas 9, 21, 22 e 23.

GOPI, E.S. **Digital Speech Processing Using Matlab**. United States: Springer, 2013. Citado 2 vezes nas páginas 14 e 18.

GROZDIĆ, Đorđe T; JOVIČIĆ, Slobodan T. Whispered speech recognition using deep denoising autoencoder and inverse filtering. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, v. 25, n. 12, p. 2313–2322, 2017. Citado 2 vezes nas páginas 9 e 21.

HINTON, Geoffrey et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. **IEEE Signal processing magazine**, v. 29, n. 6, p. 82–97, 2012. Citado 2 vezes nas páginas 9 e 18.

ITOH, Taisuke; TAKEDA, Kazuya; ITAKURA, Fumitada. Acoustic analysis and recognition of whispered speech. In: . Orlando, FL, United States: IEEE Workshop on Automatic Speech Recognition and Understanding, 2001. p. 429–432. Citado 3 vezes nas páginas 9, 16 e 18.

JOVIČIĆ, Slobodan T; ŠARIĆ, Zoran. Acoustic analysis of consonants in whispered speech. **Journal of Voice**, v. 22, n. 3, p. 263–274, 2008. Citado na página 9.

KANEKO, Takuhiro et al. Sequence-to-sequence voice conversion with similarity metric learned using generative adversarial networks. In: . Stockholm, Sweden: INTERSPEECH, 2017. Citado na página 9.

KRIZHEVSKY, Alex; SUTSKEVER, Ilya; HINTON, Geoffrey E. Imagenet classification with deep convolutional neural networks. In: . New York, United States: Communications of the ACM, 2017. v. 60, n. 6, p. 84–90. Citado na página 19.

LI, Jingjie; MCLOUGHLIN, Ian Vince; SONG, Yan. Reconstruction of pitch for whisper-to-speech conversion of chinese. In: . Singapore: The 9th International Symposium on Chinese Spoken Language Processing, 2014. p. 206–210. Citado na página 9.

LIU, Zheli et al. GMM and CNN hybrid method for short utterance speaker recognition. **IEEE Transactions on Industrial Informatics**, v. 14, n. 7, p. 3244–3252, 2018. Citado na página 18.

MOLLA, KI; HIROSE, Keikichi. On the effectiveness of MFCCs and their statistical distribution properties in speaker identification. **IEEE Symposium on Virtual Environments, Human-Computer Interfaces and Measurement Systems, 2004.(VCIMS).**, p. 136–141, 2004. Citado na página 14.

MORRIS, Robert W.; CLEMENTS, Mark A. Reconstruction of speech from whispers. **Medical Engineering and Physics**, v. 24, n. 7, p. 515 – 520, 2002. Citado 2 vezes nas páginas 9 e 16.

OLIVEIRA, Marco António da Mota. Modelização de filtro de trato vocal para reconstrução de voz disfónica. Dissertação de Mestrado Integrado, Faculdade de Engenharia da Universidade do Porto, Porto, Portugal, 2020. Citado 3 vezes nas páginas 10, 17 e 25.

OLIVEIRA, Patrícia Cristina Ramalho de. Artificial voicing of whispered speech. Dissertação de Mestrado, Faculdade de Engenharia da Universidade do Porto, Porto, Portugal, 2015. Citado na página 13.

PARMAR, Mihir et al. Effectiveness of cross-domain architectures for whisper-to-normal speech conversion. **2019 27th European Signal Processing Conference (EUSIPCO)**, p. 1–5, 2019. Citado 2 vezes nas páginas 23 e 30.

PASCUAL, Santiago; SERRA, Joan; BONAFONTE, Antonio. Time-domain speech enhancement using generative adversarial networks. **Speech Communication**, v. 114, p. 10–21, 2019. Citado 3 vezes nas páginas 9, 22 e 23.

PRAGER, Richard W; HARRISON, Thomas D; FALLSIDE, Frank. Boltzmann machines for speech recognition. **Computer Speech & Language**, Academic Press Ltd., v. 1, n. 1, p. 3–27, 1986. Citado na página 21.

RABINER, Lawrence; JUANG, Biing-Hwang. **Fundamentals of Speech Recognition**. United States: Prentice-Hall, Inc., 1993. Citado 2 vezes nas páginas 11 e 14.

RABINER, Lawrence; SCHAFER, Ronald. **Theory and applications of digital speech processing**. USA: Prentice Hall Press, 2010. Citado 6 vezes nas páginas 12, 13, 14, 15, 16 e 18.

SEARA, Izabel Christine; NUNES, Vanessa Gonzaga; VOLCÃO, Cristiane Lazzarotto. **Fonética e fonologia do português brasileiro: 2º período**. Florianópolis, SC, Brasil: LLV/CCE/UFSC, 2011. Citado 2 vezes nas páginas 12 e 13.

SHAH, Nirmesh et al. Novel mmse discogan for cross-domain whisper-to-speech conversion. In: . Hyderabad, India: Conference: Machine Learning in Speech and Language Processing (MLSLP), 2018. Citado 5 vezes nas páginas 22, 23, 24, 28 e 29.

SHARIFZADEH, Hamid R. et al. A training-based speech regeneration approach with cascading mapping models. **Computers and Electrical Engineering**, v. 62, p. 601 – 611, 2017. Citado 2 vezes nas páginas 9 e 18.

