UNIVERSIDADE FEDERAL DO PARÁ

INSTITUTO DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

Imitação da Voz Humana através do Processo de Análise-por-Síntese utilizando Algoritmo Genético e Sintetizador de Voz por Formantes

Fabíola Pantoja Oliveira Araújo

Orientador:

Prof. Dr. Aldebaro Barreto da Rocha Klautau Júnior

TD: 18/2015

UFPA / ITEC / PPGEE

Campus Universitário do Guamá

Belém-Pará-Brasil

2015

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

Imitação da Voz Humana através do Processo de Análise-por-Síntese utilizando Algoritmo Genético e Sintetizador de Voz por Formantes

Fabíola Pantoja Oliveira Araújo

Orientador:
Prof. Dr. Aldebaro Barreto da Rocha Klautau Júnior

Tese de Doutorado submetida à Banca Examinadora do Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal do Pará como pré-requisito para obtenção do título de "Doutor em Engenharia Elétrica com ênfase em Computação Aplicada".

UFPA / ITEC / PPGEE
Campus Universitário do Guamá
Belém-Pará-Brasil
2015

Dados Internacionais de Catalogação-na-Publicação (CIP)
Sistema de Bibliotecas da UFPA

Araújo, Fabíola Pantoja Oliveira, 1975-
Imitação da voz humana através do processo de análise-por-síntese utilizando algoritmo genético e sintetizador de voz por formantes / Fabíola Pantoja Oliveira Araújo. - 2015.

Orientador: Aldebaro Barreto da Rocha Klautau Júnior.
Tese (Doutorado) - Universidade Federal do Pará, Instituto de Tecnologia, Programa de Pós-Graduação em Engenharia Elétrica, Belém, 2015.

1. Síntese da voz. 2. Sistemas de processamento da fala. 3. Algoritmos genéticos. I. Título.

CDD 22. ed. 006.454

Imitação da Voz Humana através do Processo de Análise-por-Síntese utilizando Algoritmo Genético e Sintetizador de Voz por Formantes

Fabíola Pantoja Oliveira Araújo

Orientador: Prof. Dr. Aldebaro Barreto da Rocha Klautau Júnior

Banca examinadora

.....

Prof. Dr. Aldebaro Barreto da Rocha Klautau Júnior (UFPA) - Orientador

.....

Prof. Dr. Gustavo Augusto Lima de Campos (UECE) - Membro externo

.....

Prof. Dr. Antonio Marcos Lima de Araújo (IFPA/IESAM) - Membro externo

.....

Prof. Dr. Eloi Luiz Favero (UFPA) - Membro

.....

Prof. Dr. Roberto Célio Limão de Oliveira (UFPA) - Membro

.....

Prof. Dr. Glaucio Haroldo Silva de Carvalho (UFPA) - Membro

Visto:

.....

Prof. Dr. Evaldo Gonçalves Pelaes

Coordenador do PPGEE/ITEC/UFPA

À minha família com todo AMOR ...

AGRADECIMENTOS

À Deus e aos amigos espirituais por me auxiliarem a ter a tranquilidade necessária, sem emorecer perante os obstáculos.

Ao meu orientador, Aldebaro, por acreditar em mim quando eu mesma não acreditava que teria potencial para concluir este trabalho. Meus mais sinceros e profundos agradecimentos pela dedicação, infinita paciência e principalmente, pela contribuição profissional que permitiram o desenvolvimento deste.

À minha família, em especial ao meu esposo, Josivaldo, por todo apoio e incentivo principalmente nos momentos mais difíceis, e à minha filha Manuela, flor mais linda e doce do meu jardim.

Aos meus pais, Nadya e Franklin (*sempre presente*), por me incentivarem a estudar e a crescer profissionalmente e pessoalmente.

À minha amada "vozinha" Josélia por todo amor, carinho e apoio de sempre.

Às minhas irmãs, Danièlle e Josélia, pela amizade mais pura, sincera e verdadeira que tenho.

A todos os amigos que me acompanharam nessa longa jornada, em especial aos amigos e companheiros do Laboratório de Processamento de Sinais (LaPS) sem os quais não seria possível a conclusão deste trabalho.

Ao Programa de Pós-Graduação em Engenharia Elétrica (PPGEE) da Universidade Federal do Pará por possibilitar o desenvolvimento deste.

O olho vê, a lembrança revê, e a imaginação transvê.
É preciso transver o mundo.

Manoel de Barros

RESUMO

A imitação da voz através do mecanismo de *utterance copy* consiste em estimar os parâmetros de entrada de um sintetizador de voz para gerar um sinal parecido com o da voz original. Este processo distingue-se da tradicional conversão texto-fala, porém é usado em muitas áreas, especialmente, em Linguística e na Saúde. Imitar a voz humana através deste mecanismo é um problema inverso difícil, pois este mapeamento é não linear e de muitos para um. Por exemplo, existem diferentes combinações dos valores dos parâmetros de entrada do sintetizador que produzem o mesmo sinal de voz sintética. Sendo assim, realizar manualmente a imitação da voz requer uma quantidade considerável de tempo e métodos automáticos, como o proposto aqui, são de interesse. Este trabalho apresenta um arcabouço baseado em algoritmo genético (AG) para estimar automaticamente os valores dos parâmetros de entrada de um sintetizador de voz por formantes, utilizando o processo de análise-por-síntese. Os resultados apresentados compreendem a imitação de vozes sintéticas (geradas por computador) e naturais (geradas por humanos) em inglês americano, para falantes masculinos e femininos. Estes resultados são comparados com os obtidos através do Winsnoori (baseline), o único software disponível atualmente que executa a mesma tarefa. Os experimentos mostraram que o arcabouço desenvolvido (*newGASpeech*) é uma alternativa eficaz para o trabalhoso processo manual de estimar os valores dos parâmetros de entrada de um sintetizador por formantes, superando a qualidade das vozes geradas pelo baseline em relação à cinco métricas objetivas utilizadas e à avaliação subjetiva aplicada a vinte e sete ouvintes não especialistas na área de voz e nem no idioma adotado.

Palavras-chave: imitação da voz; algoritmo genético; análise-por-síntese; sintetizador por formantes.

ABSTRACT

Voice imitation through the utterance copy mechanism is estimating the value of the input parameters of a speech synthesizer to generate a similar signal with the original voice. This process is distinct from the more traditional text-to-speech, but yet used in many areas, especially, Linguistics and Health System. Imitate the human voice through this mechanism is a difficult inverse problem because the mapping is non-linear and from many to one. For instance, there are different combinations of the synthesizer input parameters values that produce the same synthetic voice signal. Therefore, perform voice imitation manually requires a considerable amount of time. In addition to automatic methods are our interest of study as well, as proposed here. This work presents our system based on Genetic Algorithm (GA) to automatically estimate the value of the input parameters of a speech formant synthesizer using the analysis-by-synthesis process. Results are presented for synthetic (computer-generated) and natural (human-generated) speech in American English, for male and female speakers. These results are compared with the ones obtained with Winsnoori, the only currently available software that performs the same task. The experiments showed that the proposed *newGASpeech* framework is an effective alternative to the laborious manual process of estimating the input parameters values of a formant synthesizer. Besides it has overcome the quality of the generated voices by the baseline if compared to five objective metrics and a subjective evaluation applied to twenty seven no-expert listeners in the speech area neither the adopted language.

Key words: voice imitation; genetic algorithm; analysis-by-synthesis; formant synthesizer.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AG - Algoritmo Genético

AGMO - Algoritmo Genético Multi-Objetivo

CC - Correlação Cruzada

CCR - *Comparison Category Rating*

CE - Computação Evolucionária

D_{LE} - Distância Espectral

DTFT - *Discrete-Time Fourier Transform*

EQM - Erro Quadrático Médio

EM - *Expectation Maximization*

EP - Erro percentual

FFT - *Fast Fourier Transform*

HTK - *Hidden Markov Models Toolkit*

HTS - *Hidden Markov Models based Speech Synthesis System*

KLSYN - *Klatt Synthesizer*

KLSYN88 - *Klatt Synthesizer version 88*

LAPS - Laboratório de Processamento de Sinais

LF - *Liljencrants-Fant*

LisTEN - *LIStenning Test ENvironment*

MOS - *Mean Opinion Score*

NSGA-II - *Non-Dominated Sorting Algorithm II*

OMO - Otimização Multi-Objetivo

PDS - Processamento Digital de Sinal

PESQ - *Perceptual Evaluation of Speech Quality*

PLN - Processamento de Linguagem Natural

RSR - Relação Sinal-Ruído

SBX - *Simulated Binary Crossover*

STS - *Speech-To-Speech*

TIMIT - *Texas Instruments - Massachussets Institute os Technology*

TTS - *Text-to-Speech*

VHLSyn - *Very High-Level Synthesis*

VODER - *Voice Operating Demonstrator*

SUMÁRIO

Capítulo 1

Introducão

1.1 Motivação e descrição geral do problema

Estimar os parâmetros de entrada de um sintetizador de voz para reconstruir um sinal e imitar uma voz alvo é um processo utilizado em muitas áreas do conhecimento, especialmente, na Linguística e Saúde, atraindo, assim, tanto interesses comerciais quanto aplicações clínicas. Como exemplos, tem-se desde a criação de uma versão virtual da voz de uma pessoa até a produção artificial das vozes de pacientes que não podem falar, normalmente, devido à trauma, doença ou cirurgia [Bangayan et al., 1997,Kain et al., 2004,Fraj et al., 2012]. Entretanto, este é um problema inverso difícil, pois o mapeamento é não linear e de muitos para um, por exemplo, existem diferentes combinações dos parâmetros de entrada do sintetizador que conduzem ao mesmo sinal de voz. Portanto, a configuração dos parâmetros de entrada para imitar uma determinada voz requer uma quantidade considerável de tempo se for realizada manualmente. Uma alternativa para isto é estimar automaticamente esses parâmetros para alimentar um sintetizador, utilizando para isso apenas um único sinal de voz de entrada.

O Klatt, anos 1980 e 1990, é um sintetizador de voz é baseado em formantes e adotado em vários estudos (como [Bangayan et al., 1997, Jinachitra and Smith III, 2005]), pois seus parâmetros de entrada que estão intimamente relacionados com os parâmetros físicos da produção da fala. Isso leva à um alto grau de interpretabilidade, essencial em alguns estudos sobre a correlação entre a acústica e a qualidade da voz, tais como: na conversão masculino/-feminino e simulações de sussurros, roquidão e sons chiados de baixa sonoridade aguda. Este sintetizador já foi usado para imitar voz natural (Anumanchipalli, 2010; Shrivastav, 2006) assim como vozes patológicas [Bangayan et al., 1997]. Existem outras técnicas de síntese de

voz assim como métodos para a obtenção dos parâmetros de entrada de um sintetizador [Liu and Kewley-Port, 2004]. No entanto, a interpretação do papel de cada parâmetro de entrada não é tão fácil como no Klatt e com isso sintetizadores alternativos parecem menos populares. O *HLSyn* é um outro sintetizador constituído por formantes que funciona como uma camada acima do Klatt, permitindo, assim, uma redução na quantidade de parâmetros utilizados para sintetizar voz. Esses dois sintetizadores mencionados foram utilizados neste trabalho. Os trabalhos correlatos citados anteriormente são apresentados na Seção 1.4.

Dessa forma, o presente trabalho consiste em estimar automaticamente os valores dos parâmetros que compõem o arquivo de entrada de um sintetizador por formantes, como o Klatt e o *HLSyn*, utilizando algoritmo genético (AG). O objetivo principal é imitar uma voz, sendo necessário encontrar a combinação de valores dos parâmetros que levem a uma voz sintética parecida o suficiente com uma voz alvo (*Speech-To-Speech system*-STS) que pode ser tanto natural quanto sintética, esta última obtida através de um sistema texto-fala (*TTS - Text-To-Speech system*). De acordo com as vozes sintéticas geradas a partir do TTS Dectalk, 25 parâmetros da versão do Klatt utilizada (KLSYN88) variam e devem ser combinados para produzir voz sendo que cada um deles tem um intervalo de valores aceitável. Para o sintetizador *HLSyn*, 13 parâmetros são combinados para produzir uma determinado trecho de voz. A tarefa de combinar valores de parâmetros pode ser considerada um problema inverso difícil devido à variedade de combinações possíveis de valores.

A Figura 1.1 ilustra o objetivo deste trabalho: a partir de um arquivo de voz dado como entrada em um sistema STS, um modelo inverso gera a combinação de parâmetros de entrada do sintetizador de voz por formantes que imita essa voz. Essa combinação de parâmetros é submetida ao sintetizador, produzindo uma voz sintética que imite a voz alvo (entrada).

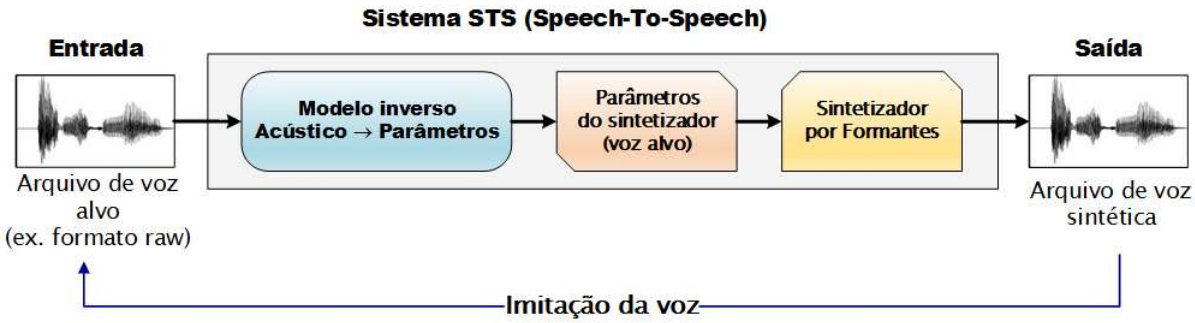

Figura 1.1: Sistema para imitação da voz.

Nas seções seguintes, são apresentadas a metodologia empregada, os objetivos e contribuições deste trabalho, assim como os trabalhos correlatos e as publicações realizadas durante

o período de desenvolvimento deste.

1.2 Metodologia e objetivos

Este trabalho concentra-se em imitar uma determinada voz utilizando para isso algoritmo genético e sintetizador de voz por formantes. Para isso, foi necessária a construção de um arcabouço baseado no processo de análise-por-síntese em que os valores dos parâmetros do sintetizador são melhorados à cada iteração por meio de uma avaliação do sinal de voz sintético produzido a partir da combinação dos parâmetros estimados. Porém, para se construir este arcabouço, foi preparado primeiramente um corpus de voz sintética foneticamente balanceado e heterogêneo (falantes masculinos e femininos) com o objetivo de estudar os parâmetros do Klatt e reduzir a dimensionalidade do problema. Além disso, houve a necessidade de escolher métricas adequadas para avaliar o sinal de voz sintético em tempo de execução do AG, uma vez que a melhora do sinal ocorre gradualmente. O estudo das métricas proporcionou a definição de uma metodologia de avaliação da voz sintética produzida ao final, composta por métricas objetivas e uma avaliação subjetiva. As vozes produzidas através do arcabouço foram comparadas com àquelas geradas através do Winsnoori. Este software foi definido como baseline por ser o único que realiza a imitação da voz, porém utilizando uma outra versão do sintetizador de Klatt. Devido à característica dos sintetizadores adotados em gerar vozes sintéticas em inglês americano, os experimentos abrangeram vozes alvo masculinas e femininas neste idioma. Os experimentos iniciais foram realizados com vozes alvo sintéticas, produzidas para 6 falantes (3 masculinos e 3 femininos) através do TTS *Dectalk*. Os valores dos parâmetros para essas vozes foram previamente conhecidos, sendo possível avaliar a convergência dos parâmetros para valores ótimos. Devido aos resultados satisfatórios nos experimentos controlados, as simulações com voz natural foram importantes para que o arcabouço fosse avaliado em relação à imitação da voz de falantes desconhecidos. Para validar os resultados obtidos na avaliação objetiva realizada nas vozes sintéticas geradas à partir das vozes naturais, foi feita uma avaliação subjetiva em que 27 ouvintes não especialistas na língua inglesa utilizaram um software para atribuir uma nota à voz avaliada, sendo esta nota baseada no método *Comparison Category Rating*. Após a decrição da metodologia empregada e dos objetivos alcançados, a próxima seção lista as principais contribuições desta tese.

1.3 Contribuições da Tese

Este trabalho possui como principal contribuição o desenvolvimento de um arcabouço baseado em algoritmo genético para estimar os parâmetros de sintetizadores por formantes, por exemplo, o Klatt e o *HLSyn*, com o objetivo de imitar uma determinada voz (sintética ou natural), independente de falante, utilizando o processo de análise-por-síntese e sem a necessidade de treinamento prévio do modelo ou de um corpus de voz. O arcabouço encontra-se gratuitamente disponível em [estimation of Klatt parameters, 2015], sendo uma alternativa melhor do que o sotware utilizado como baseline, pois este foi descontinuado pelo autor. Para avaliar a voz sintética produzida, foi desenvolvida uma metodologia de avaliação composta por métricas objetivas e teste subjetivo. Na avaliação objetiva, cinco figuras de mérito foram utilizadas sendo que uma delas permite à atribuição de uma nota à voz sem compará-la com a voz alvo , portanto, sem a necessidade de realizar o alinhamento entre os sinais alvo e sintetizado. A avaliação subjetiva foi adaptada de [P80, 1996] e aplicada somente às vozes sintéticas produzidas a partir de vozes alvo naturais. Esta avaliação se fez necessária devido ao fato desses experimentos não serem controlados , portanto, os parâmetros estimados não serem conhecidos previamente pois os falantes eram desconhecidos. Com o objetivo de restringir a dimensionalidade do espaço de busca do arcabouço, sem impactar na qualidade da voz sintética gerada como saída, foi realizado um estudo sobre os parâmetros do sintetizador de Klatt (versão KLSYN88). Para isso, foram gerados arquivos de voz através do TTS *Dectalk* para seis falantes diferentes (masculinos e femininos) utilizando frases foneticamente balanceadas. Nessa direção, foi possível verificar, por exemplo, que alguns parâmetros apresentaram valores fora da faixa de valores indicada na literatura, o sintetizador de Klatt, 1990, além de parâmetros que obtiveram valores constantes independente de falante e do tipo de quadro (sonoro ou não sonoro). A seção seguinte apresenta os trabalhos correlatos à esta tese.

1.4 Estado da Arte

O sintetizador de Klatt é amplamente utilizado em trabalhos com o objetivo de imitar voz através da síntese ou reconstruir o sinal utilizando a resíntese de uma determinada voz, seja ela natural, sintética ou patológica. Esta seção apresenta o levantamento bibliográfico das principais publicações envolvendo este sintetizador e a imitação da voz humana.

A estimação dos parâmetros do Klatt com o objetivo de reconstruir uma voz gravada sob condições moderadas de ruído é apresentada em [Jinachitra and Smith III, 2005]. Neste

trabalho, emprega-se um algoritmo EM (*Expectation-Maximization*) para estimar o modelo dos parâmetros em relação à máxima verossimilhança utilizando filtro de Kalman como suavizador na etapa de expectativa do referido algoritmo. Neste caso, a fonte de voz utilizada foi a *Rosenberg-Klatt* a qual é uma versão simplificada e derivada do modelo *Liljencrants-Fant*, com um filtro para modelar a produção da voz. O algoritmo foi aplicado em vozes masculinas cantadas do fonema */a/* com uma frequência fundamental em torno de 123Hz. Os sons resintetizados apresentaram certa similaridade e soaram naturais apesar de não serem exatamente iguais aos sons originais. Os mesmos experimentos adicionando um ligeiro ruído de 20dB resultaram em sons com aspectos não tão naturais em relação aos primeiros experimentos.

O sintetizador de Klatt foi utilizado nos trabalhos de [Anumanchipalli et al., 2010] e [Bangayan et al., 1997] para gerar voz sintética o mais próximo possível da voz alvo. Em [Anumanchipalli et al., 2010], a frequência e largura de banda das formantes do Klatt foram obtidas através do toolkit *ESPS* enquanto os coeficientes nasais, fricativos e aspirativos foram extraídos a partir do sinal de voz utilizando um modelo de mistura de gaussianas, no qual é possível identificar quando esses fenômenos articulatórios estão presentes. Os demais parâmetros como *gain*, *skew* e *aturb* foram ajustados empiricamente. Os valores dos parâmetros obtidos da base de voz *Artic rms* foram empregados para o treino do arcabouço chamado *Clustergen*, o qual realiza a síntese de voz estatístico-paramétrica [Black, 2006]. Os experimentos foram realizados com o objetivo de gerar os parâmetros do Klatt para reconstruir um sinal de voz e produzir um arquivo de voz a partir de um texto (TTS).

Os resultados mostraram que apesar da qualidade das vozes geradas serem boas, elas apresentaram algum tipo de distorção. De acordo com [Bangayan et al., 1997], o objetivo foi utilizar o Klatt para imitar vozes com patologias variando a intensidade de moderada à severa. Neste caso, o Klatt foi escolhido devido à flexibilidade em produzir vozes como sonoridade fora do normal. Os experimentos foram realizados com vinte e quatro arquivos de voz referentes à pronúncia da vogal */a/*, todos com características de ter alguma patologia. Para cada um desses arquivos de voz, foram geradas vozes sintéticas com as características patológicas da voz natural (alvo) através da modificação dos parâmetros F0, AV, OQ, SQ, TL, FL, DI e AH. Após isso, as vozes sintéticas foram classificadas e uma nota foi atribuída de 1 (voz similar à normal) a 6 (voz com patologia extremamente severa). Para avaliar a qualidade da voz patológica sintética, dez ouvintes foram escolhidos entre otorrinolaringologistas, patologistas especialistas em problemas da fala e foneticistas.

Os ouvintes escutaram os pares de voz patológica (natural e sintética) e tiveram que atribuir uma nota de avaliação em uma escala de sete pontos sendo que 1 indicava que a qualidade da voz sintética era igual ao da natural. Todos os ouvinte foram unânimes em

relatar que estavam satisfeitos com a qualidade das vozes sintéticas sendo que aquelas com patologias severas para vozes femininas foram as que apresentaram menor similaridade com a voz natural. Esse resultado foi devido ao fato do Klatt possuir uma fonte de voz mais adequada para produzir vozes masculinas e da dificuldade em si deste sintetizador em gerar vozes com uma perturbação significativa na amplitude e/ou frequência.

Shrivastav [Shrivastav and Sapienza, 2006] avaliou a qualidade da voz em relação as variações sutis de alguns parâmetros do Klatt que fazem com que a voz apresente soprosidade (sussurros). Neste caso, a manipulação da soprosidade na voz foi realizada através da variação do ruído aspirativo e a voz sintética foi avaliada considerando a relação sinal-ruído (*Signal-to-Noise Ratio* - SNR). No Klatt, esse ruído é representado através do parâmetro AH e nos experimentos deste trabalho o referido parâmetro foi variado de um em um entre 30 e 70dB. Parâmetros da fonte de voz como o quoeficiente de abertura da glote (OQ) e inclinação espectral (TL) afetam a soprosidade da voz porém não foram considerados neste trabalho porque a fonte de voz utilizada foi o modelo de *Liljencrants-Fant*. Isto mostra a flexibilidade do sintetizador de Klatt em produzir vozes com características específicas devido à facilidade em manipular parâmetros específicos na produção da fala.

Os experimentos realizados produziram sinteticamente seis instâncias da vogal */a/* para seis falantes diferentes, 3 masculinos e 3 femininos. Os valores da frequência fundamental e das três primeiras formantes (F0, F1 à F3) foram escolhidos aleatoriamente de uma base de voz disfônica. Os demais parâmetros foram ajustados com o objetivo de produzir uma voz sintética o mais próxima possível à natural. As vozes geradas através dos experimentos foram submetidos a testes com ouvintes capazes de identificar a presença da soprosidade. Os resultados mostraram que os ouvintes precisaram de um aumento de 20dB no ruído aspirativo para identificar a soprosidade em vozes que apresentavam pouco desta característica. Em contrapartida, quando as vozes apresentaram grande soprosidade, o aumento de apenas 11dB no parâmetro AH já permitiu identificar a presença do sussurro.