SHARIFZADEH, Hamid Reza; MCLOUGHLIN, Ian V; AHMADI, Farzaneh. Reconstruction of normal sounding speech for laryngectomy patients through a modified celp codec. **IEEE Transactions on Biomedical Engineering**, v. 57, n. 10, p. 2448–2458, 2010. Citado na página 16.

SHARIFZADEH, Hamid Reza; MCLOUGHLIN, Ian V; RUSSELL, Martin J. A comprehensive vowel space for whispered speech. **Journal of Voice**, v. 26, n. 2, p. e49–e56, 2012. Citado na página 17.

SILVA, Ivan; SPATTI, Danilo; FLAUZINO, Rogério. **Redes neurais artificiais para engenharia e ciências aplicadas**. [S.l.]: Artliber, 2016. Citado 2 vezes nas páginas 19 e 20.

STEVENS, Kenneth N. **Acoustic phonetics**. Cambridge, Mass: MIT press, 2000. Citado 2 vezes nas páginas 11 e 12.

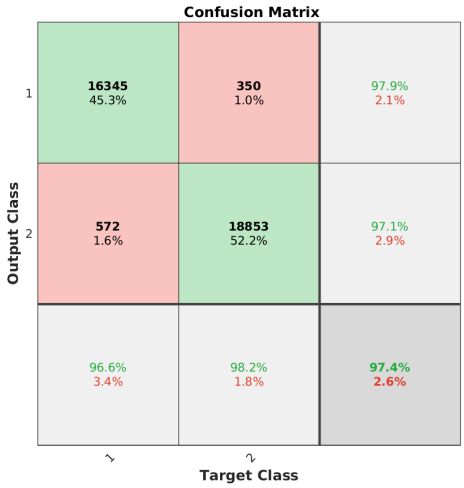
Toda, T.; Nakagiri, M.; Shikano, K. Statistical voice conversion techniques for body-conducted unvoiced speech enhancement. **IEEE Transactions on Audio, Speech, and Language Processing**, v. 20, n. 9, p. 2505–2517, 2012. Citado na página 18.

TYCHTL, Zbynìk; PSUTKA, Josef. Speech production based on the mel-frequency cepstral coefficients. In: . [S.l.]: Sixth European Conference on Speech Communication and Technology, 1999. Citado na página 14.

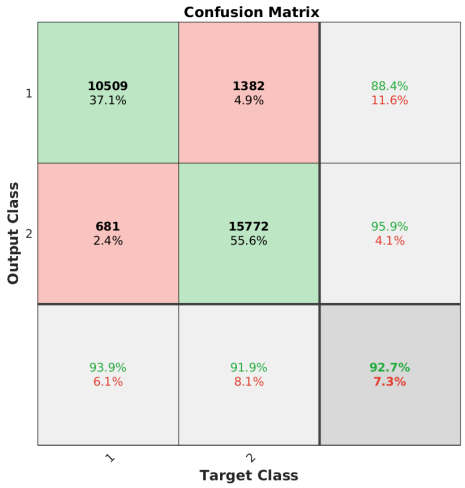
ZHAO, Yuanjun; XIA, Xianjun; TOGNERI, Roberto. Applications of deep learning to audio generation. **IEEE Circuits and Systems Magazine**, v. 19, n. 4, p. 19–38, 2019. Citado 2 vezes nas páginas 22 e 23.

Apêndices

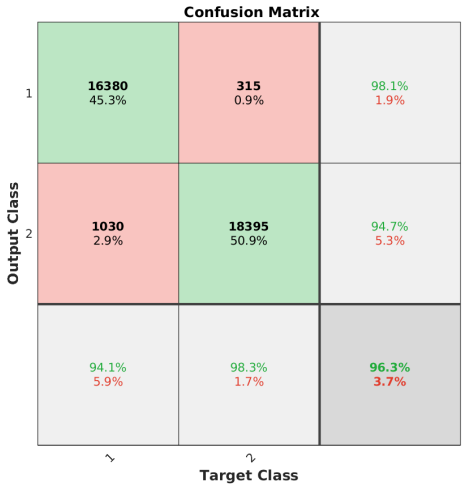
Figura 22 – Matriz de confusão utilizando agrupamento 1 com silêncio.



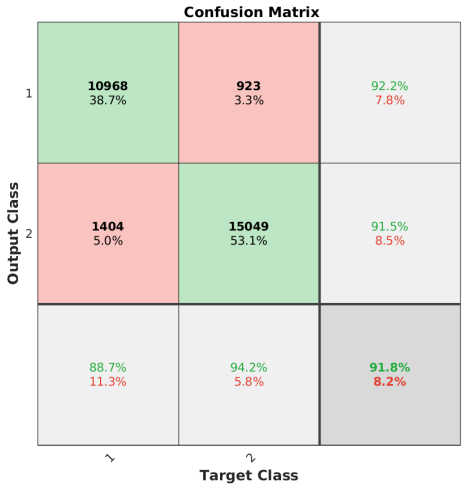
(a) MLP - Masculino



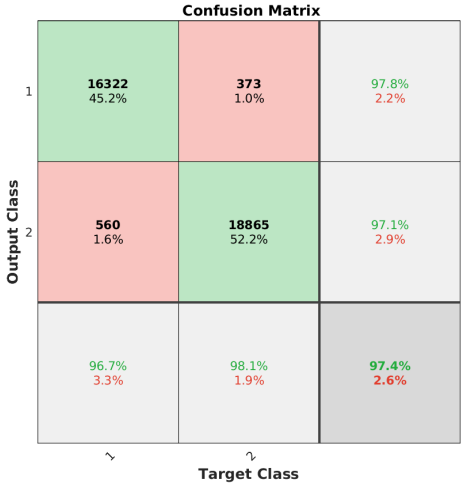
(b) MLP - Feminino



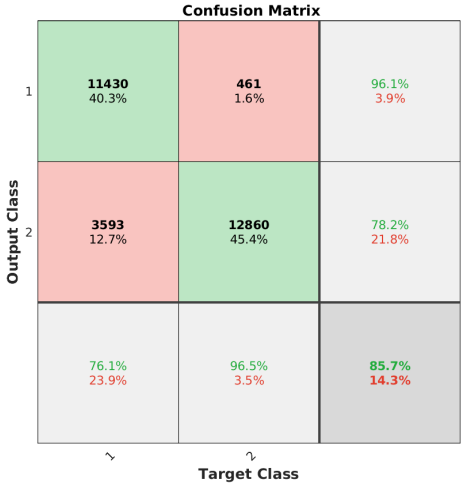
(c) GAN - Masculino



(d) GAN - Feminino



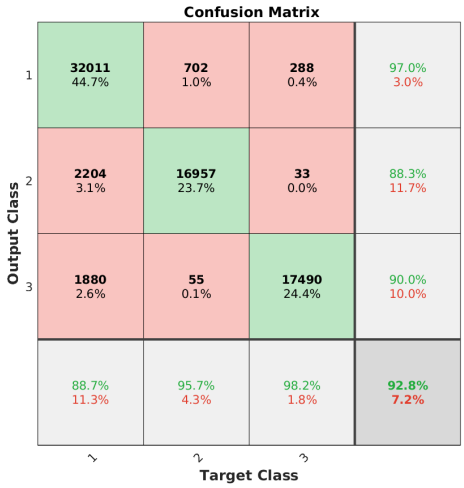
(e) DiscoGAN - Masculino



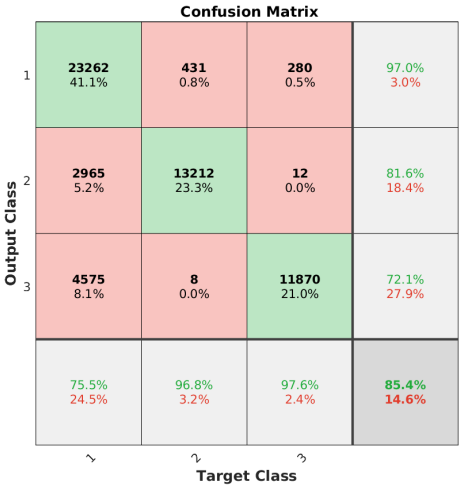
(f) DiscoGAN - Feminino

Fonte: Autoria própria

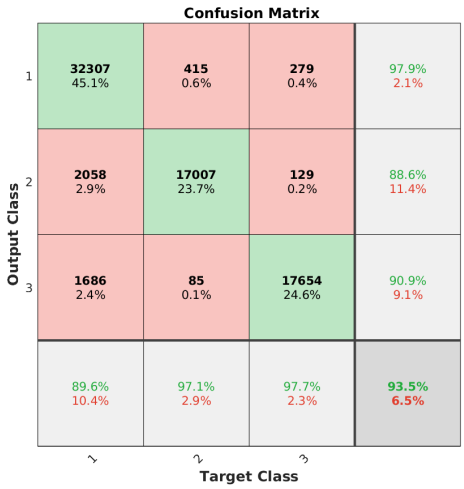
Figura 23 – Matriz de confusão utilizando agrupamento 2 com silêncio.