Existem técnicas de síntese de voz que aliam vários métodos para a obtenção dos parâmetros de entrada do sintetizador de Klatt, como o trabalho de Chang [Liu and Kewley-Port, 2004]. O objetivo deste trabalho foi resintetizar vozes através do arcabouço chamado *Straight*, eliminando a interferência da periodicidade na voz natural. Para os experimentos, a vogal */e/* isolada foi obtida através da palavra bed sintetizada pelo Klatt, utilizando uma voz feminina. As formantes foram manipuladas para esta vogal através do incremento de seus valores. Após isso, a vogal mencionada foi resintetizada através do *Straight* com o objetivo de suavizar a trajetória das formantes e realizar um estudo discriminativo em relação as modificações realizadas (incremento de valores) e o impacto disto em gerar voz sintética o

mais próximo possível da voz natural. As vozes sintéticas foram avaliadas por quatro ouvintes e notou-se que, ao resintetizar as vozes, o arcabouço adicionou ruído ao sinal de voz final. Identificou- se que isto aconteceu devido ao F0 ter um valor alto (usualmente o valor default). Ao atribuir um valor menor para este parâmetro, eliminou-se o ruído no sinal de saída.

Nos trabalhos correlatos citados nesta seção, alguns tentam melhorar a qualidade do sinal de voz, utilizando para isso a resíntese através do Klatt [Jinachitra and Smith III, 2005] e [Shrivastav and Sapienza, 2006]; enquanto outros fazem o ajuste de certos parâmetros deste mesmo sintetizador para gerar vozes patológicas [Shrivastav and Sapienza, 2006] [Bangayan et al., 1997]. Anumanchipalli [Anumanchipalli et al., 2010] teve como objetivo gerar vozes para falantes específicos através de um TTS, utilizando para isso, a síntese paramétrica a qual necessita de um treinamento prévio do modelo baseado em Modelos Ocultos de Markov. O trabalho aqui apresentado difere dos demais, pois realiza a imitação de vozes saudáveis, utilizando AG e tendo como entrada apenas o arquivo de voz que se deseja imitar (alvo). Portanto, este trabalho é um arcabouço intitulado STS que gera uma voz sintética a partir da imitação de uma voz alvo, que pode ser natural ou não, sem a necessidade de treinar algum modelo previamente com um corpus de voz.

1.5 Publicações realizadas durante o período do doutorado

As publicações abaixo relacionadas são diretamente relacionadas ao desenvolvimento da tese.

- ARAÚJO, F.; KLAUTAU, A.; SOUSA, J. Utterance copy through analysis-by-synthesis using genetic algorithm. Journal of Bazilian Computer Society, 2015.
- SOUSA, J.; ARAÚJO, F.; KLAUTAU, A. Utterance Copy for Klatt's Speech Synthesizer using Genetic Algorithm. IEEE Spoken Language Technology Workshop (SLT), 2014, p. 89-94.
- ARAÚJO, F.; KLAUTAU, A. B. R. J. Utterance Copy Through Analisys-by-Synthesis Using Genetic Algorithm. Proceedings of International Telecommunication Symposium - ITS, 2014, p. 1-5.
- TRINDADE, J. et al. A genetic algorithm with look-ahead mechanism to estimate

formant synthesizer input parameters. IEEE Congress on Evolutionary Computation (CEC), 2013, p. 3035-3042.

- Oliveira, F. P. et al. Multi-Objective Genetic Algorithm to Automatically Estimating the Input Parameters of Formant-Based Speech Synthesizers. In: Intech Open Access Publisher - (Org.). Genetic Algorithm - Book 2. . Rijeka: Intech Open Access Publisher, 2011, p.283-302.

- Borges, J. et al. GASpeech: A Framework for Automatically Estimating Input Parameters of Klatt's Speech Synthesizer. Anais do 10o Simposio Brasileiro de Redes Neurais, 2008, p. 81-86.

1.6 Estrutura da Tese

Este trabalho está estruturado em seis capítulos. O primeiro capítulo discute de forma introdutória a motivação e descrição do problema, a metodologia, os objetivos e a estrutura do trabalho. O capítulo dois aborda a síntese de voz, suas estratégias, com ênfase na síntese por formantes, além das técnicas utilizadas para imitar da voz humana. O capítulo três apresenta os conceitos de algoritmo genético e detalhamento do AG adaptado para o problema desta tese. No capítulo quatro, é apresentado o arcabouço desenvolvido para imitar a voz humana utilizando AG para estimar os parâmetros de entrada de um sintetizador de voz por formantes. O capítulo cinco apresenta os experimentos realizados assim como os resultados obtidos. Finalmente, o capítulo seis apresenta as conclusões e as propostas de trabalhos futuros são expostas.

Capítulo 2

Síntese e Imitação da Voz

2.1 Síntese de Voz

A síntese de voz consiste em produzir a fala humana artificialmente através da geração automática do sinal de voz. Aspectos como a naturalidade e a inteligibilidade são considerados quando se avalia a qualidade da voz sintética. Vários trabalhos em síntese de voz vem sendo desenvolvidos há décadas e alguns avanços foram alcançados, porém a qualidade em termos da naturalidade da voz sintética ainda apresenta lacunas principalmente no que tange as adaptações que a fala pode sofrer considerando a entonação e a emotividade associadas à expressividade do conteúdo a ser sintetizado.

2.1.1 Introdução e Histórico

A voz humana é produzida através do aparelho fonador, o qual é composto pelo diafragma, pulmões e o trato vocal (Figura 2.1). O ar ao ser expelido pelos pulmões, atravessa pelos demais órgãos componentes do trato vocal os quais sofrem movimentos e vibrações com a passagem do ar, alterando dessa maneira, o espectro do sinal emitido pelo pulmões e produzindo sons diferentes e inteligíveis pelo homem. Parte das pesquisas envolvendo a produção artificial da voz tenta de certa forma reproduzir o comportamento do aparelho fonador humano para a emissão de sons.

Os esforços em produzir voz artificialmente iniciou-se por volta do ano de 1779 quando o professor russo Christian Kratzenstein construiu um ressonador acústico, similar ao trato vocal humano, no qual era possível reproduzir os sons das vogais. Posteriormente, em 1791, Wolfgang

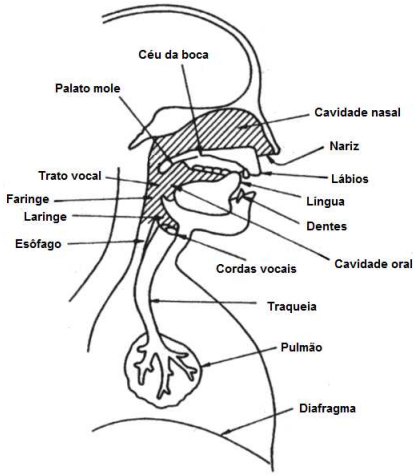

Figura 2.1: Aparelho fonador humano.

von Kempelen criou uma máquina em que era possível produzir sons simples ou combinados, com o diferencial de que a mesma possuía uma câmara de pressão, simulando os pulmões, a qual funcionava como se fosse as cordas vocais humana e um tubo de couro representando o trato vocal, permitindo através da manipulação dos seus componentes a emissão de sons das vogais e algumas consoantes. Em 1800, Charles Wheatstone reconstruiu uma nova versão da máquina de Kempelen a qual possuía um mecanismo mais sofisticado e permitia a produção dos sons das vogais e da maioria das consoantes, incluindo as nasais.

As pesquisas continuaram, porém com o objetivo de produzir sintetizadores elétricos. Em 1922, Stewart construiu um sintetizador composto por uma fonte imitando a funcionalidade dos pulmões (excitação) e dois circuitos ressonantes para modelar os ressonadores acústicos do trato vocal. Com esta máquina, foi possível unicamente a geração estática dos sons das vogais com duas formantes. O primeiro dispositivo considerado um sintetizador elétrico foi o VODER (*Voice Operating Demonstrator*), desenvolvido por Homer Dudley, em 1939. Esse sintetizador era composto por uma barra para selecionar o tipo de voz (sonora ou não sonora), um pedal para controlar a frequência fundamental e dez teclas que controlavam o trato vocal artificial. A estrutura básica do VODER é bastante similar aos sistemas baseados no modelo fonte-filtro existentes hoje em dia. Atualmente, a tecnologia envolvendo os sintetizadores de voz evoluiu e dentre as quais destacam-se as sínteses por concatenação, articulatória, por formantes (regras) e mais recentemente a síntese estatístico-paramétrica.

2.2 Sistemas de conversão texto-fala (TTS - *Text-to-Speech*)

Uma das aplicações da síntese de voz pode ser encontrada em sistemas TTS, os quais convertem um texto de entrada em uma voz artificial que seja inteligível e a mais natural possível. Portanto, a tarefa de sistemas desse tipo é bastante complexa, pois envolve a imitação (*mimicking*) de como os seres humanos realizam a leitura de um texto. Estes sistemas são compostos por dois componentes principais: *front-end* e *back-end* (Figura 2.2).

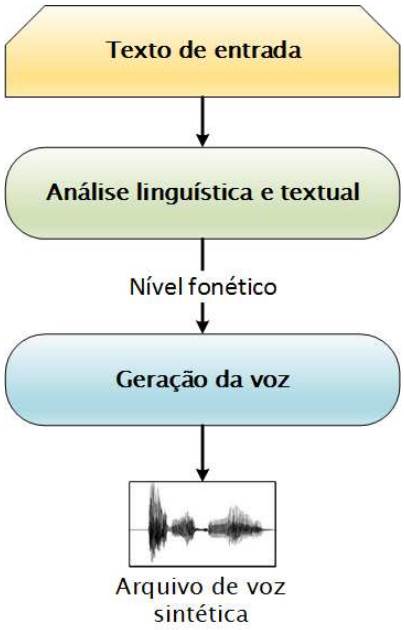

Figura 2.2: Componentes de um sistema de conversão texto-fala.

O *front-end* é responsável por receber o texto puro, converter símbolos (números e abreviações) em palavras escritas equivalentes, além de realizar a transcrição fonética de cada palavra (conversão grafema-fonema) e dividir o texto em unidades menores (fonemas, sentenças, cláusulas ou frases). Essa análise textual e linguística é feita por um módulo de Processamento de Linguagem Natural (PLN). O *back-end* realiza efetivamente a síntese de voz, convertendo a representação linguística em voz através de um componente de Processamento Digital de Sinais (PDS), transformando a informação simbólica que recebe em voz cujo som é o próximo possível do natural.

Existem duas medidas de qualidade em um sistema TTS, como, por exemplo, a segmental e supresegmental. A qualidade segmental representa a eficiência da máquina em produzir

fala soando próxima a uma voz natural, assumindo que o módulo PLN repasse informações de alta-qualidade (próxima do ser humano), ou seja, é uma medida de desempenho do módulo PDS. A qualidade suprasegmental representa a riqueza de conteúdo prosódico (entonação da voz) que a máquina é capaz de explorar, ou seja, é uma medida de desempenho do módulo PLN. Naturalidade e inteligibilidade são propriedades indispensáveis em um sistema TTS e dependem de ambos os módulos.

Os sistemas TTS podem fazer a síntese de voz através de várias estratégias como a síntese articulatória, por formantes entre outras. A seção seguinte apresenta as principais sínteses empregadas em sistemas TTS e em trabalhos que fazem imitação da voz.

2.3 Estratégias de Síntese de Voz

2.3.1 Síntese por concatenação

A síntese por concatenação consiste em agrupar unidades de voz pré-gravadas e produzir voz sintética inteligível que seja o mais próxima possível da naturalidade da voz humana. Essa técnica exige geralmente uma grande base de voz, maior capacidade de memória e pode limitar-se a um falante apenas dependendo da diversidade de falantes que compõe o corpus de voz. Ela ocorre em três fases: primeiramente a voz natural é gravada, posteriormente os segmentos (unidades) são definidos e etiquetados e, finalmente, as unidades mais adequadas são escolhidas para compor a voz gerada.

Devido ao fato das vozes pré-gravadas serem segmentadas para compor uma voz sintética, um dos aspectos importantes é encontrar o tamanho do segmento mais adequado, pois com unidades maiores a voz soa mais natural, há menos concatenações; porém, aumenta a necessidade de mais recursos de memória para realizar o processo. Caso a opção recaia em utilizar segmentos menores, aumenta a complexidade e dificuldade em relação ao processo de etiquetamento embora requeira menos recurso de memória. Os segmentos podem ser palavras, sílabas, fonemas, difones ou trifones [Lemmetty, 1999].

A concatenação por palavas é relativamente simples, entretanto a naturalidade da voz gerada fica comprometida visto que a sentença sintetizada contém palavras pronunciadas isoladamente e não apresentam aspectos de fala contínua. Neste caso, há necessidade de se ter um grande vocabulário. Já a concatenação baseada em sílabas, apesar de ser uma unidade menor que a palavra, o tamanho ainda é considerado grande para sistemas de síntese, sendo que a dificuldade maior consiste em controlar a prosódia da sentença. Os fonemas são as unidades

mais utilizadas devido ao tamanho ser menor do que os demais segmentos citados. Os difones além de conter os fonemas, possuem a transição entre os fonemas adjacentes reduzindo assim a distorção. As unidades um pouco mais longas, como trifones e tetrafones são raramente utilizadas.

Alguns problemas são encontrados ao se utilizar a síntese por concatenação, tais como:

1. Distorções em decorrência da descontinuidade nos pontos de concatenação;
2. Necessidade de grande quantidade de memória principalmente quando segmentos grandes são usados;
3. Tempo gasto na coleta e etiquetamento dos segmentos.

2.3.2 Síntese articulatória

A síntese articulatória consiste em modelar a dinâmica dos articuladores (língua, mandíbula, lábios, etc.) e das cordas vocais para produzir voz sintética de alta qualidade. Este modelo incorpora um controle individual sobre os articuladores, que se movimentam de maneira semi-independente uns dos outros. Essa relativa independência permite a modelagem da superposição de gestos articulatórios, que, normalmente, ocorre no processo natural de produção da fala.

Por ocasião da fala, os músculos do trato vocal causam a movimentação dos articuladores que por sua vez, mudam a forma do trato vocal causando a produção de diferentes sons. Os dados desse tipo de modelo são geralmente obtidos em 2D, a partir de análises de raios-X, porém o trato vocal real é naturalmente em 3D, dificultando assim a otimização desse tipo de modelo. Devido às limitações de dados, sistemas TTS eficientes, baseado neste tipo de síntese, por exemplo, ainda tem um longo caminho a percorrer, embora alguns que já existam, confirmem a potencialidade na produção de sinal com alta qualidade, limitando-se à geração de segmentos curtos de fala [Rubin et al., 1981].

2.3.3 Síntese por formantes

A síntese por formantes, ou baseada em regras, é fundamentada em um conjunto de regras usadas para determinar os valores dos parâmetros necessários para sintetizar uma fala através de um sintetizador. Este tipo de síntese é baseada no modelo fonte-filtro [Lemmetty, 1999] o qual permite a modelagem do trato vocal através de um filtro linear, como um conjunto

de ressonadores, que varia no tempo. Existem duas estruturas possíveis para os ressonadores, a saber, em cascata ou paralelo, sendo que para uma melhor performance, a combinação das duas arquiteturas pode ser utilizada. Para produzir sons inteligíveis são necessários pelo menos três formantes ou cinco, caso se queira sons com alta qualidade. Cada formante é geralmente modelada através de ressonadores de dois pólos os quais permitem a especificação da frequência e da largura de banda. Alguns dos parâmetros necessários para a síntese baseada em regras, são:

- Frequência fundamental ($F0$);
- Parâmetro de excitação (OQ);
- Grau de excitação da voz (VO);
- Frequência e amplitude das formantes ($F1...F3$ e $A1...A3$);
- Frequência de um ressonador adicional de baixa frequência (FN);
- Intensidade das regiões de alta e baixa frequência (ALF, AHF) entre outros.

Na arquitetura em cascata, os ressonadores são conectados em série e a saída de um é utilizada como entrada para o próximo, necessitando apenas das frequências das formantes como controle da informação (Figura 2.3). As vantagens desses sintetizadores consistem no fato das amplitudes das formantes relativas as vogais não precisarem de controles individuais além de serem uma boa opção para produção de sons não-nasais. Em contrapartida, esse tipo de arranjo não parece adequado para gerar sons plosivos e fricativos [Lemmetty, 1999].

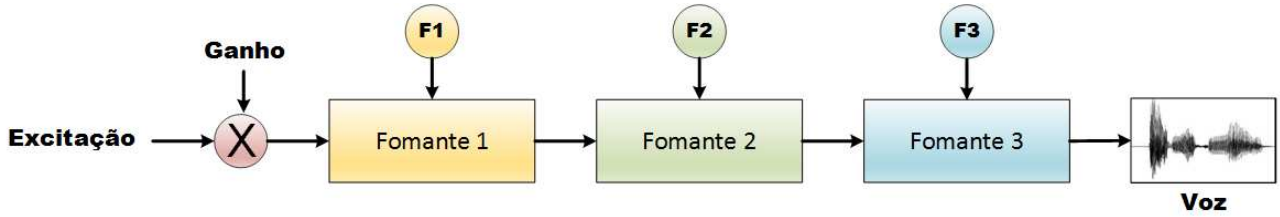

Figura 2.3: Sintetizador por formantes em cascata adaptado de [Lemmetty, 1999].

O sintetizador com a arquitetura em paralelo consiste nos ressonadores conectados paralelamente sendo o sinal de excitação aplicado a todas as formantes simultaneamente e suas saídas são resumidas para gerar o sinal final (Figura 2.4). Essa estrutura permite o controle individual da largura de banda e do ganho para cada formante. Os sons nasais e

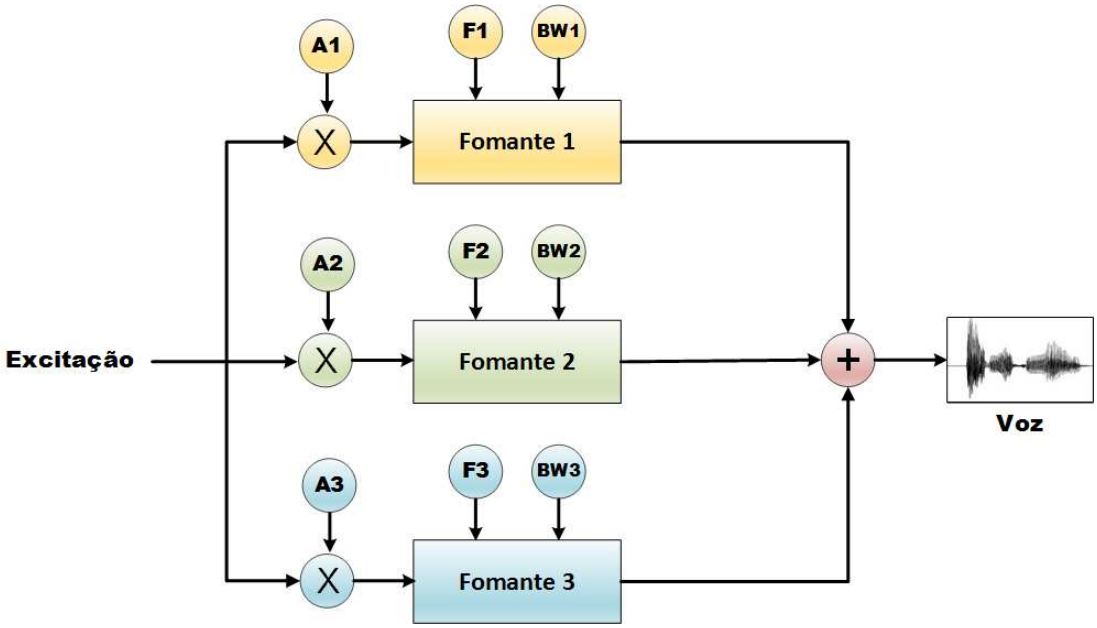

Figura 2.4: Sintetizador por formantes em paralelo adaptado de [Lemmetty, 1999].

fricativos podem ser modelados através desta arquitetura, porém os sons de algumas vogais não.

Os sintetizadores por formantes exigem pouco custo computacional e a qualidade da voz sintética obtida por eles possui alto grau de inteligibilidade, porém é difícil reproduzir exatamente o sinal de voz emitido por um falante humano. Exemplos de sintetizadores desse tipo são o Klatt e o *HLSyn*, abordados na Seção 2.4, e utilizados neste trabalho.

2.3.4 Síntese estatístico-paramétrica

A síntese estatístico-paramétrica surgiu nos anos 90 e tem crescido em popularidade tanto no meio acadêmico quanto comercial, pois sua metodologia consiste em utilizar vários parâmetros acústicos da voz em um modelo estocástico de séries temporais. Geralmente, esse tipo de síntese é baseada em Modelos Ocultos de Markov (*MOM*) e chamada, portanto, de síntese de voz baseada em MOM [King, 2011]. Esta síntese produz não apenas uma sequência de fonemas como também faz uso de vários contextos da especificação linguística, essenciais para guiar o modelo simplificado de produção da voz, o qual consiste em parâmetros de excitação e de modelagem do trato vocal [Tokuda et al., 2013]. Sendo assim, ela prediz os parâmetros da voz a partir de um sistema TTS e para isso usa geralmente um MOM do tipo esquerda-para-direita, contendo três estados e mais um conjunto de distribuições de

probabilidades, a saber, estado inicial, transição de estados e estado de saída.

Neste tipo de síntese, o treinamento prévio do modelo é necessário e realizado através da estimação da máxima verossimilhança dos parâmetros extraídos da voz. Porém, os parâmetros da voz não são gerados diretamente a partir do texto de entrada. Antes disso, o texto é convertido em uma sequência de rótulos dependentes de contexto. De acordo com a sequência dos rótulos, um MOM maior é construído a partir da concatenação dos MOM menores, dependentes de contexto. Para cada quadro de voz, um vetor de observações é gerado contendo os parâmetros do espectro e excitação da voz, os quais são sintetizados para compor uma voz artificial.

Este tipo de síntese é bastante flexível e permite entre outras vantagens, como, por exemplo, a adaptação de locutor, a interpolação de vozes de diferentes locutores, a produção de vozes em determinado estado emocional, o suporte para múltiplas línguas e a síntese de vozes cantadas. O exemplo de uma ferramenta que faz uso deste tipo de síntese é o *HTS* [Zen et al., 2007] [Tokuda et al., 2000] [HTS, 2010], o qual possui interface e funcionalidade muito semelhante à ferramenta de reconhecimento de voz chamada *HTK* [Young, 2005] [HTK, 2010].

2.4 Sintetizadores por formantes

Na síntese por formantes, dois sintetizadores destacam-se: Klatt e *HLSyn*. O *HLSyn* é baseado no Klatt, porém possui incorporado algum conhecimento especializado sobre fonética articulatória e acústica, permitindo assim a redução da quantidade de parâmetros utilizada pelo sintetizador. Nas seções seguintes, são apresentados estes sintetizadores, bem como as características e peculiaridades de cada um deles.

2.4.1 Sintetizador de Klatt

Os sintetizadores são bastante úteis em experimentos que envolvem a percepção e a produção da voz. As técnicas de síntese de voz podem ser divididas em três classes: síntese direta, simulação do trato vocal e modelo de produção da fala [Keller, 1994]. Na síntese direta, o sinal é gerado através da manipulação direta da forma de onda. Na simulação do trato vocal, a voz é produzida através da simulação do comportamento dos órgãos responsáveis pela produção da fala. Já a síntese baseada em modelo de produção da fala consiste na modelagem do trato vocal através de um filtro linear formado por um conjunto de ressonadores que variam no tempo [Lemmetty, 1999]. O filtro é excitado através de uma fonte, simulando a vibração das

cordas vocais para os sons vozeados ou a compressão do trato vocal quando se quer produzir ruído. Dessa maneira, o som é criado no trato vocal e irradiado para os lábios.