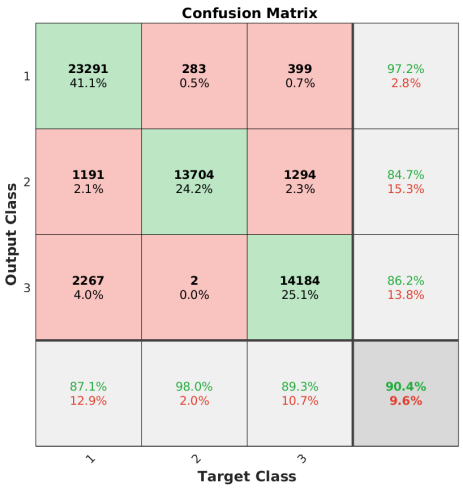
(a) MLP - Masculino



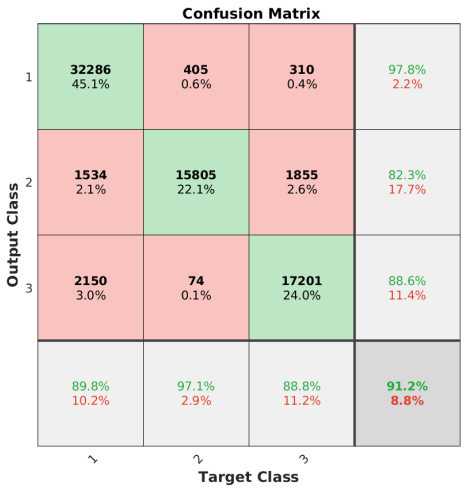
(b) MLP - Feminino



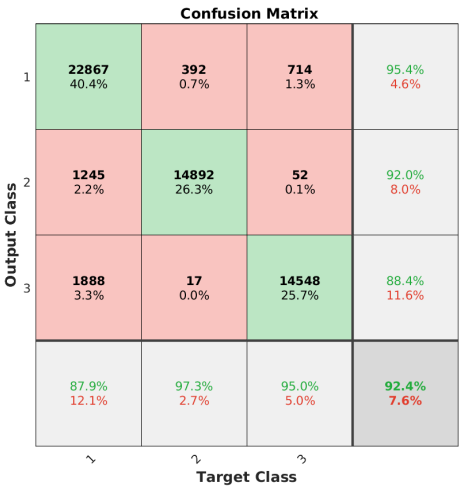
(c) GAN - Masculino



(d) GAN - Feminino



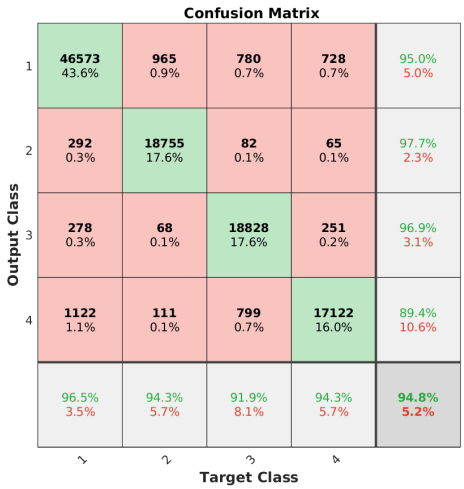
(e) DiscoGAN - Masculino



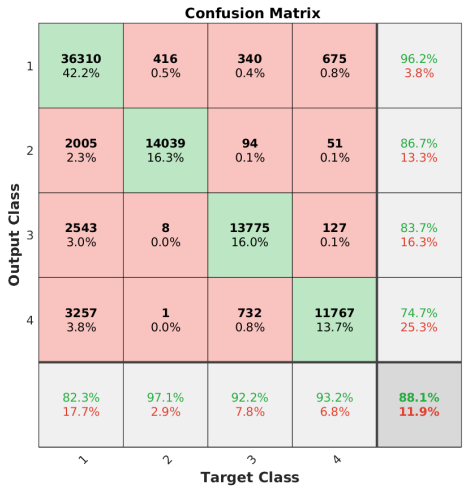
(f) DiscoGAN - Feminino

Fonte: Autoria própria

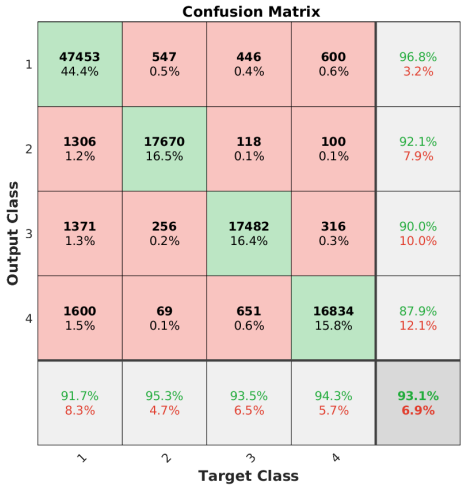
Figura 24 – Matriz de confusão utilizando agrupamento 3 com silêncio.



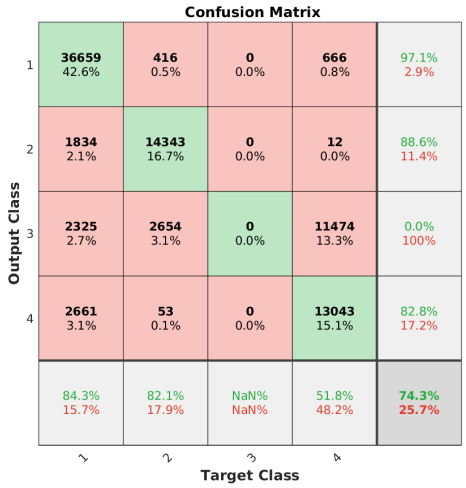
(a) MLP - Masculino



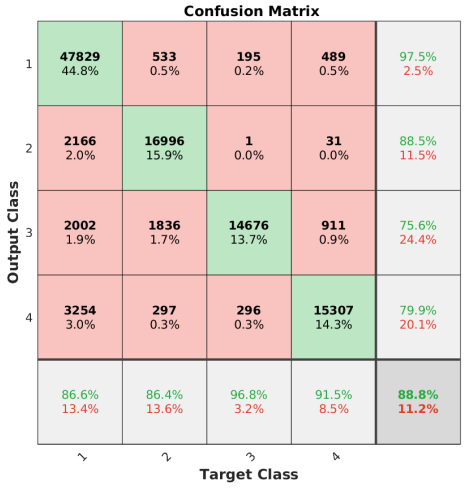
(b) MLP - Feminino



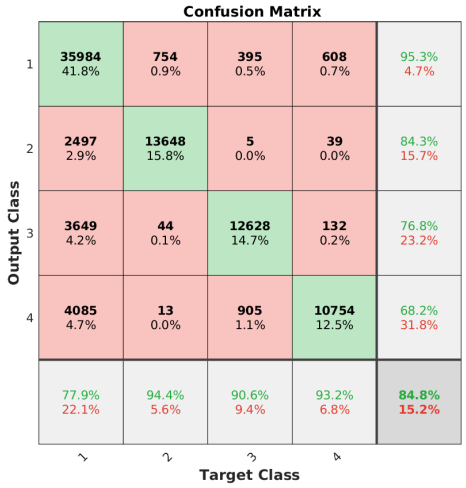
(c) GAN - Masculino



(d) GAN - Feminino



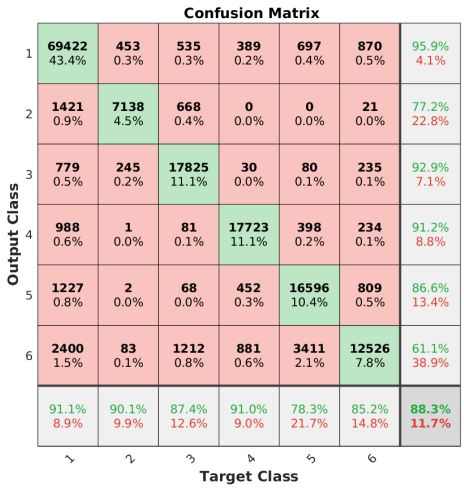
(e) DiscoGAN - Masculino



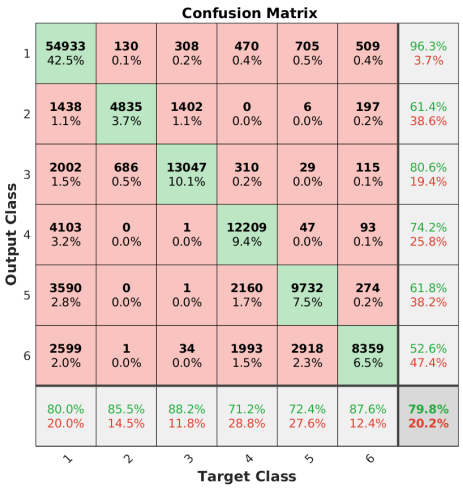
(f) DiscoGAN - Feminino

Fonte: Autoria própria

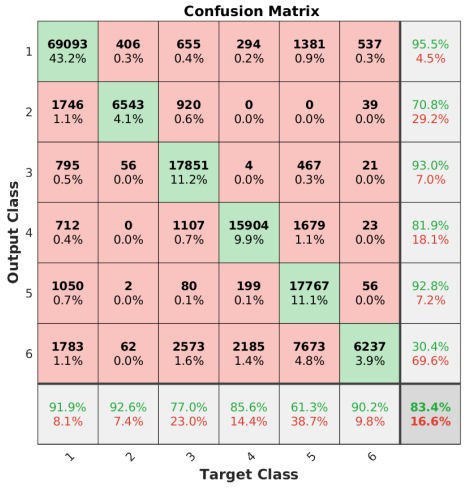
Figura 25 – Matriz de confusão utilizando agrupamento 4 com silêncio.