O sintetizador de Klatt pode funcionar em três estruturas possíveis: ressonadores dispostos em cascata, paralelo ou combinação de ambas. As duas primeiras estruturas foram abordadas anteriormente (seção 2.3.3) e a combinação de ambas (cascata/paralelo) é utilizada quando existe a necessidade de melhorar a performance do sintetizador, pois o arranjo em cascata é adequado para a produção de sons vocálicos orais e o em paralelo para sons nasais vocálicos e consonatais.

Basicamente, o Klatt funciona da seguinte forma: para cada quadro de voz, com duração variando entre cindo e dez milissegundos, uma combinação de valores dos parâmetros deve ser passado como entrada para o sintetizador com o objetivo de gerar um trecho de voz sintética. Dentre os parâmetros necessários para um sintetizador por formantes, tem-se a frequência fundamental ($F0$), o coeficiente de abertura da glote (OQ), a amplitude do vozeamento (AV), as frequências e amplitudes das formantes ($F1$ à $F3$ e $A1$ à $A3$), entre outros. Existem várias versões do Klatt, porém três softwares destacam-se: Klatt80, Klatt88 e a versão do Jon Iles (v.3.03). Apesar de ser possível gerar voz com qualidade através deste sintetizador, existem algumas dependências que ocorrem entre a fonte de voz e a função do trato vocal para a produção de determinados sons. Esse tipo de modelo de síntese implementa tais restrições através da escolha correta dos valores dos parâmetros de entrada do sintetizador, porém a quantidade de parâmetros é grande, sendo uma tarefa difícil ajustá-los manualmente para alcançar exatamente o mesmo modelo do trato vocal humano.

De acordo com [Klatt, 1980], a versão Klatt80, chamada *KLSYN*, possui trinta e nove parâmetros os quais combinados determinam as características do sinal gerado na saída. Cada parâmetro possui uma faixa de valores possíveis (mínimo e máximo). Para se ter uma voz artificial com boa qualidade deve-se variar entre 20 e 39 parâmetros, pois alguns deles são constantes para o mesmo falante tais como: AN (amplitude das formantes nasais), $A1$ (amplitude da primeira formante), FGP (frequência do primeiro ressonador glotal), entre outros. Além disso, existe a necessidade de atribuir valores a algumas variáveis que são essenciais para a inicialização do sintetizador como SW (*switch* cascata/paralelo), SR (taxa de amostragem), NWS (quantidade de amostras por trecho de voz), $G0$ (ganho) e NFC (quantidade de formantes em cascata). Nesta versão, dois tipos de fonte de voz podem ser usadas durante a produção da fala: vozeada (sons sonoros) e aspirativa (sons com ruído). Nos sons sonoros há a vibração das cordas vocais enquanto nos aspirativos existe a necessidade da geração de ruído para simular a rápida passagem do ar vindo dos pulmões através de áreas estreitas do trato vocal.

Neste trabalho, foi utilizada a última versão - Klatt88 [Klatt and Klatt, 1990], chamada também *KLSYN88*, a qual possui quarenta e oito parâmetros que devem ser configurados para a produção de cada quadro de voz sintética. Nesta versão, seis parâmetros (ANV, $A1V$, $A2V$, $A3V$, $A4V$ e ATV) não são utilizados, pois assumem um valor constante igual a zero para qualquer falante. Algumas variáveis de inicialização do sintetizador mudaram de nomenclatura em relação à versão Klatt80 como a CP (*switch* cascata/paralelo), NF (quantidade de formantes em cascata) e o ganho que foi fragmentado em três parâmetros, a saber, GV (ganho global em relação ao AV), GH (ganho global em relação ao AH) e GF (ganho global em relação ao AF). Nesta versão, existem fontes para sons sonoros, aspirativos e/ou fricativos. As amplitudes dessas fontes são controladas pelos parâmetros AV, AH e AF respectivamente. A fonte de som sonoro difere da versão anterior *KLSYN*, pois apresenta três opções: *Liljencrants-Fant*, (LF) e modelos KLGLOTT88 e trem de impulsos filtrados.

O modelo KLGLOTT88 foi a fonte escolhida para este trabalho. O arranjo dos filtros em cascata é responsável pela função de transferência do trato vocal para modelagem de sons laringeais enquanto o arranjo em paralelo contém as amplitudes de $A2F$ à $A6F$ para controlar a excitação da fonte fricativa AF. Para algumas aplicações específicas de síntese de voz, os ressonadores em cascata podem ser utilizados em conjunto com um outro arranjo de ressonadores em paralelo com o objetivo de sintetizar sons laringeais mais próximos da voz natural. Porém, esse filtro adicional geralmente não é utilizado.

A versão do Klatt do Jon Iles [Laprie and Bonneau, 2002] é baseada na versão Klatt80 [Klatt, 1980], porém incorporando algumas mudanças, como, por exemplo, a utilização apenas do arranjo de ressonadores em paralelo, aumento da quantidade de parâmetros de entrada do sintetizador em relação à versão *KLSYN*, controle da excitação do $F0$, possibilidade de utilizar uma amostra natural da forma de onda de excitação, remoção do software de síntese de voz do contexto de sistemas TTS, entre outras. Nesta versão, o arquivo de entrada consiste na configuração de quarenta e um parâmetros para produzir cada quadro de voz. Um desses parâmetros é chamado *time* e indica em que instante do arquivo de voz original os parâmetros foram estimados. Cada quadro representa dez milissegundos de áudio. Além disso, existem quatro tipos de fontes de vozeamento: trem de impulsos, simulação natural, amostra natural da excitação e excitação *Liljencrants-Fant*. Alguns parâmetros são os mesmos utilizados na versão Klatt80 como o $F0$, AV e as formantes $F1$ à $F6$, enquanto outros são duplicados com a finalidade de distinguir quais são utilizados nos filtros em cascata e em paralelo. Por exemplo, os parâmetros $B1$ e $B1P$ representam a largura de banda da formante $F1$, porém o primeiro é utilizado no arranjo em cascata e o segundo no paralelo. Esta versão do Klatt é parte de um software que faz imitação da voz chamado Winsnoori [Laprie and Bonneau, 2007].

2.4.1.1 Descrição do Klatt88

Conforme mencionado na seção anterior, a versão utilizada neste trabalho foi a Klatt88 [Klatt and Klatt, 1990](Figura 2.5). Assim como nesta e nas demais, algumas variáveis são utilizadas para a configuração inicial do sintetizador e não efetivamente como parâmetro para a síntese de voz. Dentre essas variáveis tem-se a duração da sentença (DU), taxa de amostragem (SR) e quantidade de formantes na configuração em cascata (NF). Os parâmetros usados no modelo da fonte de voz KLGLOTT88 são os seguintes: $F0$, AV, OQ, FL e DI. O $F0$ é a frequência fundamental, considerado um dos parâmetros mais importantes, pois indica quando um quadro de voz é sonoro ou não sonoro. O AV é a amplitude de vozeamento, o OQ o coeficiente de abertura, o FL a variação lenta das flutuações em $F0$ e o DI a diplofonia a qual realiza a redução da amplitude de $F0$ em períodos alternados. Conforme pode ser observado na Figura 2.6, o $F0$ e o AV possuem forte correlação pois quando $F0$ é diferente de zero o som é sonoro e o AV é diferente de zero também. A referida figura ilustra os valores de $F0$ e AV para um falante masculino referente à pronuncia da frase *The birch canoe slid on the smooth planks.*

Os parâmetros TL e AH são considerados parte da fonte de voz e responsáveis respectivamente pelo decaimento extra no espectro da voz e pela amplitude da aspiração. Sons aspirativos e fricativos são produzidos através de um gerador de ruído, o qual tem sua amplitude modulada através dos parâmetros AH e AF. Os valores desses parâmetros podem variar entre 80dB (forte ruído aspirativo ou fricativo) e 0dB (sons não-aspirativos ou não-fricativos).

Esta versão do Klatt possui um arranjo dos ressonadores em cascata para modelar o trato vocal, como as demais versões citadas anteriormente, sendo a laringe, neste caso, a fonte de voz. Existem, ainda, mais dois arranjos de filtros em paralelo: uma para controlar a amplitude dos sons fricativos e outro para produzir sons laringeais, porém este arranjo geralmente não é utilizado (Figura 2.5 - ressonadores em rosa). Para a configuração em cascata, a função de transferência é representada no domínio da frequência através de pólos e zeros. Cinco ressonadores são necessários para simular o trato vocal. Cada ressonador é representado através de uma formante n, a qual possui uma frequência F_n, e parâmetro para controlar sua largura de banda B_n. Este arranjo (em cascata) possui as formantes F_1 à F_5 e suas respectivas larguras de banda (B_1 à B_5).

Para sons nasais, pólos e zeros adicionais são inseridos no sinal através de um ressonador (RNP - representado pelos parâmetros FNP e BNP) e um anti-ressonador (RNZ - parâmetros FNZ e BNZ). Nas vogais nasalisadas, por exemplo, RNP e RNZ causam a redução da amplitude da formante F_1. Se o som não é nasalisado, ambos (ressonador e anti-ressonador)

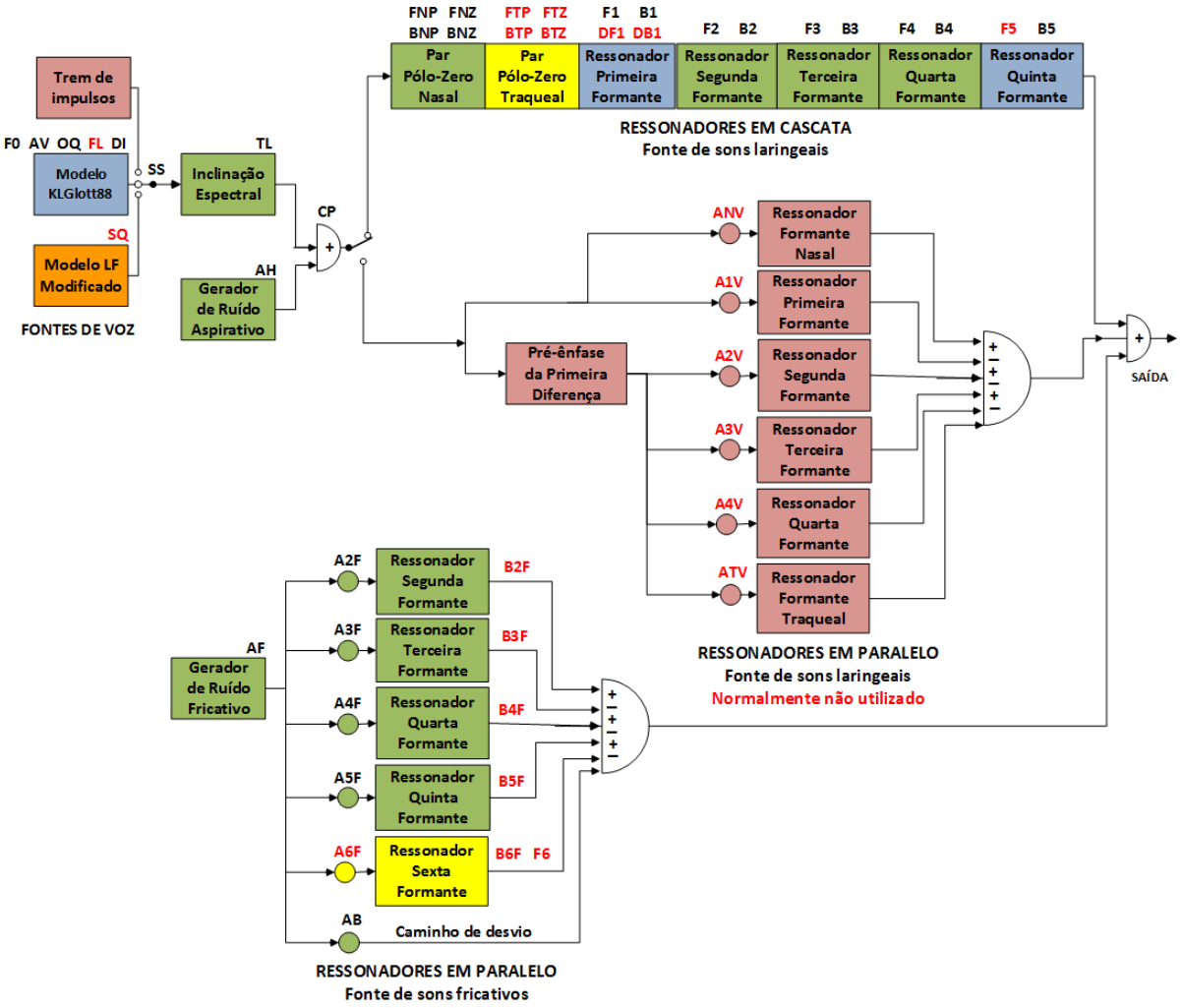

Figura 2.5: Diagrama completo do Klatt88 traduzido de [Klatt and Klatt, 1990].

são removidos ($FNP = FNZ$).

O modelo do trato vocal em paralelo tem cinco formantes (F_2 à F_6) com finalidade de sintetizar sons com alta frequência de ruído. Alguns sons não contém picos significativos no sinal, precisando, portanto, de um mecanismo que permita a passagem do sinal sem cruzar um ressonador. Esse mecanismo é chamado de *bypass* e o parâmetro AB é responsável por controlar a amplitude desses sinais. As amplitudes dos picos das formantes ($A2F$ à $A5F$) possuem valores ajustados para 60dB, aproximando-se, portanto, à configuração em cascata. Nos Apêndices A e B constam, respectivamente, uma breve definição dos parâmetros da versão utilizada e os intervalos de valores possíveis de acordo com [Klatt and Klatt, 1990]. No Apêndice C é apresentado um exemplo de arquivo de entrada do Klatt88.

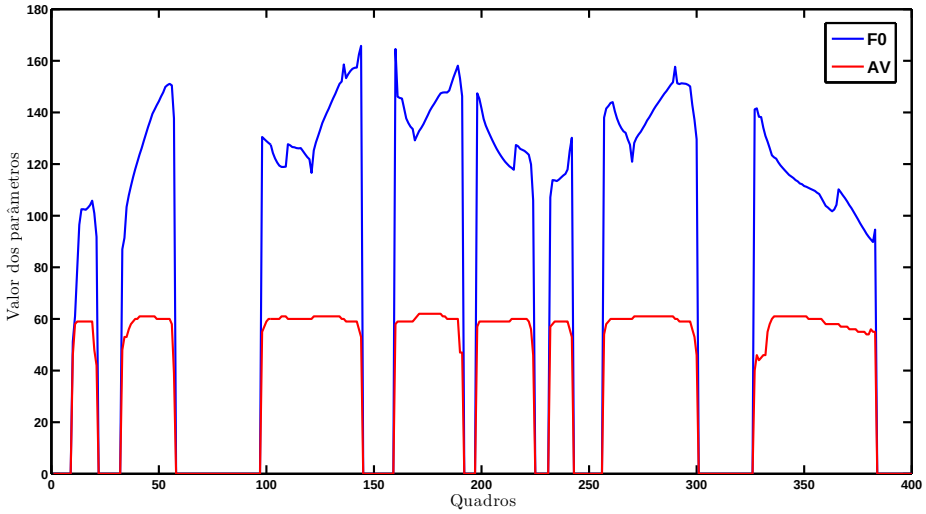

Figura 2.6: Valores de $F0$ e AV.

2.4.2 Sintetizador HLSyn

O *HLSyn* é um sintetizador que faz parte de um sistema TTS chamado *VHLSyn* (*Very High-Level Synthesis*) e foi desenvolvido a partir de regras que manipulam treze parâmetros articulatórios e acústicos, os quais são mapeados nos parâmetros do Klatt [Hanson et al., 1999]. A tabela 2.1 apresenta os parâmetros deste sintetizador assim como uma breve descrição deles.

O mapeamento dos treze parâmetros do *HLSyn* para os quarenta e oito parâmetros da versão *KLSYN88* do Klatt é realizado através de um circuito que modela o sistema de produção da fala. Através dessa modelagem, é possível obter parâmetros intermediários que representam a pressão e a passagem do ar por áreas que apresentam obstáculos no trato vocal. Estes parâmetros são, então, utilizados para calcular os parâmetros do Klatt. O principal benefício em se utilizar o *HLSyn* em relação aos tradicionais sintetizadores por formantes é a redução do número de parâmetros que são diretamente controlados sendo estes parâmetros uma representação natural do controle da produção da fala humana. A Figura 2.7 ilustra o funcionamento do TTS *VHLSyn* no qual primeiramente é informada uma sequência fonética (passo 1) a qual é submetida a regras de conversão originando marcadores fonéticos e o tempo em que eventos articulatórios importantes acontecem tais como o encerramento ou o núcleo de uma consoante ou de uma vogal (passo 2). Após essa conversão são gerados os 13 parâmetros do *HLSyn* (passo 3). Esses parâmetros por sua vez são submetidos às relações de mapeamento (passo 4) produzindo como saída os parâmetros do Klatt, os quais são sintetizados gerando a voz sintética referente à sequência fonética (passo 5).

Tabela 2.1: Os 13 parâmetros do *HLSyn*.

Parâmetros	Descrição
F0	Frequência fundamental
AG	Área média da abertura da glote
AP	Área posterior à abertura glotal
PS	Pressão subglotal
AL	Área da seção transversal da constrição dos lábios
AB	Área da seção transversal da constrição da língua
AN	Área da seção transversal da porta velofaringeal
UE	Taxa de aumento do volume trato-vocal
DC	Mudança na prega vocal ou
F1 à F4	Frequência das primeiras 4 formantes

2.5 Imitação da Voz

A geração de voz sintética com o objetivo de imitar uma voz humana possui interesses tanto comerciais, como a criação de um personagem virtual, quanto clínicos em que a pessoa está impossibilitada de comunicar-se normalmente através da fala devido algum trauma, doença ou cirurgia [Bangayan et al., 1997]. Além disso, a imitação da voz também pode ser empregada em sistemas TTS em que as características da voz de uma pessoa em um determinado idioma podem ser utilizadas para gerar voz em outro idioma, porém mantendo as características do falante da voz original [Yarrington et al., 2005]. Não obstante, existem algumas restrições que fazem com que a "clonagem" seja uma atividade pouco explorada na síntese de voz. De acordo com [Aylett and Yamagishi, 2008], são elas:

- A síntese resultante deve soar o mais natural possível para efetivamente imitar uma voz;
- O ideal é realizar a imitação da voz a partir da menor quantidade possível de material disponível;
- O estilo de voz a ser imitada requer técnicas diferentes e eventualmente uma quantidade diferente de vozes pré-gravadas.

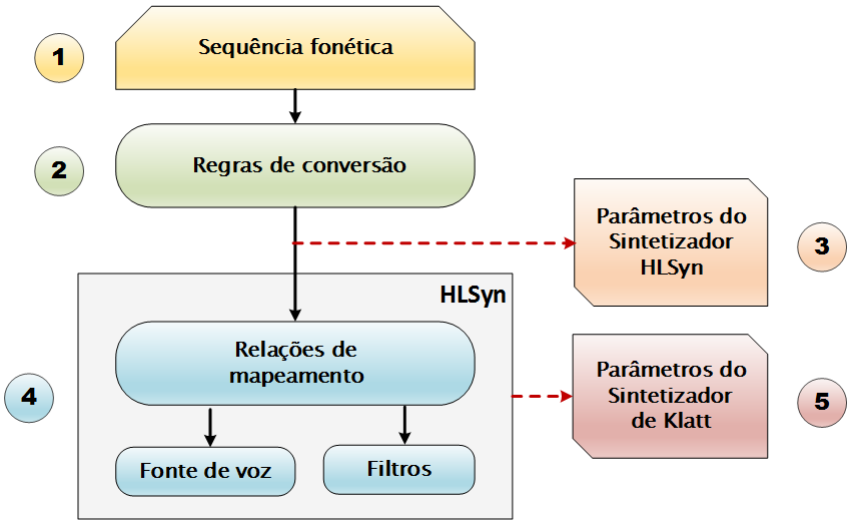

Figura 2.7: Síntese baseada em regras através do *VHLSyn* adaptado de [Hanson et al., 1999].

Nas seções seguintes são apresentados alguns trabalhos que envolvem a imitação da voz, utilizando o processo de síntese articulatória, por concatenação, além de um arcabouço híbrido que compreende as sínteses por concatenação e estatístico-paramétrica.

2.5.1 Imitação da Voz utilizando Síntese Articulatória

Segundo Howard [Howard and Huckvale, 2005], o objetivo principal é construir um sistema que seja capaz de aprender a imitar uma determinada voz utilizando seu próprio trato vocal. Para isso, um modelo inverso que faz o mapeamento entre a representação acústica da voz e os parâmetros articulatórios do sintetizador foi treinado utilizando uma técnica de regressão supervisionada e dados obtidos através de um gerador baseado em MOM. O problema maior do modelo proposto foi o fato de várias configurações do trato vocal gerarem sequências acústicas similares ou idênticas. Os testes foram feitos com vozes utilizadas no treino do modelo e vozes naturais de um falante masculino. A avaliação dos resultados foi realizada através de testes de audição e observação dos espectrogramas. O sistema apresentou bom desempenho para imitar vozes já conhecidas do modelo, porém deixou a desejar em relação às vozes naturais masculinas.

A síntese articulatória, também, é empregada para realizar o mapeamento acústico-articulatório, utilizando algoritmo genético multi-população [Brito, 2007] . Neste caso, doze subpopulações são criadas em cada geração sendo cada uma composta por vinte indivíduos. Um operador de migração é responsável por mover indivíduos de uma subpopulação para outra com o objetivo de agrupar os indivíduos mais aptos, porém mantendo a diversidade.

A função objetivo empregada permite discernir as configurações boas através da avaliação da distância acústica somada à um fator que garanta uma penalização caso os parâmetros estejam discontínuos. Os testes foram realizados com vinte e cinco vogais pronunciadas por diferentes falantes masculinos e selecionadas de uma base de sentenças em espanhol. A avaliação dos resultados foi feita por oito ouvintes de nacionalidade venezuelana que após escutarem a vogal sintética gerada deveriam digitar o que tinham escutado. A média do erro do reconhecimento da vogal ficou abaixo de 0.5 indicando, portanto, que a metodologia empregada foi bastante satisfatória.

Em trabalho mais recente [Philippsen et al., 2014], a imitação da voz é feita especificamente para sílabas, usando, para isso, um modelo recorrente, baseado em redes neurais. Esse modelo é capaz de fazer o mapeamento articulatório-acústico, e vice-versa, para sequências de consoantes-vogais, incluindo os efeitos co-articulatórios. Os modelos foram treinados inicialmente com um pequeno conjunto contendo trajetórias articulatórias e acústicas, sendo melhorados posteriormente através de metas auditivas que um módulo chamado "aprendiz" tenta imitar. Esta etapa posterior chama-se refinamento, baseado na imitação, pois os resultados acústicos das imitações, juntamente com as ações executadas pelo "aprendiz", servem como novos pares de treinamento para o modelo. Os resultados mostraram que quanto maior o conjunto de treinamento, menor o valor do erro. Além disso, o processo de refinamento reduziu significativamente o erro nos modelos articulatório-acústico e inverso, em alguns casos chegando à 94.7% essa redução. Em conjunto com a análise dos resultados baseada no erro, foi realizada uma avaliação perceptual em que ouvintes escutaram sessenta e quatro sílabas antes e depois do refinamento. Esta avaliação revelou que 64% das sílabas foram melhor reconhecidas após o refinamento, enquanto 12% ficaram menos compreensíveis.

2.5.2 Imitação da Voz utilizando Síntese por Concatenação

De acordo com Bulut [Bulut et al., 2002], a imitação abrange vozes que representem quatro estados emocionais: raiva, alegria, tristeza e neutra, utilizando a síntese por concatenação de unidades de voz chamadas difones. Cinco sentenças alvo emocionalmente imparciais foram escolhidas para os testes. O corpus de voz foi construído a partir das gravações de 357 sentenças nas quais a locutora foi uma atriz, abrangendo os quatro estados emocionais citados anteriormente. As sentenças foram segmentadas foneticamente e alinhadas através de um software específico. Os resultados foram avaliados subjetivamente por trinta e três ouvintes, sendo catorze falantes nativos do inglês e dezenove não nativos, e consistia em atribuir uma nota de 1 (ruim) a 5 (excelente), em relação ao estado emocional que a voz sintética representava. As

vozes representando o estado emocional de alegria foram as mais difíceis de serem imitadas ao passo que as de tristeza foram as mais similiares.

2.5.3 Imitação da Voz utilizando Síntese Híbrida

A imitação da voz pode, também, ser realizada de maneira combinada [Aylett and Yamagishi, 2008] em que se utiliza um sistema híbrido chamado *Cereproc* para reproduzir sinteticamente a voz de George W. Bush. Os dados utilizados foram obtidos através de áudios disponíveis sem custo na Internet. Após a escolha cuidadosa dos áudios para evitar ruídos ao fundo como aplausos e músicas, estes foram segmentados em sentenças variando o tamanho de 1 a 261 palavras. Nove vozes foram geradas através das sínteses estatístico-paramétricas, por concatenação ou utilizando o sistema híbrido citado anteriormente o qual é composto por essas duas sínteses combinadas.

Para a geração da voz através da síntese estatístico-paramétrica, Aylett [Aylett and Yamagishi, 2008] utilizou o sistema *HTS*, o qual é baseado em MOM e gera vozes sintéticas independente de falante. Características acústicas da voz como coeficientes cepstrais na escala *Mel*, log de $F0$ e medidas de aperiodicidades são utilizadas para treinar este sistema, porém ele requer mais de 10 horas de dados de voz de diferentes falantes para adaptar a geração de voz sintética a um determinado falante específico. Entretanto, as vozes sintéticas produzidas apresentaram características de robotização. Já na síntese por concatenação o sistema adotado foi o *CereVoice* o qual emprega unidades de voz chamadas difones, as quais são escolhidas por um motor de busca e combinadas para gerar uma determinada palavra ou sentença. O problema maior deste tipo de síntese é a quantidade de unidades de voz diferentes que são necessárias para a cobertura completa de todas as combinações fonéticas possíveis para uma determinada língua.

O sistema híbrido *Cereproc* tenta produzir vozes aproveitando as vantagens dos dois sistemas citados previamente, ou seja, tenta produzir vozes que soem naturais, como na síntese por concatenação, porém utilizando poucos dados para treino conforme acontece na síntese estatístico-paramétrica. Sendo assim, ele produz voz sintética através da concatenação paramétrica de unidades de voz. Para realizar os testes, noves sentenças com tamanhos variáveis (8 a 31 palavras) foram escolhidas e geradas sinteticamente. Vinte e três ouvintes tiveram que avaliar a naturalidade com que as sentenças eram proferidas e para isso era necessário atribuir uma nota variando de 1 (ruim) a 5 (excelente). O sistema proposto apresentou alguns problemas em concatenar segmentos de voz gravados em ambientes diferentes e apesar do *HTS* ter apresentado um desempenho melhor, chegou-se à conclusão que a síntese híbrida apresenta

grande potencial para a síntese de voz com boa qualidade.

2.6 Conclusões sobre o capítulo

Conforme abordado neste capítulo, existem diversas abordagens para a obtenção de voz sintética através do computador, variando desde a síntese por concatenação que requer um grande corpus previamente gravado até a síntese mais recente chamada estatístico-paramétrica, a qual é baseada em Modelos Ocultos de Markov. A imitação de uma determinada voz através da síntese não é uma tarefa fácil e os trabalhos mais representativos na área utilizam a síntese articulatória [Howard and Huckvale, 2005,Brito, 2007], por concatenação [Bulut et al., 2002] ou modelos híbridos [Aylett and Yamagishi, 2008], em que foram empregadas dois tipos de síntese, a saber, por concatenação e estatístico-paramétrica. Porém, a síntese por formantes é bastante utilizada por foneticistas devido ao alto grau de interpretabilidade dos parâmetros utilizados para modelar a fonte de voz e o trato vocal. Sendo assim, no capítulo seguinte é abordado o conceito de algoritmo genético no qual baseia-se o arcabouço desenvolvido para imitar uma determinada voz através da síntese por formantes, mais especificamente os sintetizadores de Klatt (versão KLSYN88) e *HLSyn*.

Capítulo 3

Algoritmo Genético

No presente capítulo, são apresentados os principais conceitos e aspectos sobre os algoritmos genéticos (AGs), seus operadores e as particularidades do AG empregado neste trabalho, como os conceitos de dominância, elitismo, cruzamento SBX (Simulated Binary Crossover) e a mutação polinomial. Ao final, são abordadas as conclusões deste capítulo.

3.1 Introdução

O conceito de otimização está relacionado à análise de soluções complexas, envolvendo diversas variáveis com o objetivo de quantificar a performance e medir a qualidade das decisões. De posse desta análise, a escolha recai na melhor solução, considerando, se possível, as restrições impostas pelo problema [Sivanandam and Deepa, 2008]. A otimização pode ser realizada computacionalmente, mas, para isso, faz-se necessária a modelagem do problema o mais próxima possível da realidade e a escolha do método computacional mais adequado. Porém, em se tratando de problemas complexos, nem sempre os métodos de busca encontram realmente a melhor solução global, ou seja, muitas vezes a escolha recai em soluções consideradas mínimos ou máximos locais. Os AGs são bastante eficientes como algoritmos de otimização, pois conseguem alcançar mais vezes a solução global [Sivanandam and Deepa, 2008, Verma and Kumar, 2014]. Recentemente, os AGs vendo sendo utilizados nas mais diferentes áreas [Padhye, 2012], tais como, no gerenciamento de tarefas, recursos e segurança na computação em nuvem [Singh and Kalra, 2014, Zhan et al., 2015, Sahil et al., 2015], na análise de grandes quantidades de dados [Nunez and Attoh-Okine, 2014, Khatun et al., 2015], em redes de computadores [Wang et al., 2015, Ly et al., 2015] e Internet [Agbele et al., 2012].

Embora a ampla utilização deles seja direcionada para problemas envolvendo apenas uma única função objetivo a ser minimizada ou maximizada, trabalhos recentes [Ma et al., 2015, Yuan et al., 2015, Li et al., 2015, Bansal et al., 2015, Gamal et al., 2015, Triantafyllidis et al., 2015] vem sendo desenvolvidos para encontrar a otimização de várias funções objetivo ao mesmo tempo, caracterizando esse tipo de problema como Otimização Multi-Objetivo.

3.2 Algoritmos Genéticos

Segundo Russel [Russel and Norvig, 2013], Sivanandam [Sivanandam and Deepa, 2008] e Verma [Verma and Kumar, 2014], os AGs são algoritmos estocásticos, pertencentes à área da Inteligência Computacional, especificamente, à Computação Evolucionária (CE), a qual busca, através de técnicas inspiradas na natureza, o desenvolvimento de sistemas inteligentes que imitem aspectos do comportamento humano, tais como, evolução e adaptação. Eles possuem as técnicas de busca e otimização inspiradas no princípio Darwiniano da evolução natural das espécies e na genética, utilizando, assim, a seleção natural e a reprodução genética através dos operadores de cruzamento e mutação, ou seja, os indivíduos, que compõem um AG, são avaliados e aqueles mais aptos são selecionados e terão a chance de maior longevidade, perpetuando, assim, seu material genético por várias gerações através da recombinação entre eles. Considerando um problema complexo para ser resolvido através de um AG, por exemplo, este deve ser modelado através de uma função matemática em que os indivíduos mais aptos obterão maior ou menor valor, dependendo se o objetivo é minimizar ou maximizar a função [Verma and Kumar, 2014].

Em uma população, podem existir vários indivíduos, sendo cada um deles, correspondente à uma possível solução do problema (ou função) e sua indentidade é composta por um ou vários cromossomos. Por exemplo, se a função que modela o problema contém três variáveis, cada uma é representada por um cromossomo e a concatenação deles compõe um indivíduo. Um cromossomo é composto por vários caracteres (genes), cada um destes encontra-se em uma determinada posição (locus), com seu valor determinado (alelo). Na Figura 3.1, tem-se no item (a) um gene, em (b) um alelo, em (c) um cromossomo com 4 genes, em (d) um indivíduo composto por 3 cromossomos e em (e) uma população com cinco indivíduos.

De acordo com Sivanandam [Sivanandam and Deepa, 2008] e Padhye [Padhye, 2012], as etapas envolvidas em um AG estão ilustradas na Figura 3.2. O algoritmo começa com a inicialização da população e posterior verificação para se identificar quais são os indivíduos mais aptos. Estes indivíduos são selecionados para o cruzamento e cada gene que compõe o

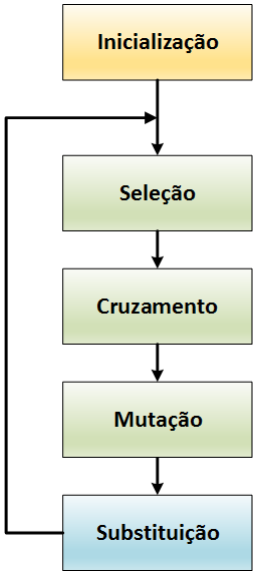

Figura 3.1: Composição de uma população.

cromossomo pode sofrer mutação. Após a etapa de mutação, é realizada uma nova avaliação dos indivíduos e aqueles com maior grau de aptidão, ou seja, com maior valor para a função objetivo, garantirão a sobrevivência para a nova população (etapa de substituição). Os ciclos evolutivos são repetidos até que seja alcançado algum critério de parada que pode ser, por exemplo, a quantidade máxima de gerações, perda de diversidade da população com indivíduos muito semelhantes ou a convergência do processo de otimização que consiste em o algoritmo alcançar o resultado esperado na função objetivo. Um AG é composto por vários elementos [Reeves and Rowe, 2003], porém para este trabalho destacam-se:

1. Problema de otimização;
2. Codificação do indivíduo;
3. Função objetivo ou *fitness*;
4. Seleção, cruzamento, mutação e elitismo;
5. Características da população e número de gerações.

3.2.1 Problema de otimização

Conforme enfatizado anteriormente, os AGs apresentam bons resultados quando aplicados a problemas complexos que caracterizam-se por:

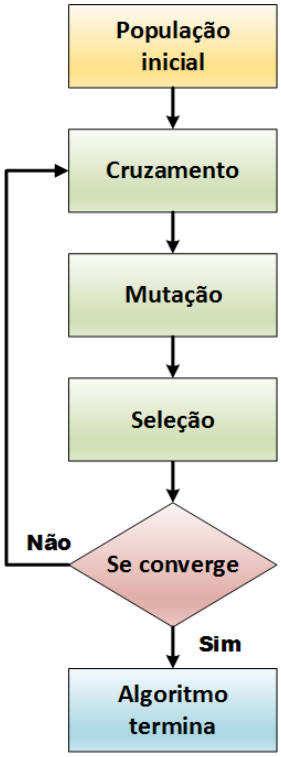

Figura 3.2: Etapas envolvidas em um algoritmo genético.

- Possuir vários parâmetros que precisam ser combinados em busca da melhor solução;
- Problemas com muitas restrições ou condições que não podem ser modelados matematicamente;
- Problemas com grande espaço de busca.

A otimização consiste em achar a solução melhor ou ótima entre as várias opções factíveis. Considerando que o problema a ser resolvido deve primeiramente ser formulado através de uma função matemática, a melhor solução recai nos valores das variáveis que minimizam ou maximizam esta função, chamada de função objetivo, enquanto satisfazem algumas restrições. Sendo assim, um problema de otimização é baseado em três pontos importantes [Verma and Kumar, 2014]:

1. Função de otimização, também chamada de função objetivo ou *fitness*;
2. Conjunto de variáveis que afetam a função de otimização;
3. Conjunto de restrições que devem ser satisfeitas.

Em problemas que envolvem a otimização com um objetivo (uma função) apenas, o AG tentará encontrar uma solução ótima global que pode ser o valor mínimo ou máximo, dependendo, portanto, se a função deve ser minimizada ou maximizada. Nas otimizações multi-objetivo , ou seja, considerando mais de uma função objetivo, a tarefa passa a ser a busca por uma ou mais soluções factíveis, sendo que nenhuma delas pode ser dita melhor que as outras, levando em consideracão todos os objetivos, já que alguns objetivos são conflitantes. Nesse caso, pode-se ter um conjunto de soluções relacionadas, conhecidas como soluções eficientes de Pareto, as quais são comparadas através do conceito de dominância, abordado na seção 3.4.

3.2.2 Codificação do indivíduo

A codificação do problema consiste em como representar os genes do indivíduo [Sivanandam and Deepa, 2008], ou seja, em como as soluções do problema serão representadas. O cromossomo é um conjunto de genes e cada um deles representa algum dado. Dependendo do tipo de problema e do que se deseja manipular geneticamente, um indivíduo pode ser estruturado através de cromossomos codificados de maneira binária, real, inteira, hexadecimal, entre outras [Sivanandam and Deepa, 2008]. A codificação possui algumas restrições, tais como, os cromossomos que compõem o indivíduo devem ser do mesmo tipo e tamanho [Verma and Kumar, 2014].

Na codificação binária, os indivíduos são representados por conjuntos de valores binários (*bits*), os quais podem ser transformados em valores discretos (inteiros) ou contínuos (reais) através de uma conversão. Cada bit do conjunto pode representar alguma característica da solução. Este tipo de codificação é bastante utilizada, porém acarreta um aumento no tamanho do cromossomo, e, consequentemente, do indivíduo, caso o mesmo represente um número real com a precisão de muitas casas decimais, por exemplo.

A codificação real utiliza os próprios números reais para representar o cromossomo, sem necessitar, portanto, de uma conversão conforme abordado anteriormente. Isso agiliza a execução e deixa a formulação mais próxima da realidade do problema. Da mesma maneira que a real, existe a possibilidade da codificação inteira e hexadecimal. Esta última utiliza números hexadecimais ($0^{\smile}9, A^{\smile}F$) para codificar o cromossomo.

3.2.2.1 Função objetivo ou *fitness*

A função objetivo de um AG é definida com o intuito de avaliar quais indivíduos melhor representam a solução para o problema. Esta função avalia a aptidão dos indivíduos que irão guiar o processo de busca com o intuito de atingir a melhor solução, ou seja, obter em cada geração indivíduos cada vez mais aptos. Os indivíduos com melhor valor na função objetivo serão escolhidos para compor a próxima população. Esta avaliação é sempre realizada antes de compor uma população e esse processo evolutivo é repetido até que algum critério de parada seja alcançado. A busca pela melhor solução para o problema compreende várias etapas, entre elas, a seleção, o cruzamento e a mutação. Existem várias maneiras de executar essas etapas, sendo algumas delas abordadas nas seções seguintes.

3.2.2.2 Seleção

O processo de seleção visa escolher indivíduos aptos para realizar o cruzamento, ou seja, simula o mecanismo de seleção natural no qual os mais aptos terão chance de gerar mais descendentes para a próxima geração que os menos aptos e ,dessa maneira, perpertuar por várias populações seu material genético. Considerando f_i a função de avaliação do indivíduo i na população corrente, a probabilidade dele ser selecionado é proporcional a Equação 3.1, dado que N é o tamanho da população.

$$p_i = \frac{f_i}{\sum_{j=1}^{N} f_i} \tag{3.1}$$

Não obstante, por ocasião da seleção, nem todos os indivíduos com baixa aptidão devem ser descartados, pois eles podem conter características genéticas benéficas para a criação de indivíduos que possam vir a ser a melhor solução para o problema. Além disso, eliminar completamente esses indivíduos faz com que a população perca a sua diversidade, pois os indivíduos ficarão cada vez mais semelhantes, prejudicando, assim, o desempenho satisfatório do algoritmo. Para forçar que não apenas os mais aptos sejam escolhidos, utiliza-se a pressão seletiva na qual a variação da aptidão média da população é induzida pelo método de seleção. A intensidade dela por ser medida por:

$$I = \frac{M^* - M}{\sigma} \tag{3.2}$$

onde M é a avaliação média (aptidão) dos indivíduos da população anterior, M^* é o valor da avaliação média (aptidão) dos indivíduos da população atual, após a seleção, e σ é o desvio

padrão das avaliações das populações antes da seleção. Dentre os métodos de seleção mais utilizados destacam-se o proporcional (roleta), por torneio, por *rank* e por truncamento.

Na seleção proporcional cada indivíduo possui uma fatia proporcional à sua aptidão relativa, na qual a probabilidade de seleção é diretamente proporcional ao valor de sua função objetivo, de acordo com a Equação 3.3. A roleta é rodada tantas vezes quanto for a quantidade de indivíduos na população e, neste caso, o indivíduo com maior fatia proporcional tem maior chance de ser selecionado. Nessa seleção, a desvantagem é que podem ser selecionados indivíduos iguais, encolhendo a variabilidade da população e gerando, assim, problemas de convergência prematura.

$$p_i = \frac{f_i}{NM} \tag{3.3}$$

Na Equação 3.3, p_i é a probabilidade de seleção de um indivíduo i, f_i é a aptidão do mesmo e N é o tamanho da população.

Na seleção por torneio, um grupo de n indivíduos é escolhido aleatoriamente e entram em torneio para serem escolhidos como parte da próxima geração. O indivíduo vencedor é escolhido através de uma probabilidade k, definida, previamente, e somente ele é inserido na população seguinte. Este processo é repetido N vezes até que a nova população esteja formada.

Na seleção por *rank*, os indivíduos i são classificados de acordo com um *rank*, baseado no valor da função de avaliação. Os indivíduos com maior *rank* serão escolhidos. Isso evita a manutenção de superindivíduos ao mesmo tempo que mantém a pressão seletiva. Após a realização da classificação, o valor da aptidão é alterado de acordo com a posição no *rank*, calculada através da Equação 3.4. Aos indivíduos classificados como melhor e pior, são atribuídas as aptidões máxima (max) e mínima (min), respectivamente. Esses valores são determinados pelo usuário, porém, segundo Blickle [Blickle, 1996], existem as restrições de que $max = 2 - \min$ e $\min \geq 0$. Os demais indivíduos têm os valores de aptidão linearmente distribuídos entre os valores min e max, de acordo com a sua aptidão relativa na ordenação ($i = 1$ corresponde ao pior elemento).

$$E(i,n) = \min + \max - \min \times \frac{rank(i,n) - 1}{N - 1} \text{ onde } N \text{ é o tamanho da população.} \tag{3.4}$$

Na seleção por truncamento, somente uma porcentagem x da população será escolhida. Essa porcentagem varia entre 0 e 100% e existe a necessidade de classificar decrescentemente os indivíduos de acordo com a sua aptidão. Aqueles que se encontrarem a partir da posição 1 até à

relativa a porcentagem escolhida ($x\%$) serão escolhidos. Quando a porcentagem escolhida recai em valores pequenos, ocasiona uma perda da diversidade da população apesar de proporcionar uma convergência mais rápida. Esta seleção, assim como a anterior, possue a desvantagem do tempo gasto para realizar a ordenação.

3.2.2.3 Cruzamento

Após a etapa de seleção, realiza-se a recombinação ou cruzamento dos indivíduos com o objetivo de combinar material genético e gerar indivíduos com as melhores características dos pais. Dessa maneira, dois indivíduos são escolhidos aleatoriamente para pais e o resultado da recombinação é a obtenção do filho. A recombinação ocorre considerando um operador ou taxa de cruzamento com probabilidade p_c e novos indivíduos, apesar de possuírem características genéticas de seus pais, são diferentes de ambos. Quanto maior a taxa de cruzamento, maior a quantidade de novos indivíduos que serão introduzidos na população. Se o valor da taxa de cruzamento for alto, pode acontecer a perda de indivíduos já bem adaptados, por exemplo. Ajustando-a para um valor menor, o AG pode tornar-se muito lento até convergir. O valor ideal para esta taxa depende do problema, podendo ser um valor fixo ou variável ao longo das gerações. Dentre as várias maneiras de realizar o cruzamento para cromossomos com representação por números reais, destacam-se o aritmético e o heurístico.

No cruzamento aritmético, os cromossomos c_1 e c_2 gerados a partir dos pais p_1 e p_2 são dados pelas Equações 3.5 e 3.6, onde $\beta \in U(0,1)$, sendo $U(0,1)$ uma distribuição uniforme.

$$c_1 = \beta \times p_1 + (1 - \beta) \times p_2 \tag{3.5}$$

$$c_2 = (1 - \beta) \times p_1 + \beta \times p_2 \tag{3.6}$$

No cruzamento heurístico, é necessário conhecer o valor da função objetivo dos pais. Os descendentes são originados a partir de uma interpolação linear entre os pais usando a aptidão e favorecendo o pai mais bem adaptado. A Equação que define o cruzamento é dada por 3.7, que o valor da variável r está entre 0 e 1 e a aptidão do indivíduo p_1 é maior que a aptidão do indivíduo p_2 ($f(p_1) > f(p_2)$). Existem ainda outras variações de cruzamento como o parcial, cíclico, sequencial entre outros [Soares, 1997].

$$c = p_1 + r(p_1 + p_2) \tag{3.7}$$

3.2.2.4 Mutação

A mutação é utilizada para aumentar a diversidade da população com a finalidade de explorar melhor o espaço de busca e evitar a convergência prematura. Este operador escolhe alguns genes nos cromossomos para serem alterados, sendo que a quantidade de indivíduos que sofrerão alteração em seu código genético depende da taxa de mutação p_m. Entretanto, deve-se evitar uma taxa de mutação muito alta, um vez que esta pode tornar a busca essencialmente aleatória, prejudicando fortemente a convergência para uma solução ótima. A mutação pode ser, por exemplo, do tipo uniforme, não-uniforme, gaussiana, *creep*, entre outras.

Nas mutações uniforme e gaussiana, um gene selecionado é substituído por outro gerado aleatoriamente seguindo distribuições uniforme e gaussiana, respectivamente. Na mutação não-uniforme, os novos valores possíveis do gene obedece uma distribuição não-uniforme e na *creep* um valor aleatório é acrescentado ou subtraído do gene, obtido através de uma distribuição normal $N(0, \sigma^2)$, onde a variância assume um valor pequeno.

3.2.2.5 Elitismo

O elitismo é um operador utilizado para manter as melhores soluções para as próximas gerações, garantindo assim a preservação dos melhores indivíduos. Uma maneira simples de implementar o elistimo é passar uma determinada porcentagem ($x\%$) de indivíduos para a próxima geração, sendo os demais indivíduos necessários para completar a população obtidos através do cruzamento e da mutação da geração atual. Este operador deve ser muito bem ajustado, pois, caso contrário, o algoritmo pode convergir precocemente e ficar preso em um extremo local.

3.2.3 Características da população e número de gerações

Antes de começar a execução de um AG, é necessária a criação da população inicial, que, tipicamente, é formada por indivíduos gerados aleatoriamente. Porém, mesmo nessa aleatoriedade, os indivíduos podem ser "semeados" com bons cromossomos para agilizar o processo de evolução.

Uma outra característica que deve ser observada em uma população é o seu tamanho, pois este afeta o desempenho global e a eficiência do AG. Populações muito pequenas oferecem uma cobertura igualmente pequena do espaço de busca, afetando o desempenho e a eficiência do algoritmo. Em contrapartida, populações suficientemente grandes fornecem uma melhor

cobertura do domíinio do problema e previnem a convergência prematura para soluções locais, porém são necessários recursos computacionais maiores ou um tempo maior de processamento do problema.

O número de gerações representa a quantidade de ciclos de evolução que o AG executará até encontrar um critério de parada preestabelecido. Quando o número de gerações é pequeno, há uma queda no desempenho do algoritmo, pois ele não terá ciclos para evoluir a população satisfatoriamente. Já um valor exorbitante, exige mais recurso computacional, porém fornece mais tempo para uma melhor cobertura do domínio do problema.