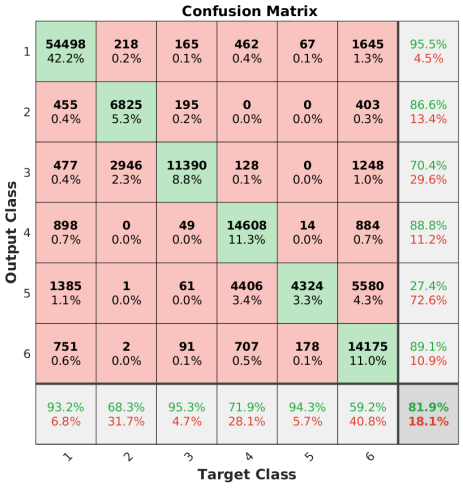
(a) MLP - Masculino



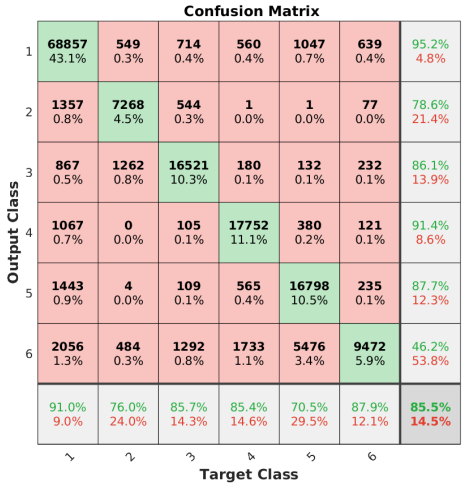
(b) MLP - Feminino



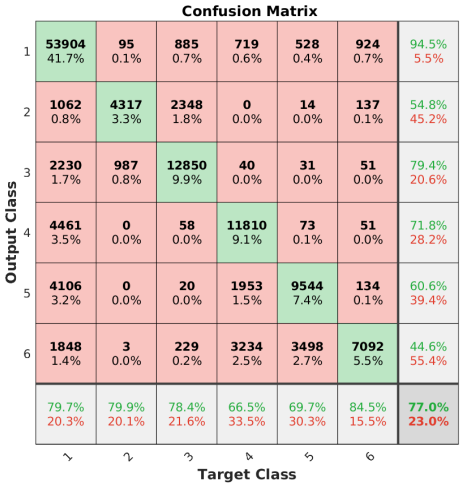
(c) GAN - Masculino



(d) GAN - Feminino



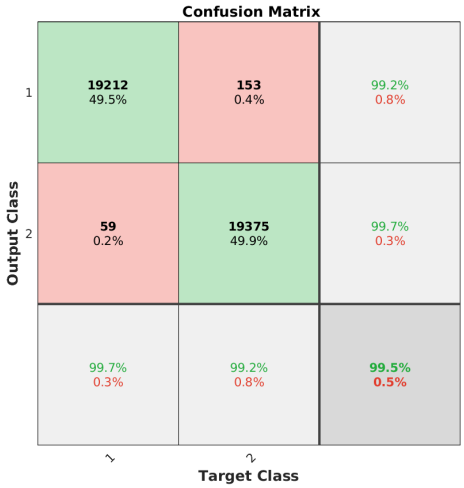
(e) DiscoGAN - Masculino



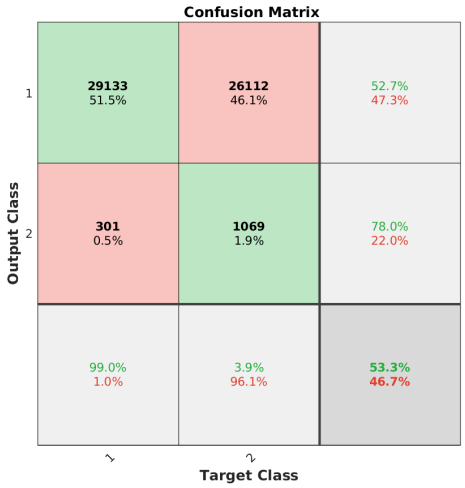
(f) DiscoGAN - Feminino

Fonte: Autoria própria

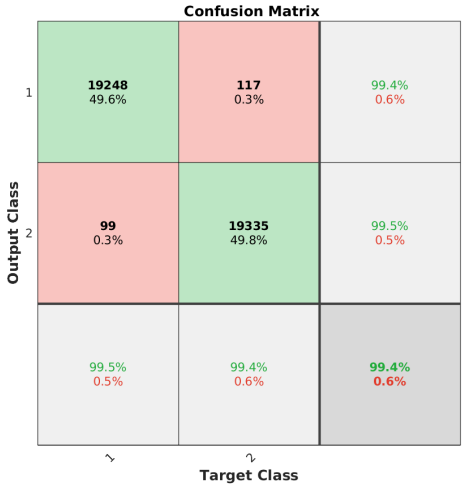
Figura 26 – Matriz de confusão utilizando agrupamento 2 sem silêncio.