3.3 Problema de Otimização Multi-Objetivo

Um problema de otimização é dito multi-objetivo quando possui várias funções objetivo que devem ser maximizadas ou minimizadas, obedecendo um determinado número de restrições que qualquer solução viável deve obedecer, incluindo até mesmo as soluções consideradas ótimas. De acordo com Deb [Deb, 2001], um problema de OMO pode ser caracterizado pela Equação 3.8 .

$$\left.\begin{array}{lll} \text{Maximizar/Minimizar} & f_m(x), & m = 1, 2, ..., M; \\ \text{sujeito à} & g_j(x) \geq 0 & j = 1, 2, ..., J; \\ & h_k(x) = 0, & k = 1, 2, ..., K; \\ & x_i^{(L)} \leq x_i \leq x_i^{(U)}, & i = 1, 2, ..., n. \end{array}\right\} \tag{3.8}$$

onde $\boldsymbol{x}$ é um vetor de n variáveis de decisão: $\boldsymbol{x} = (x_1, x_2, ..., x_n)^T$. As restrições das variáveis de limite (x_i) restringem cada variável de decisão entre o limite inferior $x_i^{(L)}$ e superior $x_i^{(U)}$. Esses limites representam o espaço de variáveis de decisão, ou simplesmente, o espaço de decisão. Os termos $g_j(x)$ e $h_k(x)$ são funções de restrição e uma solução qualquer x que não satisfaça todas restrições e os *2n* limites será considerada um solução não-factível. Caso contrário, é considerada uma solução factível. O conjunto de todas as soluções viáveis denomina-se região viável, espaço de busca ou simplesmente S. As funções objetivo $f_1(x)$, $f_2(x)$, ..., $f_M(x)$ podem ser maximizadas ou minimizadas dependendo da característica do problema.

3.4 Dominância e Soluções Eficientes de Pareto

No geral, os algoritmos de otimização multi-objetivo baseiam suas buscas no conceito de dominância, no qual duas soluções são comparadas para verificar se alguma estabelece relação de dominância sobre a outra. A partir dessa análise é montado um conjunto de soluções eficientes de Pareto. Considerando um problema com M funções objetivos, onde $M > 1$, a relação de dominância entre duas soluções $x^{(1)}$ e $x^{(2)}$ pode ser definida como : a solução $x^{(1)}$ domina a solução $x^{(2)}$ e vice-versa, ou não há dominância entre elas. Para uma solução $x^{(1)}$ dominar outra $x^{(2)}$ é necessário que duas condições sejam satisfeitas [Deb, 2001].

1. A solução $x^{(1)}$ é não pior que $x^{(2)}$ em todos os objetivos, ou $f_i(x^{(1)})$ **não** $\prec f_i(x^{(1)})$ para todos $j = 1, 2, ..., M$ objetivos;
2. A solução $x^{(1)}$ é estritamente melhor que $x^{(2)}$ em ao menos um objetivo, ou $f_j(x^1) \succ f_i(x^{(2)})$ para ao menos um $j \in 1, 2, ..., M$.

onde considera-se que o operador $\prec$ denota pior e o operador $\succ$ denota melhor. Se qualquer uma das condições acima é violada, a solução $x^{(1)}$ não domina a solução $x^{(2)}$. Se $x^{(1)}$ dominar $x^{(2)}$ ($x^{(1)} \succ x^{(2)}$) pode-se afirmar que:

- $x^{(2)}$ é dominada por $x^{(1)}$;
- $x^{(1)}$ não é dominada por $x^{(2)}$;
- $x^{(1)}$ é não pior do que $x^{(2)}$.

Na Figura 3.3, tem-se duas funções objetivo f_1 e f_2, nas quais f_1 deve ser maximizada e f_2 minimizada, e cinco soluções com valores diferentes. Como as duas funções são importantes, é difícil encontrar uma solução ótima que satisfaça ambas. Neste caso, o conceito de dominância auxilia na decisão pela melhor solução por ocasião da comparação entre elas. Comparando as solucões 1 e 2, considera-se que as condições de dominância foram satisfeitas e a solução 1 domina a solução 2. Avaliando a dominância entre 1 e 5, esta última é melhor em relação ao primeiro objetivo e não é pior em relação a solução 1 no segundo objetivo. Portanto, a solução 5 domina a 1.

A razão para a optimalidade de muitas soluções é que nesse conjunto não existe uma única solução que pode ser considerada melhor em todos os objetivos. A este conjunto é dado o nome de soluções eficientes de Pareto.

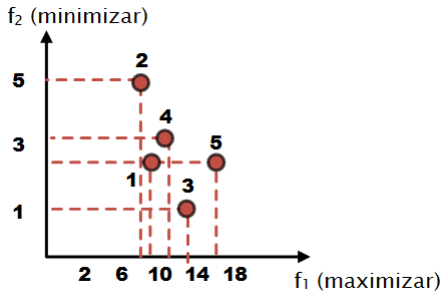

Figura 3.3: Conceito de dominância, adaptado de [Deb, 2001].

O conceito de dominância pode ser aplicado para definir conjuntos de soluções eficientes locais e globais. O conjunto ótimo de Pareto local é definido quando, para cada elemento x pertencente ao conjunto $\bar{P}$, não existir uma solução y na sua vizinhança que domine outro elemento do conjunto $\bar{P}$, caracterizando as soluções pertencentes a $\bar{P}$ como um conjunto eficiente local de Pareto. A curva que junta estas soluções é conhecida como frente ótima de Pareto [Raghuwanshi and Kakde, 2004]. Se não existe solução no espaço de pesquisa que domina qualquer membro no conjunto $\bar{P}$, então as soluções pertencentes ao conjunto $\bar{P}$ constituem um conjunto ótimo global de Pareto.

Na presença de múltiplas soluções eficientes de Pareto, é difícil escolher uma única solução sem nenhuma informação adicional sobre o problema. Devido a isso, é importante achar quantas soluções eficientes de Pareto forem possíveis, obedecendo aos seguintes objetivos:

1. Guiar a busca o mais perto possível para a região ótima de Pareto global e;
2. Manter a diversidade da população na frente de Pareto ótima.

3.5 Algoritmo NSGA-II - *Non-Dominated Sorting Genetic Algorithm II*

O NSGA-II (Non-Dominated Sorting Genetic Algorithm II) é um algoritmo genético multi-objetivo (AGMO) capaz de encontrar soluções bem espalhadas sobre a frente ótima de Pareto, necessitando de baixo esforço computacional. Deb [Deb et al., 2000] propôs este método como uma modificação do algoritmo original [Srinivas and Deb, 1994]. Dentre as características principais, destacam-se o elitismo, a atribuição de *rank* e a distância da multidão.

O elitismo é utilizado como mecanismo para a preservação e usabilidade das melhores

soluções encontradas previamente em gerações posteriores. Através do *rank*, o algoritmo realiza o ordenamento das soluções não-dominantes da população. A distância da multidão utiliza um operador de seleção por torneio para preservar a diversidade entre as soluções não dominadas nos estágios de execução posteriores para obter um bom espalhamento das soluções.

A Figura 3.4 ilustra o funcionamento do NSGA-II no qual a população Q_t é criada a partir da população pai P_t, onde ambas possuem N indivíduos e são combinadas para juntas formar a população R_t, de tamanho $2N$. Após essa junção, é então realizado um ordenamento das melhores soluções para classificar a população inteira R_t. Apesar de requerer maior esforço computacional, o algoritmo permite checar uma não-dominância global entre as populações P_t e Q_t. Com a finalização do ordenamento das soluções não-dominantes, um novo conjunto P_{t+1} é criado e preenchido por soluções de diferentes frentes não-dominadas $(F_1, F_2, ..., F_n)$. O preenchimento começa com as melhores soluções da primeira frente não-dominada, seguindo com as frentes subsequentes. Como somente N soluções podem ser inseridas na nova população, as demais soluções restantes são simplesmente descartadas. Cada conjunto F_i deve ser inserido em sua totalidade na nova população (P_{t+1}) e quando $|P_{t+1}| + |F_i| > N$ o algoritmo introduz um método chamado de distancia de multidão (*crowding distance*), onde são preferidas as soluções mais dispersas do conjunto F_i e retiradas as demais.

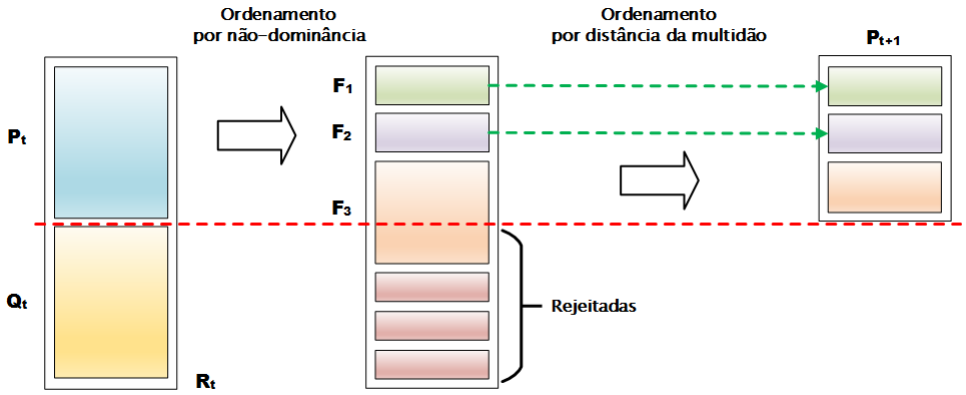

Figura 3.4: Esquema dos algoritmo NSGA-II.

Para verificar a distância da multidão, primeiramente, é calculada a distância média dos dois pontos, de ambos os lados desse ponto, levando em consideração todos os objetivos [Deb, 2001, Carvalho and Araújo, 2009]. A quantidade d_i serve como uma estimativa do tamanho do maior cubóide que inclui o ponto i sem incluir qualquer outro ponto da população, sendo esta chamada distância da multidão. Na Figura Figura 3.5, a distância da i-ésima solução na sua frente de Pareto (pontos preenchidos) é a média do comprimento lateral do cubóide desenhado pelas linhas tracejadas.

O operador que realiza a comparação da multidão incorpora uma modificação no método de seleção por torneio que leva em conta a distância da multidão de uma solução (*crowded tournament selection operator*). Portanto, uma solução i é considerada ganhadora em um torneiro contra uma solução j, se obedecer as restrições seguintes:

1. A solução i possui um melhor *rank* de não-dominância na população;
2. Se ambas soluções estão no mesmo nível, mas i tem uma distância de multidão maior, $d_i > d_j$;

Considerando duas soluções com diferentes níveis de não-dominância, os pontos escolhidos são aqueles com menor nível. Se ambos os pontos pertencem a mesma frente, então são escolhidos pontos localizados em uma região com menor número de pontos, ou seja, soluções com maiores distâncias de multidão.

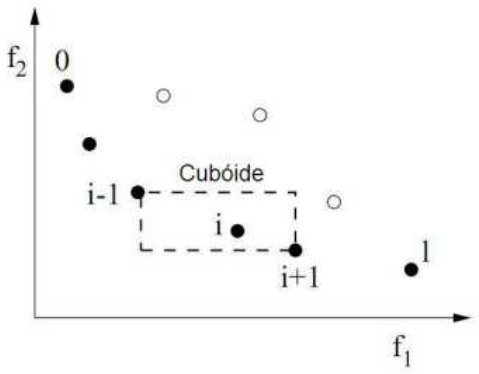

Figura 3.5: Distância da multidão adptada de [Deb, 2001, Carvalho and Araújo, 2009].

3.5.1 Cruzamento e Mutação

O operador de cruzamento adotado pelo NSGA-II chama-se SBX (*Simulated Binary Crossover*) e simula o cruzamento com um ponto de corte (Seção 3.2.2.3) utilizado na codificação binária aplicado à codificação real e utilizando uma distribuição de probabilidades em torno de dois pais para criar as soluções descendentes. Esse tipo de codicação mostrou-se bastante adequada para resolver problemas de otimização quando o espaço de busca é contínuo ou existem várias soluções ótimas.

O operador SBX inicia com o cálculo do fator de propagação β o qual é obtido a partir da Equação 3.9 na qual c_1 e c_2 são filhos criados a partir dos pais p_1 e p_2.

$$\beta_i = \left| \frac{c_1 - c_2}{p_1 - p_2} \right| \tag{3.9}$$

A partir de β pode-se calcular a distribuição de probabilidade em torno dos pais que é dada pela Equação 3.10 [Raghuwanshi and Kakde, 2004].

$$P(\beta_i) = \begin{cases} 0.5(\eta+1)\beta_i^{\eta}, \ se \ \beta_i \leq 1 \\ 0.5(\eta+1)\beta_i^{\frac{1}{\eta+2}}, \ caso \ contrário \end{cases} \tag{3.10}$$

Se o índice de distribuição η for um valor alto existe uma probabilidade grande de obter filhos próximos aos pais. Caso contrário, as soluções obtidas para os filhos será distante dos pais. O número aleatório u_i é obtido a partir da distribuição de probabilidade $P(\beta)$ e utilizado para calcular a Equação 3.11.

$$\beta q_i = \begin{cases} (2u_i)^{\frac{1}{\eta+1}}, \ se \ u_i \leq 5 \\ \left[\frac{1}{2(1-u_i)}\right]^{\frac{1}{\eta+1}}, \ caso \ contrário \end{cases} \tag{3.11}$$

Considerando βq_i calculado anteriormente (Equação 3.11), os filhos y_i^1 e y_i^2 são obtidos através da Equação 3.12 na qual x^{i^1} e x^{i^2} são os pais.

$$y_i^1 = 0.5 \left[(1+\beta q_i) x^{i^1} + (1-\beta q_i) x^{i^2} \right] \quad y_i^2 = 0.5 \left[(1-\beta q_i) x^{i^1} + (1+\beta q_i) x^{i^2} \right] \tag{3.12}$$

O operador de mutação empregado no NSGA-II chama-se mutação polinomial [Raghuwanshi and Kakde, 2004] e segue um pouco a metodologia empregada no cruzamento SBX. A probabilidade de mutação é obtida através da variável P_m, sendo n o número de variáveis de decisão e η_m o índice de distribuição o qual pode ter qualquer valor não negativo. Para cada variável de decisão x_i são definidos os limites inferior e superior $\left[x_i^{Lower}, x_i^{Upper}\right]$. Esta mutação ocorre da seguinte maneira [Hamdan, 2012]: cada variável de decisão X_i tem uma probabilidade P_m de ser perturbada. Considerando uma variável aleatória r a qual possui valor entre o intervalo 0 e 1, se $r \leq P_m$ então calcula-se δ_1 e δ_2 conforme as Equações 3.13 e 3.14 respectivamente.

$$\delta_1 = \frac{X_i - X_i^{Lower}}{X_i^{Upper} - X_i^{Lower}} \tag{3.13}$$

$$\delta_2 = \frac{X_i^{Upper} - X_i}{X_i^{Upper} - X_i^{Lower}} \tag{3.14}$$

Na sequência, um novo valor aleatório é sorteado para r com o objetivo de calcular δ_q que é utilizado para se obter o novo valor do indivíduo X_i, conforme as Equações 3.15 e 3.16.

$$\delta_q = \begin{cases} [(2r) + (1-2r) \times (1-\delta_1)^{\eta_m+1}]^{\frac{1}{\eta_m+1}} - 1, \ se\ r \leq 0.5 \\ 1 - [2(1-r) + 2(r-0.5) \times (1-\delta_2)^{\eta_m+1}]^{\frac{1}{\eta_m+1}}, \ caso\ contrário \end{cases} \tag{3.15}$$

$$X_i = X_i + \delta_q(X_i^{Upper} - X_i^{Lower}) \tag{3.16}$$

3.6 Conclusões sobre o capítulo

Neste capítulo, foram apresentados os conceitos e o funcionamento básico de algoritmos, especificamente do algoritmo NSGA-II, o qual foi utilizado no arcabouço, desenvolvido para estimar os valores dos parâmetros de entrada de sintetizadores por formantes. O NSGA-II apresentou características pertinentes para que fosse aplicado ao problema aqui apresentado, pois permite a codificação real, utiliza o mecanismo de elitismo assim como permite impor limites (superiores e inferiores) aos valores dos genes modelados no cromossomo, conforme é abordado no capítulo seguinte.

Capítulo 4

Imitação da Voz utilizando Algoritmo Genético

Este capítulo apresenta a descrição deste trabalho que compreende: o estudo realizado sobre o sintetizador de Klatt, o software utilizado como baseline para a pesquisa e o arcabouço desenvolvido, chamado *newGASpeech*, baseado em um procedimento de análise-por-síntese e em algoritmo genético para a solução do problema.

4.1 Descrição do problema

Como citado, o presente trabalho tem como principal objetivo estimar os valores dos parâmetros de entrada de um sintetizador por formantes, como o Klatt e o *HLSyn*, por exemplo, visando imitar a voz humana. Este problema é considerado difícil, uma vez que os parâmetros especificam a temporização da fonte e os valores dinâmicos para todos os filtros. Dependendo da quantidade de parâmetros envolvidos, a possibilidade de combinações possíveis pode ser muito grande e inviável de ser realizada manualmente, pois cada parâmetro possui um vasto intervalo de valores cabíveis. De acordo com a Figura 4.1, é necessário estimar inicialmente valores para os parâmetros de entrada do sintetizador, submetê-los para a síntese e, em seguida, avaliar a voz sintetizada em relação à voz alvo, através de algum mecanismo de comparação. Após essa verificação, caso a voz sintética ainda não esteja similar à voz alvo, os valores dos parâmetros devem ser ajustados, ou seja, novos valores são reestimados, é feita a síntese da voz e posterior comparação (análise-por-síntese), até que a voz gerada esteja o mais próxima possível da voz alvo.

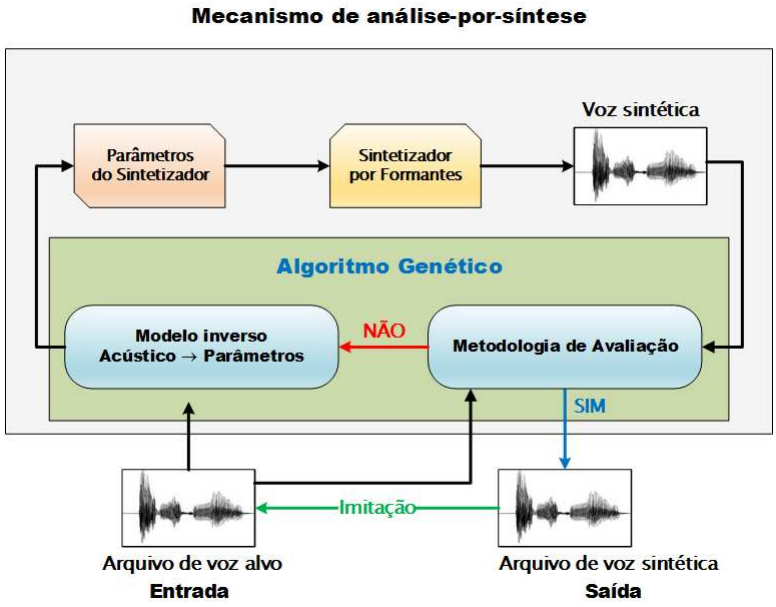

Figura 4.1: Descrição do problema.

A dificuldade maior consiste em extrair os valores dos parâmetros de um sintetizador de voz a partir de um arquivo de voz (sistema STS - *Speech-To-Speech*), ou seja, realizar uma análise acústica e, a partir de um modelo inverso, extrair os parâmetros do sintetizador. Esses parâmetros podem ser gerados atualmente através de sistemas TTS, como o Dectalk [Hallahan, 1995], por exemplo, apenas para vozes sintéticas e específicas para alguns falantes previamente definidos no software. Algumas ferramentas e técnicas que utilizam processamento de sinais surgiram para tentar extraí-los a partir de arquivos de voz e não obtê-los, apenas, através de sistemas TTS, como é o caso do software utilizado como baseline (abordado na Seção 4.3).

Considerando a complexidade do problema, o objetivo é desenvolver um arcabouço para estimar automaticamente os valores dos parâmetros de entrada de um sintetizador por formantes, através de um mecanismo de análise-por-síntese, com o objetivo de gerar voz sintética o mais próxima possível da voz alvo. Neste trabalho, os sintetizadores adotados são Klatt (versão KLSYN88) e *HLSyn*. A seção a seguir, aborda particularidades do sintetizador de Klatt em virtude do *HLSyn* ser baseado nele para gerar vozes sintéticas.

4.2 Estudo sobre o Sintetizador de Klatt

Na Seção 2.4.1.1, especificou-se a versão do sintetizador chamada Klatt88 ou KLSYN88. Por ser a mais recente, essa versão é composta por quarenta e oito parâmetros e cada um deles

possui um intervalo de valores próprios. Dependendo do parâmetro, esses intervalos podem ser bem extensos, acarretando um aumento considerável na dimensão do espaço de busca do problema. Sendo assim, foi realizado um estudo nos arquivos do Klatt obtidos através do TTS *Dectalk*, com o objetivo de identificar quais parâmetros apresentam comportamento estático, além de ratificar os intervalos de valores possíveis para cada um deles.

Para isso, de acordo com [Rothauser et al., 1969], duzentas e quarenta sentenças em inglês foram submetidas à esse TTS (ver apêndice D), vozes sintéticas foram produzidas para seis falantes diferentes e agrupadas em duas categorias, tais como, masculina e feminina. Histogramas de todos os parâmetros do Klatt foram gerados e, a partir deles, foi possível identificar que o TTS impôs variação nos valores de apenas vinte e cinco parâmetros, independente da categoria do falante, os quais estão grafados em preto na Figura 4.2. Os demais parâmetros (em vermelho) mantiveram-se constantes sendo que FL, $DF1$, $DB1$ e $A6F$ apresentaram valor igual a zero. Os parâmetros com valores constantes, diferente de zero, estão listados na Tabela 4.1 e aqueles que variaram ao longo do tempo são apresentados na Tabela 4.2. Os intervalos de valores dos parâmetros AF, $B1$, FNP e FNZ, Klatt - versão 1990 [Klatt and Klatt, 1990], tiveram que ser expandidos.

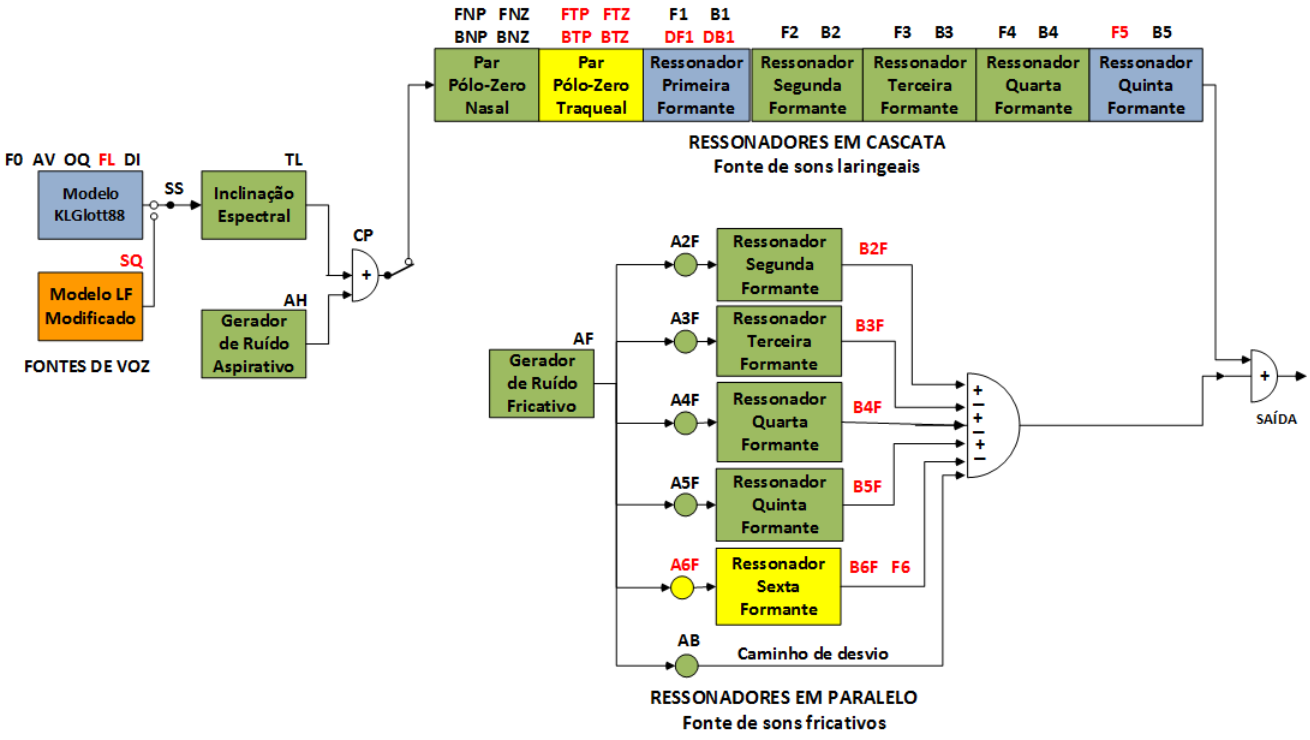

Figura 4.2: Versão Klatt88 adaptado de [Klatt and Klatt, 1990].