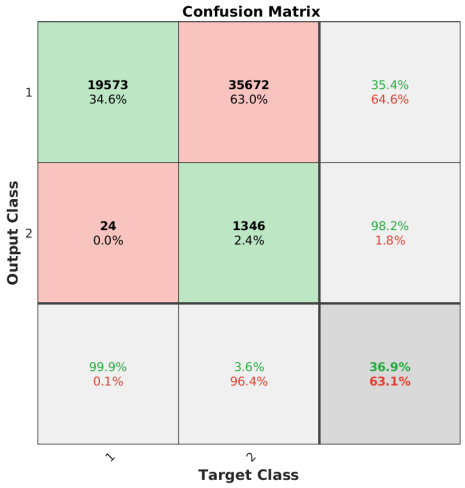
(a) MLP - Masculino



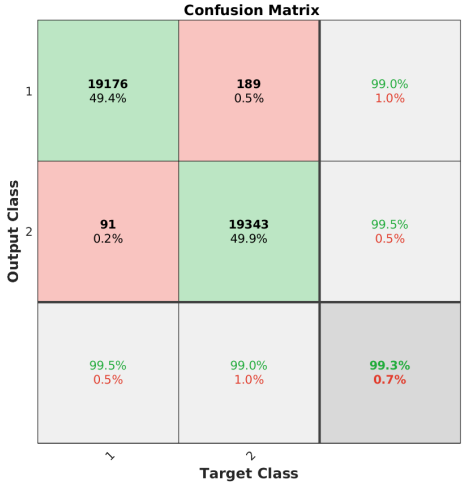
(b) MLP - Feminino



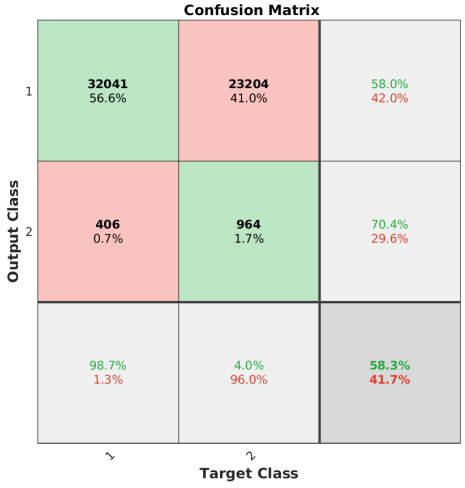
(c) GAN - Masculino



(d) GAN - Feminino



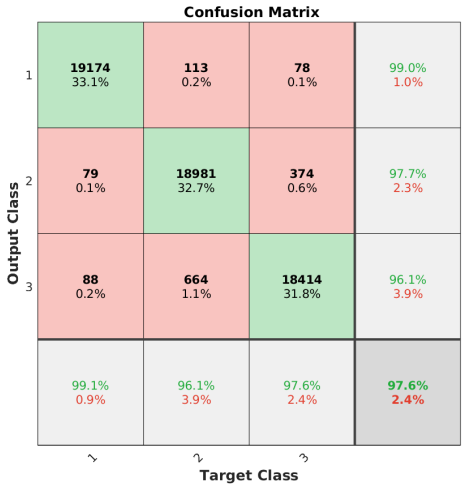
(e) DiscoGAN - Masculino



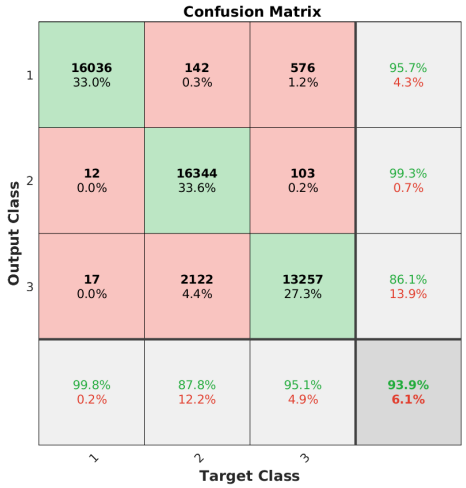
(f) DiscoGAN - Feminino

Fonte: Autoria própria

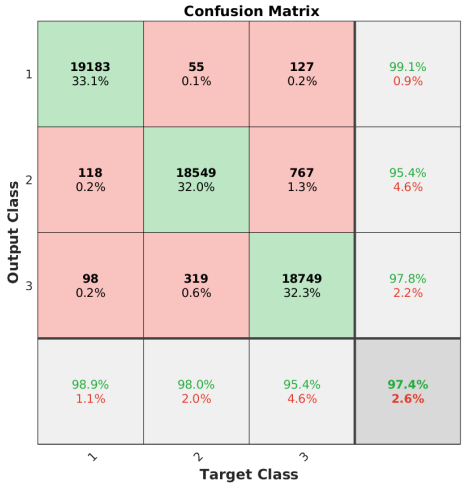
Figura 27 – Matriz de confusão utilizando agrupamento 3 sem silêncio.