A Figura 4.3 apresenta o exemplo de um arquivo de entrada do Klatt88 em que ele é

Tabela 4.1: Parâmetros do Klatt com valores constantes diferente de zero.

Parâmetro	Valor	Parâmetro	Valor
F5	4500	BTZ	200
F6	4990	B2F	250
B6	1000	B3F	320
FTP	1000	B4F	350
BTP	200	B5F	500
FTZ	1000	B6F	1500

Tabela 4.2: 25 parâmetros variantes na versão Klatt88.

P.	Min	Max	Unid.	P.	Min	Max	Unid.
F0	0	5000	Hz	F4	2400	4990	Hz
AV	0	80	dB	B4	3000	4990	Hz
OQ	0	99	%	B5	100	1500	Hz
TL	0	41	dB	FNP	450	870	Hz
DI	0	100	%	BNP	40	1000	Hz
AH	0	80	dB	FNZ	180	1000	Hz
AF	0	70	dB	BNZ	40	1000	Hz
F1	180	1300	Hz	A2F	0	80	dB
B1	30	1040	Hz	A3F	0	80	dB
F2	550	3000	Hz	A4F	0	80	dB
B2	40	1000	Hz	A5F	0	80	dB
F3	1200	4800	Hz	AB	0	80	dB
B3	60	1000	Hz	-	-	-	-

formado por várias linhas e colunas. Cada linha representa um segmento (quadro) de voz e cada coluna é um parâmetro do Klatt.

Figura 4.3: Arquivo de entrada do Klatt88.

4.3 Software Winsnoori

O Winsnoori, versão 1.34, é um software desenvolvido com os objetivos de prover a análise e edição de sinais de voz para pesquisadores e estudantes que trabalham nesta área. A ferramenta contém basicamente as seguintes funcionalidades:

- Edição de sinais de voz;
- Realiza o cálculo e visualização do espectograma;
- Permite anotações fonéticas e ortográficas nos sinais de voz;
- Contabiliza dinamicamente o resultado de várias análises espectrais;
- Monitora a trajetória das formantes;
- Extrai os parâmetros do sintetizador de Klatt (versão Jon Iles).

A vantagem dessa ferramenta é a interface gráfica para o sintetizador de Klatt, versão de Jon Iles (Seção 2.4.1), que permite a síntese através da cópia do sinal, ou seja, determina a cada quadro, a frequência e amplitude das formantes, permitindo a reprodução do sinal o mais próximo possível do sinal original. O usuário tem a opção de extrair automaticamente a trajetória das formantes através de um algoritmo ou editar manualmente esta trajetória diretamente no espectograma. Além disso, a ferramenta possui um detector de frequência fundamental que foi desenvolvido a partir de um algoritmo proposto por Martin [Martin, 1981].

O Winsnoori realiza a síntese através da cópia do sinal, gerando voz sintética inteligível e bem próxima à voz natural. Não obstante, pequenas imperfeições características da voz também são "copiadas"com essa técnica, ou seja, durante a cópia não há mecanismos para a melhora do sinal. A versão citada funciona na plataforma Windows e foi descontinuada pela empresa desenvolvedora.

4.4 Metodologia para Imitar Voz através de Algoritmo Genético

4.4.1 Descrição da metodologia

Com o objetivo de automatizar a imitação da voz, foi desenvolvido um arcabouço, o qual faz uso de algoritmo genético e do procedimento de análise-por-síntese. Adaptou-se o NSGA-II para que este trabalhasse *Intraframe* e *Interframe*. Considerando que um arquivo de voz é composto por vários quadros, na metodologia *Intraframe*, assume-se que cada quadro é um problema convencional de AG. Sendo assim, por exemplo, se a voz alvo tem duração de um segundo e cada quadro de dez milisegundos (sem superposição), então cem problemas de AG são resolvidos independentemente.

Já na metodologia *Interframe*, os melhores indivíduos da última população do quadro t (obtiveram *rank* =1) são copiados para inicializar parte da população do quadro $t+1$, pois os valores dos parâmetros do sintetizador de voz por formantes não apresentam mudança considerável de um quadro de voz para o outro subsequente. Isso acontece devido ao fato dos parâmetros do sintetizador modelar aspectos do aparelho fonador que são dinâmicos, porém com variações sutis.

Na metodologia *Interframe*, uma grande quantidade de indivíduos pode estar aptos à cópia para inicializar parte da população do quadro seguinte. Nessa perspectiva, configura-se em um arquivo de entrada a porcentagem máxima dos indivíduos, que podem ser copiados para a população inicial do próximo quadro. Os demais indivíduos necessários para compor a quantidade total de indivíduos da população são inicializados randomicamente [Couto and Borges, 2008] [Borges et al., 2008].

O arcabouço possui três opções para que a execução do algoritmo seja finalizada. São elas:

- **Convergência:** A execução é encerrada quando há a convergência da função objetivo para o valor esperado.
- **Número máximo de gerações:** Este critério é utilizado em AGs tradicionais e finaliza a execução quando o número máximo de gerações é alcançado. Este valor é configurado pelo usuário.
- **Número de gerações sem evolução:** Neste critério, quando o quadro atinge uma

porcentagem do número máximo de gerações sem evoluir, a simulação para. Este valor é configurado pelo usuário e leva em consideração o grau de diversidade da população através da existência de soluções dominadas, conceito abordado na seção 3.4. Se existirem soluções dominadas, então, a população continua evoluindo. Caso contrário, a população não está evoluindo e a quantidade de gerações nessa condição é contabilizada. Configura-se um parâmetro de entrada do AG, que é a porcentagem de gerações sem evoluir em relação à quantidade total de gerações. Caso o algoritmo alcance essa porcentagem, ele para de executar.

Um indivíduo no *newGASpeech* é composto por um vetor de parâmetros e, para cada quadro, um único indivíduo deve ser escolhido para compor o arquivo de saída, com vários quadros de voz, que é sintetizado ao final. Pode-se encontrar mais do que uma solução viável e, neste caso, o arcabouço é configurado para escolher a solução mais eficiente que minimiza ou maximiza as funções objetivo. Caso ele não encontre indivíduos com essa característica, o processo de decisão, pelo mais efeiciente, é realizado conforme o NSGA-II nativo (seção 3.5), baseado no elitismo, *rank* e distância da multidão.

Três funções objetivo estão implementadas no arcabouço. São elas: log da Distância Espectral (D_{LE}), Erro Quadrático Médio (EQM) e a Correlação Cruzada (CC). Considera-se que quanto menor for o valor das três funções objetivo, melhor é o indivíduo, ou seja, busca-se minimizar os valores das funções.

O D_{LE}, também conhecida como distorção espectral, é a medida da distância entre dois espectros, calculada em dB. Assumindo sinais discretos no tempo, $x[n]$ e $y[n]$, a D_{LE} é dada por:

$$D_{\mathrm{LE}} = \sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi}\left[10\log_{10}\frac{|X(e^{j\Omega})|}{|Y(e^{j\Omega})|}\right]^2 d\Omega} \tag{4.1}$$

onde $X(e^{j\Omega})$ e $Y(e^{j\Omega})$ são as DTFT dos sinais alvo e sintético, respectivamente.

Quando a Equação 4.1 é aproximada usando-se FFT (*Fast Fourier Transform*), o espectro da potência é calculado para os sinais discretizados da voz alvo ($x[n]$) e sintética ($y[n]$). Em seguida, o log do espectro dos sinais é obtido para inseri-los na Equação 4.1.

O EQM mede as variações do sinal referente à voz sintética em relação ao sinal da voz alvo [Imbens et al., 2005]. O cálculo é realizado através da Equação 4.2, a seguir.

$$EQM = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(x(j)-y(j))^2} \tag{4.2}$$

onde n é o número de amostras por quadro, $x(j)$ e $y(j)$ são as amostras de índice j de cada quadro da forma de onda da voz alvo e da voz sintética, respectivamente.

A CC mede a semelhança entre dois sinais. Considerando duas sequências $x(i)$ e $y(i)$ em que $i = 0, 1, 2...N-1$ são as amostras dos sinais. A correlação no deslocamento de tempo l é definida como:

$$r(l) = \sum_{i=1}^{N} [(x_{(i+l)} - \overline{x})(y_{(i)} - \overline{y})] \tag{4.3}$$

onde $\overline{x}$ e $\overline{y}$ são as médias dos sinais x e y, respectivamente.

Nos AGs tradicionais, os valores das probabilidades de cruzamento e mutação são fixos, predefinidos antes de iniciar a execução do algoritmo. Entretanto, essa opção pode ser ineficiente, em alguns casos, uma vez que pode levar o algoritmo a mínimos locais. Por essa razão, Ho [Ho et al., 1999] propôs uma heurística para que esses parâmetros pudessem ter seus valores adaptados, porém controlados. Essa estratégia visa variar essas probabilidades, iniciando com valores altos e decaindo para valores mais baixos, considerando, dessa maneira, que no início existe pouca informação sobre o domínio do problema e deve-se ter uma maior diversidade da população (*exploring*). Ao final do processo de otimização, já se possui um certo conhecimento do domínio e as melhores soluções precisam ser aprofundadas (*exploiting*). Sendo assim, o (newGASpeec) pode ser executado com valores fixos para as probabilidades de mutação e cruzamento ou adaptativos conforme as Equações 4.4 e 4.5.

$$p_m^{n+1} = p_m^n - p_m^n x \delta_m \tag{4.4}$$

$$p_c^{n+1} = p_c^n - p_c^n x \delta_c \tag{4.5}$$

onde δ_m e δ_c são as taxa de decaimento para a mutação e o cruzamento, respectivamente, considerando um valor inicial configurado para as probabilidades de mutação e cruzamento (p_m^0 e p_c^0) e valores mínimos que eles podem assumir ($min(p_m)$ e $min(p_c)$). Ambas as taxas de decaimento são configuradas pelo usuário e informadas ao AG, através do arquivo de configuração dado como entrada.

O processo de análise-por-síntese inicia com os arquivos de voz e de configuração, dados como entrada (Figura 4.4 - Etapa 1). O arquivo de configuração contém informações importantes como o número máximo de gerações, a quantidade de indivíduos nas gerações, as probabilidades iniciais de cruzamento e mutação entre outras. O arquivo de voz é segmentado em quadros de aproximadamente 5 ms e, para cada um deles, um AG completo com várias gerações é executado (Etapa 2). Os parâmetros para sintetizar um quadro compõe um

cromossomo o qual possui seu valor de função objetivo calculado a partir de uma ou mais funções objetivo. Cada quadro, ao longo da execução do AG, existe um cuidado especial para reinicializar o estado do sintetizador (memória de seus ressonadores digitais, assim por diante) a cada interação. Após a avaliação de toda a população, um *rank* é atribuído à cada indivíduo e aqueles com melhores valores de *rank* são selecionados para sofrer cruzamento e mutação. Como resultado, uma nova população é gerada e submetida novamente à avaliação, seleção, cruzamento e mutação. Todo o processo é repetido até que o algoritmo atinja um dos critérios de parada (Etapa 2), descritos anteriormente. Os melhores indivíduos de cada quadro compõem um arquivo de entrada do sintetizador o qual é sintetizado ao final, gerando um arquivo de voz sintética que visa imitar a voz alvo (Passos 3-6).

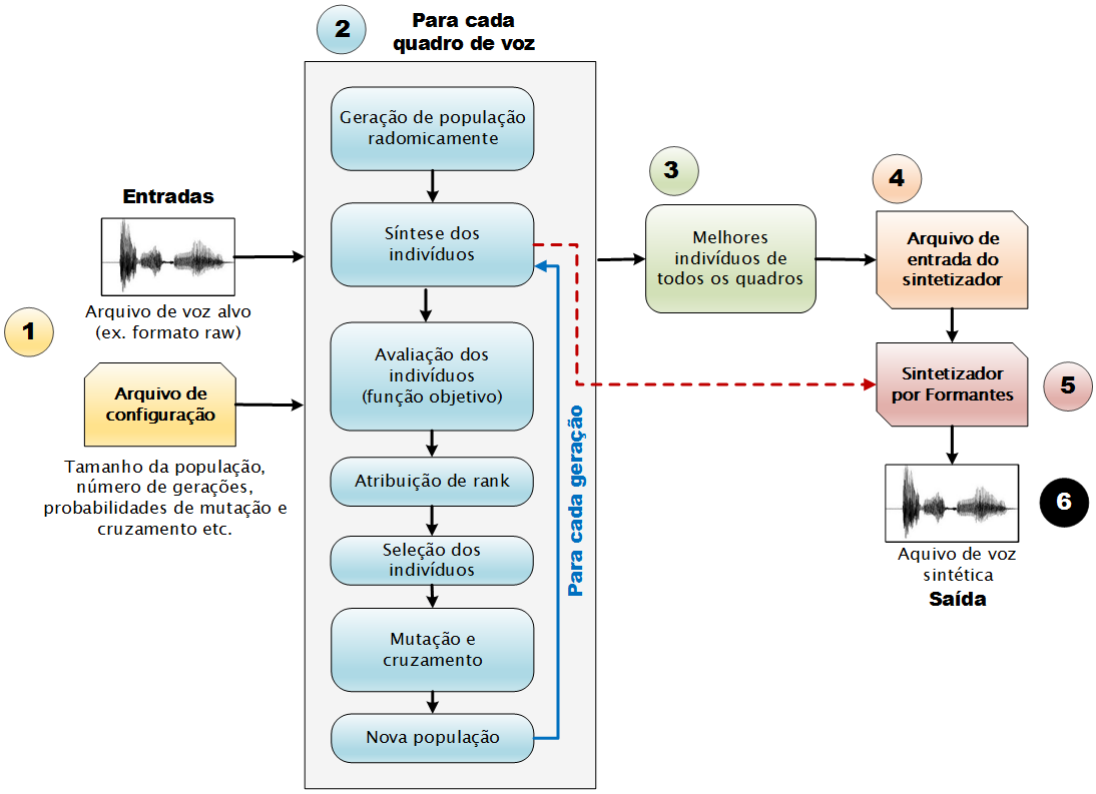

Figura 4.4: Fluxograma do funcionamento do *newGASpeech*.

Os genes que compõem o cromossomo são divididos em quatro seções, como, por exemplo, gene de vozeamento (G_v), fonte de voz (F_v), trato vocal (T_v) e Look-Ahead (L_a). Estas particularidades são abordadas na seção a seguir.

4.4.2 Codificação e Decodificação do Cromossomo

Cada cromossomo no *newGASpeech* possui genes codificados através de números reais, contendo informações sobre os valores dos parâmetros de entrada do sintetizador. Os genes são organizados em seções (Figura 4.5) e a quantidade de genes em cada uma delas depende da configuração adotada, ou seja, da quantidade de parâmetros que são estimados. A Tabela 4.3 apresenta as informações sobre as seções que constituem um cromossomo. As Figuras 4.6 e 4.7 ilustram todos os genes que podem compor as seções F_v e T_v, respectivamente.

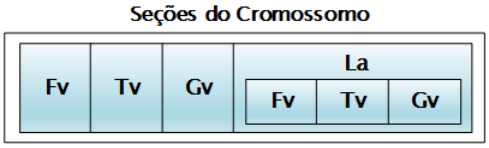

Figura 4.5: Seções do cromossomo.

Tabela 4.3: Seções que compõem o cromossomo do *newGASpeech*.

Símbolo	Descrição
F_v	Seção que armazena os parâmetros responsáveis por modelar a fonte de voz.
T_v	Seção que armazena os parâmetros que modelam o trato vocal.
G_v	Bit que representa a classe do segmento de voz (sonoro ou não sonoro).
L_a	Contém as seções G_v, F_v and T_v dos quadros seguintes (Look-aheads).

Figura 4.6: Estrutura da seção fonte de voz.

O AG desenvolvido tem a opção de executar em modo subdomínio, o qual permite que, apenas, alguns parâmetros sejam estimados. Quando esta opção é utilizada, é necessário definir

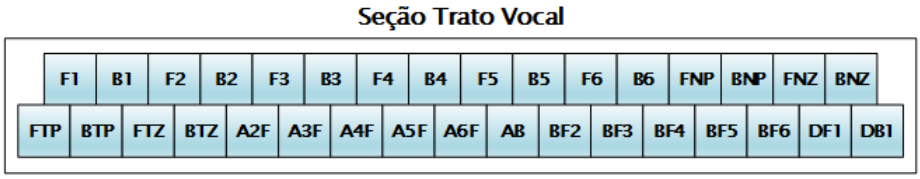

Figura 4.7: Estrutura da seção trato vocal.

quais são estes parâmetros (a serem estimados), pois os demais são copiados a partir de um arquivo extra de entrada do sintetizador, informado durante a configuração do arcabouço. Por exemplo, se um sistema TTS tem como saída um arquivo com os parâmetros do sintetizador, estes podem ser assumidos como os "valores corretos" para a voz sintética, gerada através desse sistema. Os parâmetros, que são copiados, a partir do arquivo com "valores corretos", não são considerados no processo de busca do AG. O número total de genes presentes no cromossomo é dado pela Equação 4.6.

$$N_g = q_f + q_t, \tag{4.6}$$

onde q_f e q_t denotam a quantidade de genes necessários para representar a fonte de voz e o trato vocal, respectivamente. A quantidade total de genes pode variar dependendo da configuração informada no arquivo de entrada citado anteriormente. Por exemplo, se o arcabouço está configurado para executar na opção "subdomínio", estimando apenas os valores dos parâmetros da fonte de voz $F0$, AV, OQ e os parâmetros do trato vocal $F1$, $B1$, $F2$, $F3$, $B3$ e $F4$, a quantidade total de genes é $N_g = 9$ sendo $q_f = 3$ (parâmetros da fonte de voz) e $q_t = 6$ (parâmetros do trato vocal). Caso o *newGASpeech* estime os valores de todos os parâmetros que variam segundo o TTS Dectalk (Seção 4.2), então $N_g = 25$ sendo $q_f = 7$ e $q_t = 18$.

O símbolo ζ denota o processo de codificação de cada parâmetro do sintetizador nos genes do cromossomo e ζ^{-1} o processo inverso, decodificação dos genes nos parâmetros do sintetizador.

No processo de codificação $\zeta(F_v)$ da seção F_v, os genes armazenam os valores dos parâmetros (variáveis) responsáveis por modelar a fonte de voz. A Figura 4.6 representa os possíveis genes da seção F_v. Cada gene representa um parâmetro o qual possui o seu próprio intervalo de valores possíveis $[Min, Max]$. Os intervalos de valores dos parâmetros da fonte de voz estão ilustrados na Tabela 4.4 e foram extraídos de [Klatt and Klatt, 1990]. Portanto, na decodificação de um gene presente em F_v, o mapeamento(μ) pode ser definido como:

$$i = \mu\{\zeta(g), [Min, Max]\} \tag{4.7}$$

onde g representa cada gene de F_v.

Tabela 4.4: Parâmetros da fonte de voz.

Parâmetro	Min	Max
F0	0	5000
AV	0	80
OQ	10	99
TL	0	41
FL	0	100
DI	0	100
AH	0	80
AF	0	80

Para o processo $\zeta^{-1}(T_v)$, a decodificação dos genes que modelam o trato vocal. A codificação realizada em T_v é a mesma descrita anteriormente para F_v. A Tabela 4.5 mostra os genes que compõem o trato vocal e seus respectivos intervalo de valores possíveis de acordo com Klatt [Klatt and Klatt, 1990]. A Figura 4.7 é a representação dos genes que podem compor a seção T_v.

4.4.2.1 Gene de Vozeamento

Na síntese de voz, existem duas classes de sons, a saber, sonoros (vozeado) e não sonoros. O sonoro é caracterizado por um sinal periódico gerado por uma fonte de voz, simulando a vibração das cordas vocais. Em contrapartida, o não sonoro é caracterizado por um sinal randômico gerado através de uma fonte de ruído e simulando, por exemplo, um som nasal [Klatt and Klatt, 1990, Rabiner, 1989]. Sendo assim, um sinal de voz pode ser assumido rudemente como sendo composto por sons sonoros e não sonoros, além de regiões contendo silêncio.

O sintetizador de Klatt possui parâmetros para controlar a geração de som, sendo que dois deles destacam-se por serem correlacionados $F0$ e AV. O primeira representa dez vezes o valor da frequência fundamental e o segundo a amplitude de vozeamento [Klatt and Klatt, 1990]. Quando o $F0$ é zero, o som é não sonoro e, consequentemente, o valor de AV também

Tabela 4.5: Parâmetros do trato vocal.

Parâmetro	Min	Max	Parâmetro	Min	Max
F1	180	1300	FNP	180	500
B1	30	1000	BNP	40	1000
DF1	0	100	FNZ	180	800
DB1	0	400	BNZ	40	1000
F2	550	3000	FTP	300	3000
B2	40	1000	BTP	40	1000
F3	1200	4800	FTZ	300	3000
B3	60	1000	BTZ	40	2000
F4	2400	4990	A2F	0	80
B4	100	1000	A3F	0	80
F5	3000	4990	A4F	0	80
B5	100	1500	A5F	0	80
F6	3000	4990	A6F	0	80
B6	100	4000	AB	0	80

é zero. Se $F0$ é diferente de zero, o som produzido é sonoro e AV também é diferente de zero. Devido a esse comportamento, o intervalo de valores válidos para eles foi modificado no arcabouço desenvolvido (Tabela 4.6) e um gene de vozeamento foi inserido na estrutura do cromossomo. Este gene caracteriza o cromossomo como sonoro ou não podendo este último ser não vozeado ou silêncio. Ele é composto por um bit e seu funcionamento é descrito a seguir:

Tabela 4.6: Novo intervalo de valores possíveis para $F0$ e AV.

Parametro	Min	Max
F0	200	5000
AV	10	80

- Se o gene de vozeamento é igual a zero, o cromossomo é não sonoro e os valores de $F0$ e AV é zero.
- Caso contrário, o gene de vozeamento é igual a um e o cromossomo é sonoro. Neste caso, $F0$ e AV possuem valores diferentes de zero, assumindo qualquer valor compreendido entre os intervalos mencionados na Tabela 4.6.

O gene de vozeamento é inicializado randomicamente de acordo com uma distribuição uniforme e, ao longo da execução do AG, sofre cruzamento e mutação conforme a metodologia descrita a seguir.

- **Cruzamento**: se o gene de vozeamento é 0 então $F0$ e AV possuem valor zero também. Caso contrário, se o valor do gene é 1, o cruzamento é aplicado aos parâmetros $F0$ e AV.
- **Mutação**: se o gene de vozeamento é 0, $F0$ e AV possuem valor zero. Entretanto, quando o gene é 1 e não sofreu mutação, os valores de $F0$ and AV podem sofrer mutação devido ao fato do quadro ser vozeado. Se o valor do gene de vozeamento é zero e após a mutação é modificado para o valor 1, os valores de $F0$ and AV são escolhidos radomicamente com a mesma distribuição na qual foram inicializados.

4.4.3 Mecanismo de Look-ahead

Usualmente, o sinal de voz sintetizado pelo Klatt é composto por vários quadros. Um quadro não pode ser tratado independentemente, pois ele pode, potencialmente, influenciar os próximos quadros. Portanto, uma combinação de valores dos parâmetros de entrada do sintetizador pode ser boa para o quadro corrente, porém impactar negativamente o sinal dos próximos quadros. O mecanismo de Look-ahead [Trindade et al., 2013] permite a avaliação da síntese do quadro corrente em conjunto com os próximos n_f quadros, aumentando, assim, o espaço de busca do problema. A Figura 4.5 ilustra a estrutura do cromossomo considerando a seção de Look-ahead (L_a). Com isso, a L_a armazena informações sobre as seções G_v, F_v e T_v dos quadros seguintes e seu tamanho depende da quantidade de quadros necessários adiante para resolver o problema.

A quantidade mínima de quadros de Look-ahead é baseada no intervalo de tempo t_0 entre os impulsos para gerar a excitação da voz. Sendo assim, o número de amostras do sinal de voz T_0 correspondente à t_0 segundos é $T_0 = t_0 f_s$ no qual f_s é a taxa de amostragem em Hz. Para qualquer sinal periódico, a frequência fundamental $f_0 = 1/t_0$ (em Hz) é um sobre o período fundamental t_0. Para se obter T_0 em função do parâmetro $F0$, é necessário observar que o Klatt utiliza o parâmetro inteiro $F0 = 10f_0$ para representar f_0 [Klatt and Klatt, 1990]. Por isso, o número aproximado da quantidade de amostras que separa os impulsos da voz é dada por $T_0 = f_s/(F0/10)$.

Neste trabalho, $f_s = 11025$ e cada quadro é representado por setenta e uma amostras. A Figura 4.8 ilustra o sinal da fonte de voz para quatro quadros, sendo N o quadro corrente. Neste caso, o quadro N tem $F0$ com valor igual a 943 e, de acordo com a equação T_0, o próximo impulso irá ocorrer 116.9 amostras adiante a partir do início do período deste quadro, ou seja, no quadro $N + 2$. No estudo realizado, a partir dos parâmetros do Klatt, obtidos pelo TTS *Dectalk*, a média do valor de $F0$ foi de 975.5 para os quadros sonoros de voz masculina, ou seja, o valor de $F0$ escolhido para o quadro corrente irá impactar o segundo quadro adiante $(N + 2)$. Para os falantes femininos a média de $F0$ é maior, aproximadamente 1595.5, em relação aos masculinos. Aplicando a equação T_0 para calcular a quantidade de amostras para o próximo impulso, o valor encontrado foi 69.1. Isto indica que para os falantes femininos apenas $N + 1$ quadros de Look-ahead são necessários, pois o valor de $F0$ escolhido para o quadro corrente irá impactar diretamente o próximo quadro.

A avaliação do quadro de Look-ahead é o mesmo realizado para o quadro corrente, ou seja, é feita uma comparação entre os sinais do aquivo de voz alvo e sintético através das métricas EQM, D_{LE} e/ou CC, descritas na Seção 4.4.1. Todos os quadros avaliados (corrente

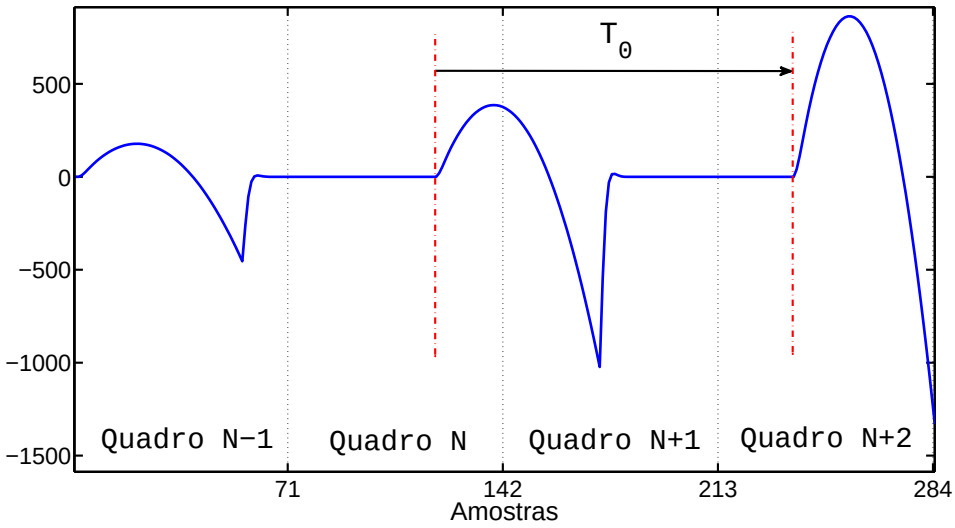

Figura 4.8: O valor escolhido para o parâmetro $F0$ pode influenciar os quadros seguintes.

e Look-aheads) possuem a mesma importância durante a avaliação.

Conforme mencionado anteriormente, são necessários 48 parâmetros para sintetizar um quadro de voz. Cada parâmetro possui seu próprio intervalo de valores válidos. Para a configuração de um simples quadro, a combinação de números possíveis é extensa. Ao estimar o valor dos parâmetros para o quadro atual e os subsequentes (quadros de Look-ahead), aumenta-se bastante o espaço de busca pela melhor solução, dificultando a convergência rápida do AG para os valores ótimos. A dimensão do espaço de busca pela solução do problema é abordada na seção seguinte.

4.4.4 Dimensionalidade do espaço de busca

Cada parâmetro do sintetizador de Klatt possui seu próprio intervalo de valores possíveis [Klatt and Klatt, 1990] sendo estes restritos à números inteiros. Para um simples quadro, a dimensão do espaço de busca S da solução é dada pela Equação 4.8 a seguir.

$$S = \prod_{n=1}^{N_p} (S_n - I_n + 1) \tag{4.8}$$

no qual N_p é o número de parâmetros a ser estimado, S_n e I_n são respectivamente os valores inteiros dos limites superior e inferior do parâmetro n. Por exemplo, se o arcabouço está estimando os valores de 25 parâmetros para um quadro e cada parâmetro assume-se $S_n -$

$I_n + 1 = 50$ valores distintos, $\forall n$, o espaço de busca é dado por $S = (51)^{25} \approx 5 \times 10^{42}$. Esta dimensão é ainda maior quando o espaço de busca inclui os quadros de Look-ahead e pode ser calculada por $S^{(n_f+1)}$. Sendo assim, ferramentas como o *newGASpeech* são importantes para resolver automaticamente este tipo de problema.

4.5 Conclusões sobre o capítulo

O problema apresentado não é trivial e apesar de já existirem trabalhos correlatos como [Jinachitra and Smith III, 2005] [Anumanchipalli et al., 2010] [Bangayan et al., 1997] [Shrivastav and Sapienza, 2006] [Liu and Kewley-Port, 2004] [Laprie and Bonneau, 2007], eles diferem um pouco da proposta ou precisam aperfeiçoar a técnica utlizada. O *Winsnoori* [Laprie and Bonneau, 2007], por exemplo, precisa melhorar a síntese baseada na cópia, pois apesar dele extrair satisfatoriamente os parâmetros do Klatt (versão Jon Iles), a voz sintética gerada apresenta alguma distorção (zumbido). Aliado à isso, este software foi descontinuado pelo autor. O arcabouço desenvolvido estima os valores dos parâmetros de um sintetizador por formantes através do processo de análise-por-síntese, ou seja, à cada iteração do AG é feita a comparação entre os sinais de voz sintético (estimado) e alvo, com o objetivo de gerar uma voz artificial que imite uma voz (sintética ou natural). Para isso, o AG desenvolvido possui acoplado dois sintetizadores por formantes (Klatt e HLSyn), mostrando dessa forma a flexiblidade da metodologia implementada. A escolha por qual sintetizador utilizar fica à cargo do usuário pois esta informação é passada através do arquivo de configuração de entrada. As simulações realizadas assim como os resultados obtidos são apresentados no capítulo seguinte.

Capítulo 5

Experimentos e Resultados

Neste capítulo são apresentados os experimentos realizados de imitação da voz através do arcabouço desenvolvido (*newGASpeech*) e do software Winsnoori (baseline). Foram utilizadas como alvo tanto vozes sintéticas quanto naturais (masculinas e femininas), além de dois sintetizadores por formantes (Klatt e HLSyn). Os resultados obtidos foram avaliados através de uma metodologia que abrange métricas objetivas e avaliação subjetiva.

5.1 Introdução

Os experimentos realizados foram divididos em dois grupos: vozes alvo sintéticas e naturais, sendo todas elas no idioma inglês (americano). As vozes sintéticas adotadas abrangeram seis falantes (três masculinos e três femininos) e foram geradas através do TTS DECtalk. As palavras utilizadas por falante estão listadas na Tabela 5.1. Além disso, uma análise da dimensão do espaço de busca versus o valor do erro, obtido entre os sinais alvo e sintético, foi realizada para os experimentos com vozes sintéticas (ver seção 5.3). Para esses mesmos experimentos, uma outra análise foi feita em relação aos erros percentual e absoluto dos valores estimados dos parâmetros (ver seção 5.5.1). Já os arquivos de voz natural usados compreenderam a pronúncia dos dígitos de 0 a 9 para três falantes masculinos e foram obtidos através do corpus do TIDIGITS [TIDIGITS, 2014]. A principal motivação em usar vozes sintéticas é ter controle sobre os experimentos, uma vez que considera-se correto os valores dos parâmetros de entrada do sintetizador obtidos através do TTS DECtalk. As vozes naturais foram utilizadas para testar o arcabouço desenvolvido com falantes desconhecidos.

Devido à dificuldade em avaliar objetivamente as vozes sintéticas obtidas através do

Tabela 5.1: Lista de palavras para falantes masculinos e femininos (vozes sintéticas).

Índice	Frank	Harry	Paul	Betty	Ursula	Wendy
1	air	awe	are	air	bean	death
2	dill	earl	end	dill	earl	fern
3	hurl	gill	is	hurl	tang	is
4	jam	them	no	jam	them	then
5	who	wish	there	who	wish	there

newGASpeech, foi definida uma metodologia de avaliação que abrangeu cinco figuras de mérito (métricas objetivas), a saber, Erro Quadrático Médio, log da Distância Espectral, Relação Sinal-Ruído, PESQ e P.563, sendo que as duas primeiras métricas foram utilizadas também como funções objetivo no arcabouço.Neste trabalho, os resultados obtidos foram comparados objetivamente e subjetivamente com aqueles produzidos pelo Winsnoori (baseline). O detalhamento sobre a metodologia de avaliação adotada é apresentado na seção 5.2.

Para os experimentos, o newGASpeech foi configurado com número de gerações, quantidade de indivíduos, taxas iniciais de cruzamento e mutação conforme especificado na Tabela 5.2. As taxas de cruzamento e mutação configuradas foram adaptativas (seção 4.4.1) e poderiam ser decrementadas em 0.01 à cada iteração do AG. O decremento só ocorre se a população apresentar diversidade e até que a taxa mínima seja igual a 0.01. O arcabouço foi configurado para executar no modo Interframe e 10% dos melhores cromossomos do quadro anterior foram copiados para inicializar parte da população do próximo quadro. Para as simulações com o Klatt, optou-se pela configuração de subdomínio estimando apenas os vinte e cinco parâmetros, que variam conforme estudo realizado nos arquivos do Klatt, gerados através do TTS DECktalk (ver seção 4.2). Já nas simulações com o sintetizador HLSyn, a opção de subdomínio não foi utilizada e todos os treze parâmetros foram estimados. As simulações através do arcabouço desenvolvido foram executadas em um *cluster* com 8 nós sendo um nó principal [1](*master*) e os demais secundários [2](*slaves*).

[1]Processador Xeon E5450, 1 TiB de HD, 02 portas Gigabit e 8 GiB de Memória RAM

[2]Processador Xeon E5450, 80 GiB de HD SAS, 02 portas Gigabit e 4 GiB de Memória RAM

Tabela 5.2: Configuração do *newGASpeech*.

Parâmetro	Valor
Número de gerações	800
Tamanho da população (indivíduos)	5000
Taxa inicial de cruzamento	50%
Taxa inicial de mutação	30%

5.2 Metodologia para avaliação dos resultados

Na literatura, não consta metodologia padronizada para avaliar a qualidade da voz sintética, apenas algumas referências individuais de realizar avaliações objetivas e subjetivas [Kondo, 2012]. Sendo assim, as Seções 5.2.1 e 5.2.2 apresentam as métricas objetivas adotadas neste trabalho e a avaliação subjetiva realizada.

5.2.1 Métricas para avaliação objetiva

Conforme mencionado anteriormente, os arquivos gerados pelo TTS DECtalk (vozes sintéticas) e aqueles obtidos através do corpus do TIDIGITS (vozes naturais) foram utilizados como entrada (voz alvo) pelo *newGASpeech* e pelo Winsnoori. Para o cálculo das métricas objetivas, os sinais alvo e sintético foram alinhados levando em consideração a correlação cruzada (Equação 4.3). Para cada voz sintética produzida, as seguintes métricas objetivas foram medidas: RSR (Relação Sinal-Ruído), EQM, D_{LE}, PESQ (*Perceptual Evaluation of Speech Quality*) [Pes, 2001] e P.563 [P56, 2004].

A métrica RSR foi calculada em decibels (dB) entre as pontências do sinal alvo ($P_{sinal\,alvo}$) e do ruído ($P_{\text{ruído}}$), conforme a Equação 5.1. O valor da RSR deve ser o maior possível, indicando assim que a potência do sinal é maior que a do ruído.

$$RSR = 10 \times \log 10 \frac{P_{sinal\,alvo}}{P_{\text{ruído}}} \tag{5.1}$$

onde a $P_{\text{ruído}}$ é obtida através do EQM (Equação 4.2) entre os sinais alvo e o sintético.

As Equações 4.2 e 4.1 foram utilizadas para o cálculo do EQM e da D_{LE}, respectivamente, além de serem utilizadas como funções objetivo no arcabouço. Para essas duas métricas, quanto menor o valor calculado é melhor. O PESQ e o P.563 são métricas utilizadas

para avaliar a degradação na qualidade da voz em telefonia. Elas atribuem uma pontuação à voz avaliada que varia de 1 a 5, sendo 1 "má qualidade" e 5 "excelente qualidade". A Figura 5.1 ilustra o funcionamento do PESQ, em que os arquivos de voz original e degradada são dados como entrada. Ambos são submetidos a modelos perceptuais que geram uma representação interna dos sinais (original e degradado). A diferença entre essas representações internas determina a diferença audível entre eles e gera uma pontuação para a voz degradada através de cálculos, como, por exemplo, a média não-linear ao longo do tempo e frequência de perturbação simétrica e assimétrica, realizados por um modelo cognitivo.

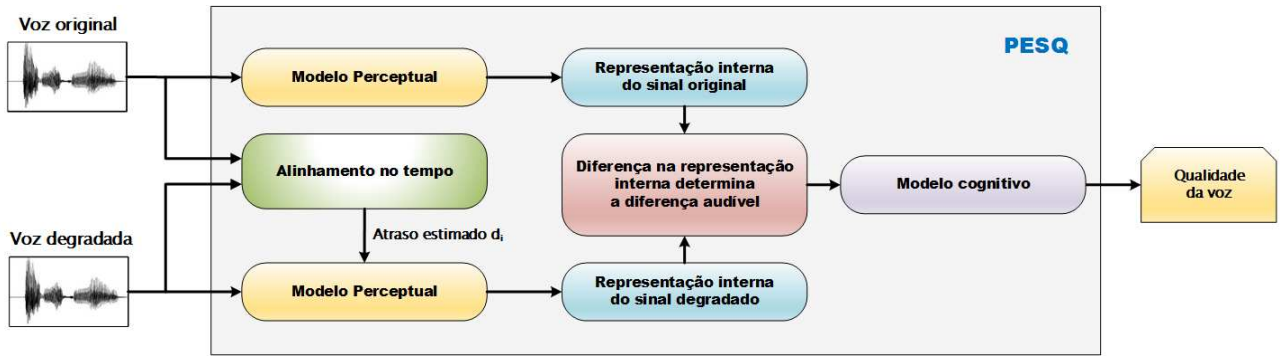

Figura 5.1: Funcionamento do PESQ adaptado de [Pes, 2001].

Uma importante diferença entre essas métricas está no fato de que o P.563 não precisa de um arquivo alvo para atribuir uma pontuação para o arquivo avaliado. O P.563 é composto por três módulos (ver Figura 5.2), tais como, pré-processamento, estimação da distorção e mapeamento perceptual. Na etapa de pré-processamento, o sinal de voz dado, como entrada, é normalizado, filtrado e submetido à detecção de atividade de voz (DAV). Na estimação da distorção, um modelo de trato vocal é construído com o objetivo de identificar variações anormais na voz. Por fim, na etapa do mapeamento perceptual, um modelo de referência psico-acústico é utilizado para estimar a degradação do sinal de voz, sendo que, para isso, ele utiliza como referência um sinal reconstruído a partir da voz dada como entrada.

O P.563 é interessante para o problema de imitação da voz aqui abordado, pois, as demais métricas, requerem um razoável alinhamento dos sinais (alvo e sintético) no tempo, enquanto esta métrica não. Entretanto, a desvantagem dela é não ser apropriada para sinais de voz com duração menor que três segundos [P56, 2004] e, devido a esse fator, ela não foi utilizada para avaliar os sinais sintéticos produzidos a partir dos arquivos do TIDIGITS dados como entrada (vozes alvo) no *newGASpeech* e no Winsnoori.

Deve-se notar que nenhuma destas métricas está perfeitamente correlacionada com avaliações subjetivas. Testes de audição informais indicam que existe uma correspondência

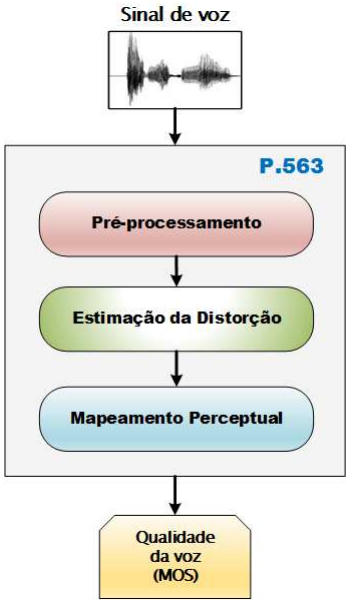

Figura 5.2: Funcionamento simplificado do P.563 adaptado de [Malfait et al., 2006].

entre o resultado global das métricas objetivas e uma avaliação subjetiva, como o MOS (*Mean Opinion Score*), por exemplo. Sendo assim, após o cálculo das métricas objetivas, foi realizada uma avaliação subjetiva nos resultados obtidos através dos experimentos, apenas, com voz natural, pois os falantes eram desconhecidos. A metodologia utilizada nesta avaliação está descrita na seção seguinte.

5.2.2 Avaliação subjetiva

A avaliação subjetiva foi realizada através do método CCR (*Comparison Category Rating*), no qual são atribuídas pontuações [P80, 1996] para as vozes que se deseja avaliar a qualidade em relação à voz alvo. A pontuação, que pode ser atribuída, está listada na Tabela 5.3. De acordo com [P80, 1996], a quantidade de ouvintes para esse tipo de avaliação deve ser no mínimo seis.

O ideal é que sejam doze ou mais ouvintes e que as sessões de avaliação não demorem mais do que seis minutos, para não cansar os ouvintes. Neste trabalho, vinte e sete ouvintes, não especialistas na língua inglesa, usaram o software LisTEn (*LIStening Test ENvironment*) [Schafer et al., 2011] para avaliar subjetivamente noventa arquivos de vozes sintéticas que foram geradas. Para avaliar cada voz sintética, o ouvinte escutou, através do LisTEn, primeiramente, a voz alvo e, na sequência, a voz avaliada. Com esse mesmo software, ele atribuiu uma nota, variando de -3 a 3, para as vozes sintéticas e cada uma delas foi avaliada por três ouvintes

diferentes.

Tabela 5.3: Pontuação utilizada em testes subjetivos através do método CCR.

Pontuação	Qualidade
3	Bem melhor
2	Melhor
1	Ligeiramente melhor
0	Aproximadamente o mesmo
-1	Ligeiramente pior
-2	Pior
-3	Muito pior

5.3 Dimensionalidade do espaço de busca

O EQM foi uma métrica importante para guiar o arcabouço, assim como para avaliar os resultados obtidos. Sendo assim, foram realizados experimentos com o objetivo de verificar como a dimensão do espaço de busca influencia no valor da função fitness quando o EQM é utilizado. Para isso, foi desenvolvido no arcabouço a opção de informar, por quadro, um intervalo restrito de valores (próximo ao valor correto) para cada parâmetro. Os experimentos abrangeram as cinco palavras do falante Paul (ver Tabela 5.1) e definiu-se uma nova configuração, mais compacta, para o newGASpeech, descrita a seguir. A partir dos valores corretos dos parâmetros para cada quadro, eles foram variados em ± 2%, 4%, 8%, 16% e 32%.

1. Tamanho da população: 1000 indivíduos;
2. Quantidade de gerações: 300;
3. Taxa inicial de cruzamento: 50%;
4. Taxa inicial de mutação: 30%.

Considerando que os experimentos abrangeram no máximo a estimação de vinte e cinco parâmetros, a maior dimensão do espaço de busca ocorreu quando os valores dos parâmetros foram variados em pm 32%. A variável Z representa essa dimensão máxima e foi calculada

através da Equação 4.8. O valor de Z é aproximadamente $6.86e{+}102$, incluindo as dimensões de dois quadros de Look-ahead necessários para os falantes masculinos, conforme explanado na Seção 4.4.3. As demais dimensões do espaço de busca, para a variação dos valores dos parâmetros em pm 2, 4, 8 e 16, foram normalizadas em Z. A Figura 5.3 ilustra a média do EQM calculado para todas as palavras do falante especificado. Pode-se observar que o espaço de busca é grande e o EQM só diminui significativamente quando este espaço é reduzido em várias ordens de magnitude.

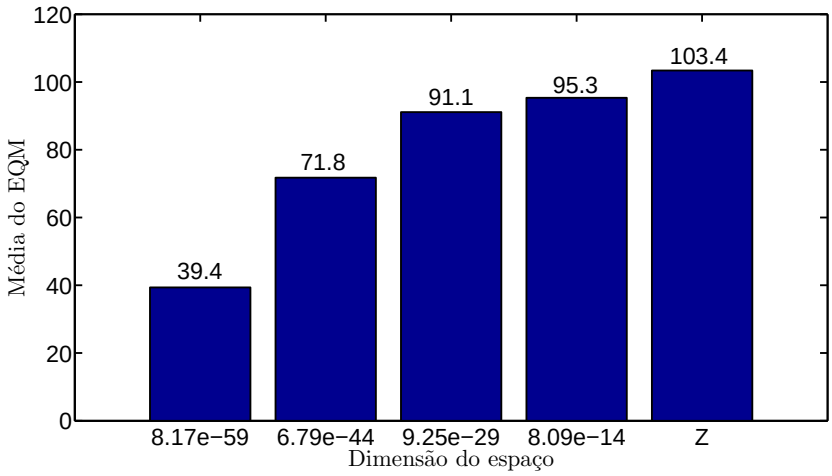

Figura 5.3: Média do EQM para todas as 5 palavras.

5.4 Avaliação objetiva dos experimentos mono-objetivo e multiobjetivo com vozes sintéticas

O arcabouço desenvolvido foi baseado no algoritmo NSGA-II sendo portanto um AGMO (Seções 3.5 e 4.4.1). Três funções objetivo foram implementadas: o EQM (Equação 4.2), a D_{LE} (Equação 4.1) e a CC (Equação 4.3), porém o log da Distância Espectral (D_{LE}) não foi utilizado nesses experimentos pois o mesmo leva em consideração apenas a magnitude do sinal, desprezando a fase em que ele se encontra.