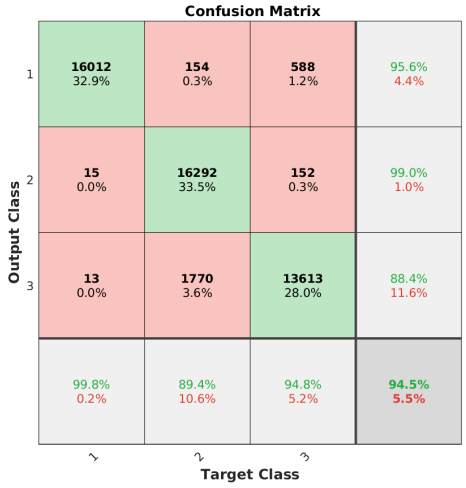
(a) MLP - Masculino



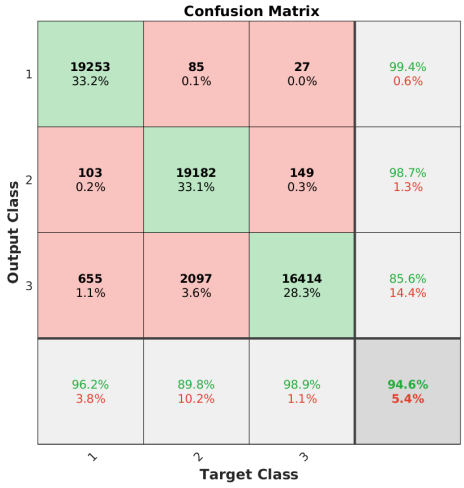
(b) MLP - Feminino



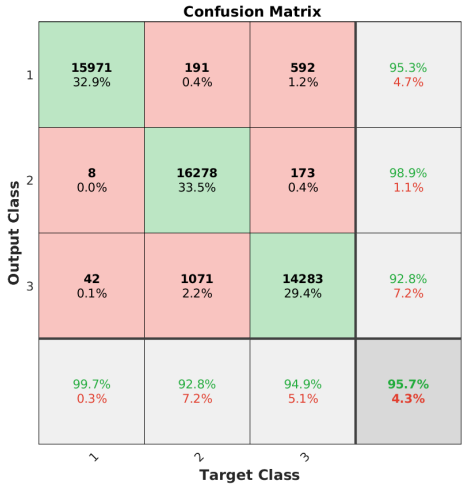
(c) GAN - Masculino



(d) GAN - Feminino



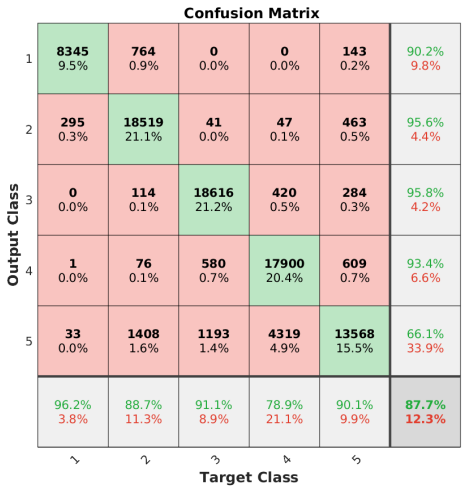
(e) DiscoGAN - Masculino



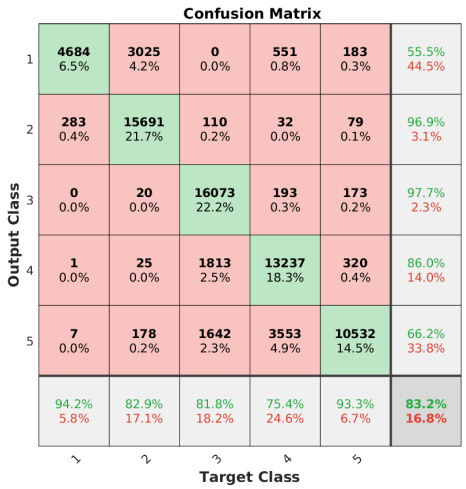
(f) DiscoGAN - Feminino

Fonte: Autoria própria

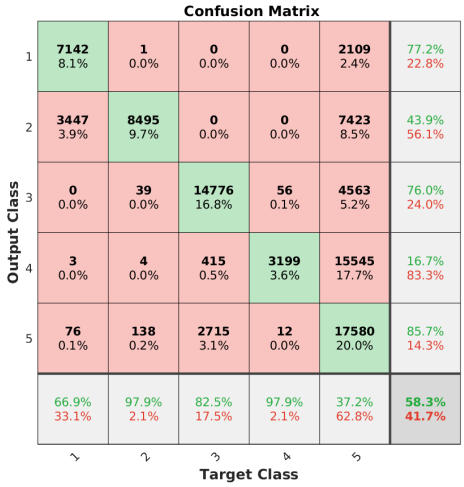
Figura 28 – Matriz de confusão utilizando agrupamento 4 sem silêncio.



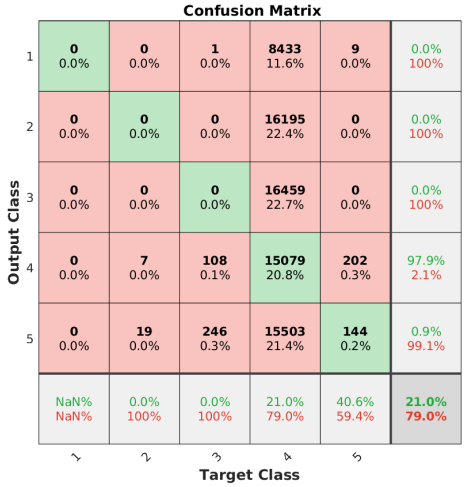
(a) MLP - Masculino



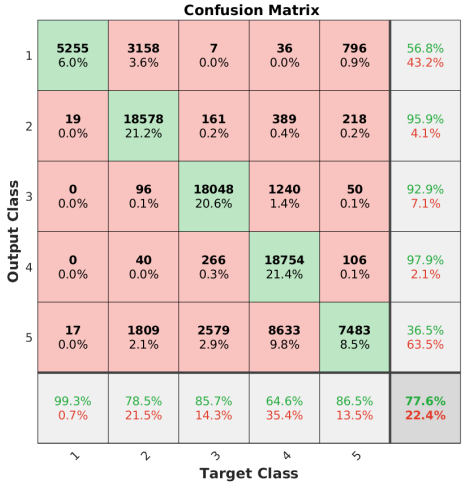
(b) MLP - Feminino



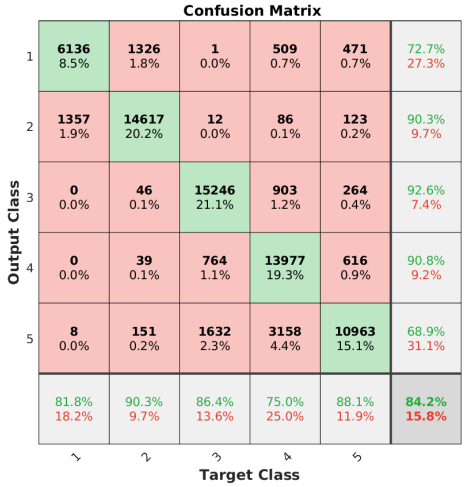
(c) GAN - Masculino



(d) GAN - Feminino



(e) DiscoGAN - Masculino



(f) DiscoGAN - Feminino

Fonte: Autoria própria

Anexos

ANEXO A – REPRODUÇÃO DAS TAREFAS DE GRAVAÇÃO (DYNAVOICER)

Procedimento: Leitura de palavras –dizer cada palavra 3 vezes em fala normal e 3 vezes em fala sussurrada, pela ordem apresentada;

Col. 1	Col. 2	Col. 3	Col. 4	Col. 5	Col. 6
Nuca 14	Ripa 19	Sala 24	Zaro 29	Asa 34	Velho 38
Lupa 15	Laje 20	Juba 25	Assa 30	Rita 35	Hora 39
Chiba 16	Face 21	Chama 26	Luta 31	Haja 36	Buda 40
Minha 17	Pica 22	Vida 27	Jarra 32	Ache 37	Viga 41
Guga 18	Vaze 23	Acha 28	Fisga 33		

Procedimento: Leitura de frases –ler cada frase 3 vezes em fala normal e 3 vezes em fala sussurrada;

A Marta e o avô vivem naquele casarão rosa velho.

Sofia saiu cedo da sala.

A asa do avião andava avariada.

Agora é hora de acabar.

A minha mãe mandou-me embora.

O Tiago comeu quatro peras.

Procedimento: Leitura do texto 1 vez em fala normal e 1 vez em fala sussurrada