Para avaliar a eficácia do arcabouço em operar multiobjetivo e o uso das probabilidades de cruzamento e mutação com valores adaptativos e fixos, foram realizadas 4 simulações, conforme especificado a seguir:

1. 1obj_EQM_Adap: Mono-objetivo com EQM como função objetivo e os valores adap-

tativos das probabilidades de cruzamento e mutação.

2. 2*obj_EQM_CC_Adap*: Multiobjetivo com duas funções objetivo: o EQM e a CC. Os valores das probabilidades de cruzamento e mutação foram adaptativos.

3. 1*obj_EQM_Fixo*: Mono-objetivo com EQM como função objetivo e os valores fixos das probabilidades de cruzamento e mutação.

4. 2*obj_EQM_CC_Fixo*: Multiobjetivo com duas funções objetivo: o EQM e a CC. Os valores das probabilidades de cruzamento e mutação foram fixos.

A configuração do arcabouço para esses experimentos está especificado na Tabela 5.4 a seguir. Foram estimados os 25 parâmetros que variam para o TTS *DECtalk* considerando a síntese da palavra "are" pronunciada pelo falante Paul.

Tabela 5.4: Configuração do *newGASpeech* para comparação entre experimentos mono-objetivo e multiobjetivo.

Parâmetro	Valor
Número de gerações	80
Tamanho da população (indivíduos)	500
Taxa inicial de cruzamento	10%
Taxa inicial de mutação	10%

As vozes sintéticas produzidas foram avaliadas segundo a metodologia de avaliação objetiva definida na Seção 5.2.1. Os valores obtidos são apresentados na Tabela abaixo.

Tabela 5.5: Avaliação objetiva para as simulações mono-objetivo e multiobjetivo.

Simulação	Avaliação objetiva				
	EQM	D_{LE}	CC	RSR	PESQ
1obj_EQM_Adap	1353.9	5.3	0.9	4.7	2.4
2obj_EQM_CC_Adap	1581.3	6.7	0.9	3.5	2.8
1obj_EQM_Fixo	815.5	3.7	0.9	6.9	3.5
2obj_EQM_CC_Fixo	789.7	3.8	0.9	6.4	3.4

Conforme observado, os valores obtidos nas métricas são bem próximos quando se utiliza o *newGASpeech* mono-objetivo ou multiobjetivo com os valores das probabilidades adaptativos, estes decrementados conforme abordado na Seção 4.4.1 (Equações 4.4 e 4.5). Portanto, optou-se neste trabalho utilizá-lo apenas mono-objetivo sendo a função objetivo o EQM, como no trabalho recente [Neto et al., 2013] em que essa mesma métrica foi usada como função objetivo para avaliar sinais de voz sintéticos.

Os experimentos com das probabilidades fixas de cruzamento e mutação foi uma sugestão da banca avaliadora e conforme pode ser observado, as métricas EQM e D_{LE} foram menores e a RSR e o PESQ foram maiores, indicando assim que as probabilidades com valores fixos podem gerar resultados melhores do que com valores adaptativos.

5.5 Experimentos com vozes sintéticas

Os experimentos com vozes sintéticas foram referentes as palavras e falantes da Tabela 5.1. As Figuras 5.4 e 5.5 mostram os resultados usando gráficos "boxplot" para falantes masculinos e femininos. Nos testes realizados, os resultados obtidos pelo *newGASpeech* foram melhores do que aqueles através do WinSnoori, para todos os falantes, de acordo com a metodologia de avaliação objetiva adotada (Seção 5.2.1). Os resultados ao se utilizar o arcabouço foram similares (Tabela 5.6) tanto para falantes masculinos quanto femininos. O D_{LE} variou muito pouco tanto para as vozes sintéticas geradas pelo arcabouço quanto pelo baseline. Porém, considerando a média de valores entre falantes femininos e masculinos, o EQM foi 87.3% menor para o *newGASpeech* e a RSR foi maior do que 11dB indicando dessa maneira que a potência dos sinais sintéticos foi maior do que o ruído. Já o PESQ e o P.563 foram 55.9% e 65.9% maiores para o AG considerando os falantes masculinos, 37.9% e 77.8% maiores também para os falantes femininos. A diferença maior nos resultados pode ser notada para a falante *Wendy* a qual obteve menor média de EQM (91.6%), maiores médias de PESQ (28.4%) e P.563 (78.3%) apesar da D_{LE} não ter sido o menor valor (91.2% menor). Essas porcentagens refletem a variação das métricas em relação ao baseline (Winsnoori). Considerando que um AG foi executado para cada quadro de voz, a média de tempo execução para cada quadro foi de 12 minutos.

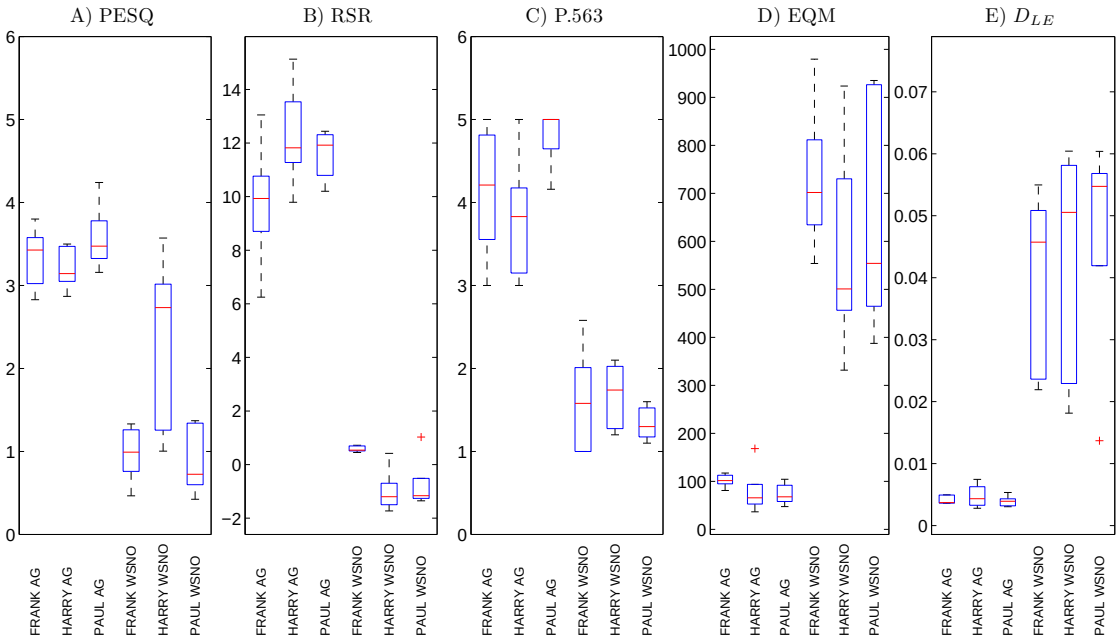

Figura 5.4: Valores do A) PESQ, B) RSR C) P.563 D) e EQM E) para os falantes masculinos.

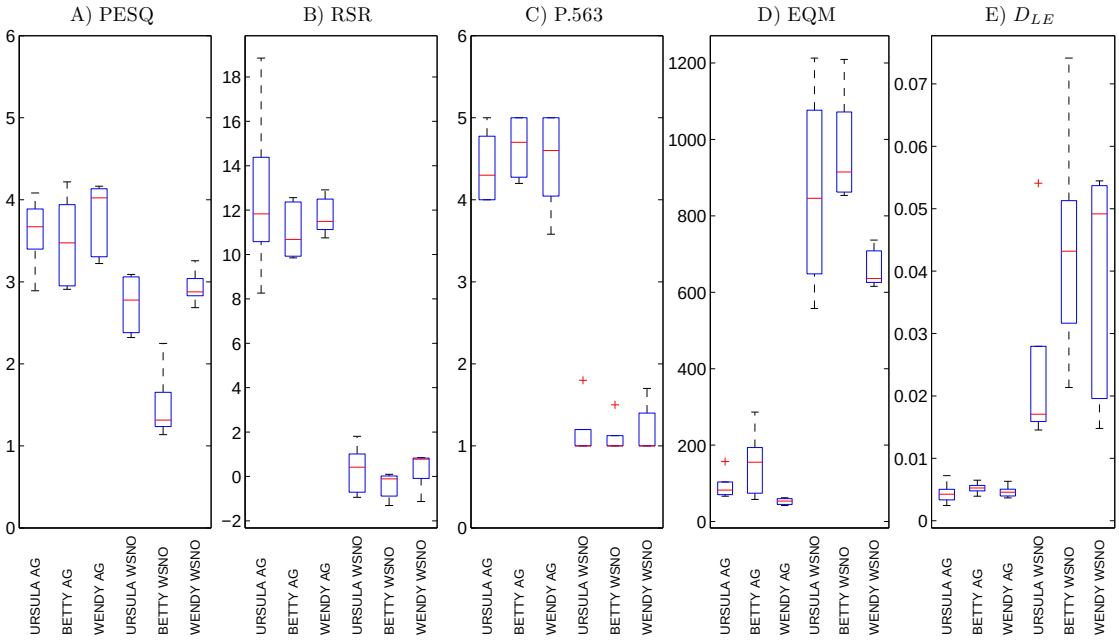

Figura 5.5: Valores do A) PESQ, B) RSR C) P.563 D) e EQM E) para os falantes femininos.

Tabela 5.6: Média das métricas para falantes masculinos e femininos.

Sexo	Masculino		Feminino	
Média das métricas	AG	Wsno	AG	Wsno
EQM	78.3	585.6	97	799
D_{LE}	0.004	0.05	0.004	0.036
PESQ	3.4	1.5	3.7	2.3
P.563	4.4	1.5	4.5	1
RSR	11.2	-0.6	11.3	0.4

5.5.1 Erros percentual e absoluto dos parâmetros estimados pelo *newGASpeech* a partir de vozes sintéticas

Dado que os valores corretos dos parâmetros são conhecidos para as vozes alvo sintéticas devido ao fato delas terem sido geradas através do TTS DECtalk, para todos os experimentos com esse tipo de voz, usando o arcabouço desenvolvido, foram calculadas as médias do erro percentual (Equação 5.2) e do erro absoluto (Equação 5.3). O objetivo com essas métricas é observar quais parâmetros foram melhores estimados e os piores também.

$$EP = \frac{1}{n}\left|\frac{\hat{p_k} - p_k}{p_k}\right| \times 100\% \tag{5.2}$$

$$EA = \frac{1}{n}|\hat{p_k} - p_k|\% \tag{5.3}$$

Nessas equações, k é a quantidade de parâmetros estimada, sendo, neste trabalho, o valor máximo é $k = 25$ parâmetros, que variam segundo o TTS DECtalk. O $\hat{p_k}$ é o valor do parâmetro (k) do Klatt, estimado pelo AG, e p_k o valor correto do parâmetro. O cálculo dos erros foi feito para todas as palavras e falantes listados na Tabela 5.1. Para cada palavra, a média do EP e do EA foram calculados, apenas, para os quadros sonoros, evitando, assim, os quadros não sonoros (silêncio ou não vozeados). Especialmente no EP, os frames sonoros, em que o valor correto do parâmetro (p_k) é zero, não foram desconsiderados no cálculo. As Tabelas 5.7 e 5.8 apresentam os valores obtidos dos erros. Para os parâmetros que compõem a fonte de voz, o parâmetro DI alcançou $EP = 0$ e o $F0$ obteve $EP = 1$ para os falantes femininos. Já para os falantes masculinos, esses mesmos parâmetros apresentaram EP maior ($DI = 7$ e $F0 = 28$). O alto EP do $F0$ para os falantes masculinos justifica-se pelo fato

do valor escolhido para o quadro atual influenciar, apenas, no segundo quadro consecutivo, conforme abordado na seção 4.4.3. Esta situação não ocorre com os falantes femininos, pois o valor do $F0$ escolhido para o quadro atual, influencia diretamente no quadro seguinte. Os demais parâmetros da fonte de voz modelam o impulso glotal produzido pelo $F0$, portanto, é normal que os valores de EP para estes parâmetros também sejam maiores para os falantes masculinos que femininos.

Nos parâmetros que fazem parte dos ressonadores em cascata, os mesmos parâmetros obtiveram $EP = 0$ tanto para os falantes masculinos quanto femininos. Isso deve-se ao fato dos parâmetros $A2F$ a $A4F$ possuirem valor igual a zero na maior parte dos frames, conforme observa-se, por exemplo, nos histogramas apresentados no Apêndice D.

Tabela 5.7: Erro percentual dos parâmetros estimados pelo *newGASpeech*.

Ramo	Parâmetros		
	EP	Masculino	Feminino
Fonte de Voz	$EP = 0$	-	DI
	$EP < 10\%$	AV, OQ e DI	F0, AV e OQ
	$10\% < EP < 50\%$	F0, TL e AH	AH and TL
Ressonadores em Cascata	$EP = 0$	A2F, A3F, A4F, A5F e AB	A2F, A3F, A4F, A5F e AB
	$EP < 10\%$	F1, F2 e F3, F4, B4 e B5	F2, F4, B4, B5 e FNP
	$10\% < EP < 50\%$	B3, FNP, BNP e FNZ	F1, F3, B3 e BNP
	$EP > 50\%$	AF, B1, B2 e BNZ	AF, B1, B2, FNZ e BNZ

As Figuras 5.6 e 5.7 ilustram os valores dos erros por parâmetro. O destaque nas figuras, tanto para os falantes masculinos quanto femininos, é para os parâmetros AF, $B1$, $B2$ e BNZ os quais obtiveram valor de erro alto se comparado aos demais parâmetros. Para explicar esse comportamento, esses parâmetros foram submetidos à uma análise de sensbilidade, a qual é descrita na seção seguinte.

Tabela 5.8: Erro absoluto dos parâmetros estimados pelo *newGASpeech*.

Ramo	Parâmetros		
	EP	Masculino	Feminino
Fonte de Voz	$EA \leqslant 10$	AV, OQ, TL e DI	AV, OQ, TL e DI
	$10 \leqslant EA \leqslant 50$	AH	F0 e AH
	$EA \geqslant 200$	F0	-
Ressonadores em Cascata	$EA \leqslant 50$	AF, F1, B4, BNP, A2F, A3F, A4F, A5F e AB	AF, B4, BNP, A2F, A3F, A4F, A5F e AB
	$50 \leqslant EA \leqslant 100$	B1, B3 e B5	F1, B3, B5 e FNP
	$100 \leqslant EA \leqslant 200$	F2, B2 e FNP	F2 e B2
	$EA \geqslant 200$	F3, F4, FNZ e BNZ	F3, F4, FNZ e BNZ

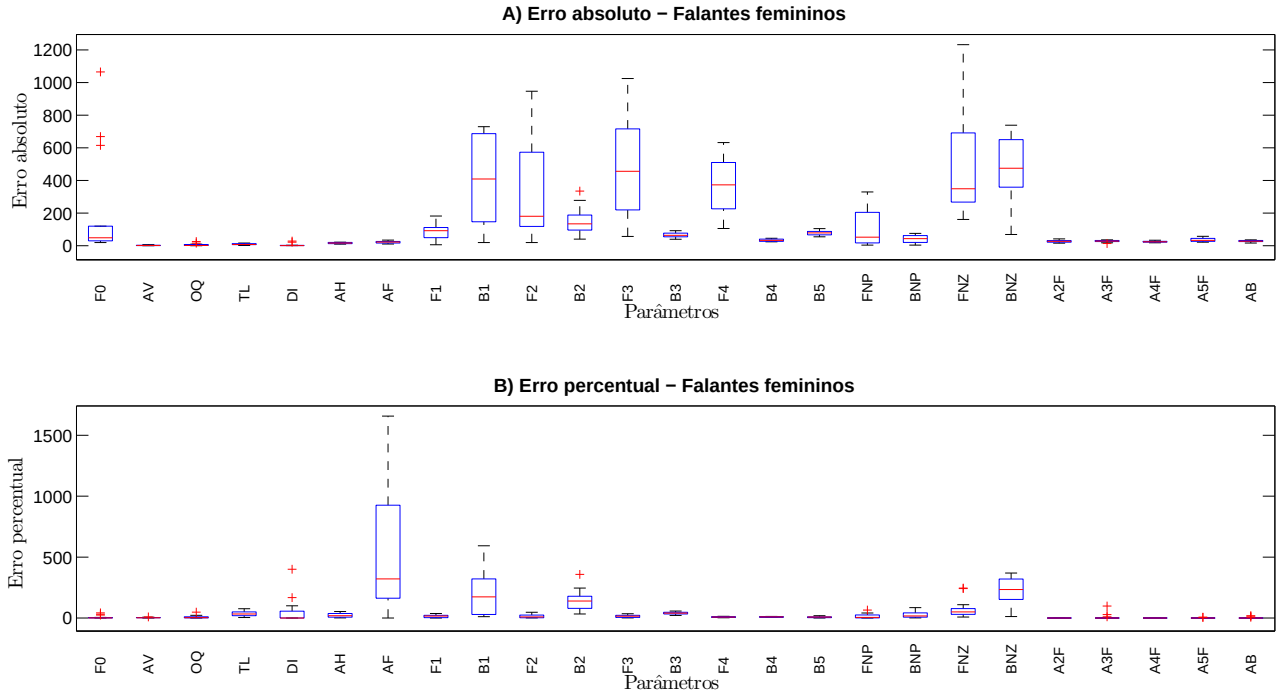

Figura 5.6: A) Erro absoluto B) Erro percentual para falantes femininos.

5.5.1.1 Sensibilidade dos parâmetros com alto erro percentual

Para realizar a análise de sensibilidade foi escolhida a palavra "are" do falante Paul (Tabela 5.1) por possuir apenas 130 quadros de voz, uma das menores quantidade de quadros em relação as demais palavras. Os parâmetros AF, $B1$, $B2$ e BNZ foram variados indivi-

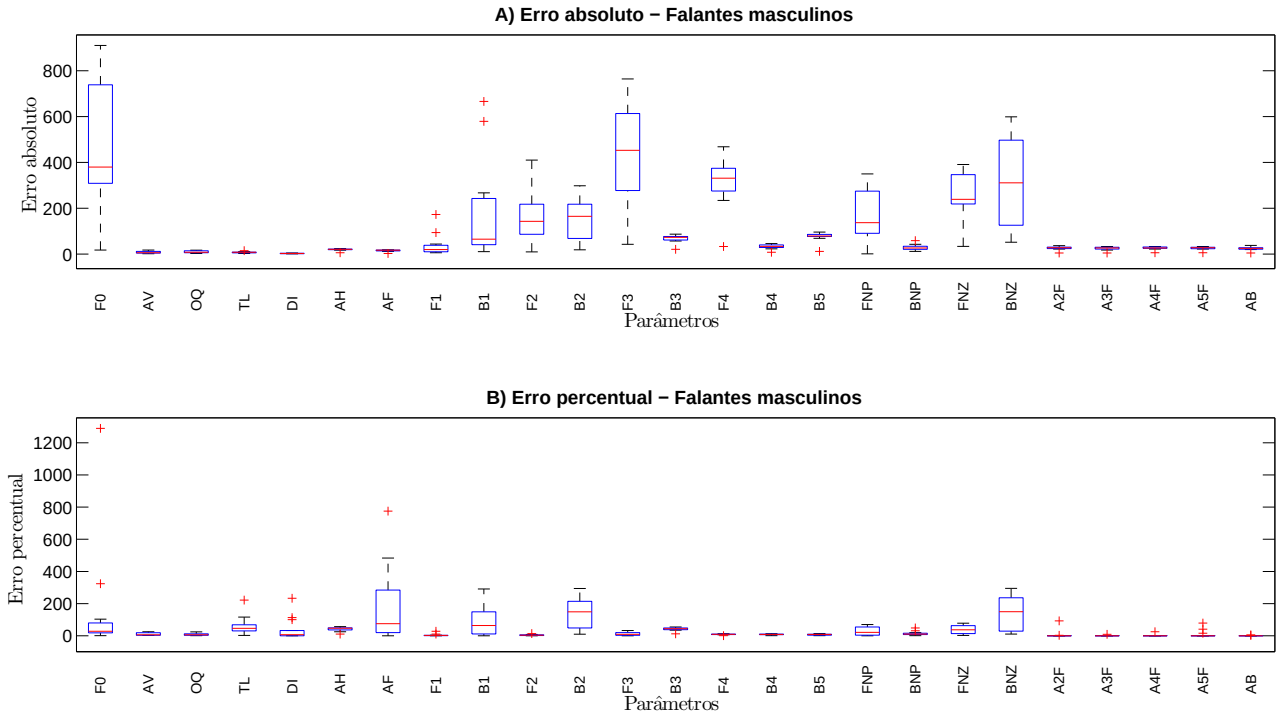

Figura 5.7: A) Erro absoluto B) Erro percentual para falantes masculinos.

dualmente em 10, 20 e 50 por cento em relação ao valor correto. Apenas os quadros sonoros ($F0$ e AV diferentes de zero), no intervalo de 10 em 10, sofreram esta variação. Esses novos arquivos do Klatt foram sintetizados e calculou-se o EQM em relação ao arquivo de voz alvo (Tabela 5.9). Nota-se que apesar de existir uma variação no valor do erro quando se aumenta ou diminui o valor desses parâmetros, eles não influenciam fortemente o sinal de voz gerado se comparado com o parâmetro $F0$, por exemplo. O parâmetro AF permanece com $EQM = 0$, independente da variação, pois pode ser observado em seu histograma (Apêndice D), na maior parte dos quadros ele apresenta valor zero também.

A sensibilidade reduzida desses parâmetros em afetar o EQM, quando eles são variados, pode ser identificada também através da convergência lenta deles para o valor correto. Por exemplo, as Figuras E.1 à E.4 (Apêndice D) ilustram a convergência de AF, $B1$, $B2$ e BNZ no primeiro quadro de voz sonoro (quadro 3) da palavra “are” para o falante Paul. Neste caso, considera-se simulações para estimar os 25 parâmetros utilizando uma população de 500 indivíduos e 80 gerações. Os valores corretos para cada um dos parâmetros no quadro 3 estão listados na Tabela . Nota-se nas figuras mencionadas que o parâmetro AF é o menos sensível pois com o passar das gerações não concentrou sua população próximo ao valor

Tabela 5.9: EQM do sinal de voz para a variação dos parâmetros.

Parâmetro	Porcentagem de variação					
	+10%	+20%	+50%	−10%	−20%	−50%
F0	796.9	653.5	761.8	685.5	752.1	563
AF	0	0	0.07	0	0	0
B1	23.6	45.4	97.6	27.6	58.5	183.5
B2	8.3	16.1	35.3	9.5	19.5	55.8
BNZ	22.1	43.7	105.9	22.5	45.3	115.5

correto do parâmetro. Nos demais parâmetros, ainda houve uma mudança na escala onde se concentraram-se os indivíduos porém ainda estavam longe de atingir o valor correto.

Tabela 5.10: Valor dos parâmetros no quadro 3.

Parâmetro	Valor
AF	10
B1	200
B2	90
BNZ	200

A variação do parâmetro $F0$ para mais ou para menos acarreta a diminuição do EQM conforme pode ser observado na Tabela 5.9. A Figura 5.8 ilustra o sinal da fonte de voz e o EQM para os quadros sonoros de 3 à 10 quando se varia o $F0$ para mais 10%, 20% e 50%. Pode se notar que apesar dos impulsos estarem desalinhados, isso não determinante para impactar no valor do EQM.

5.6 Experimentos com vozes naturais

Nesses experimentos, as vozes naturais do corpus do TIDIGITS [TIDIGITS, 2014] utilizadas totalizaram 30 sinais de voz (dígitos de 0 à 9 para 3 falantes masculinos) e foram usadas como voz alvo tanto para o *newGASpeech*, utilizando os sintetizadores Klatt e o HLSyn [estimation of Klatt parameters, 2015], como para o Winsnoori. Nesses experimentos não

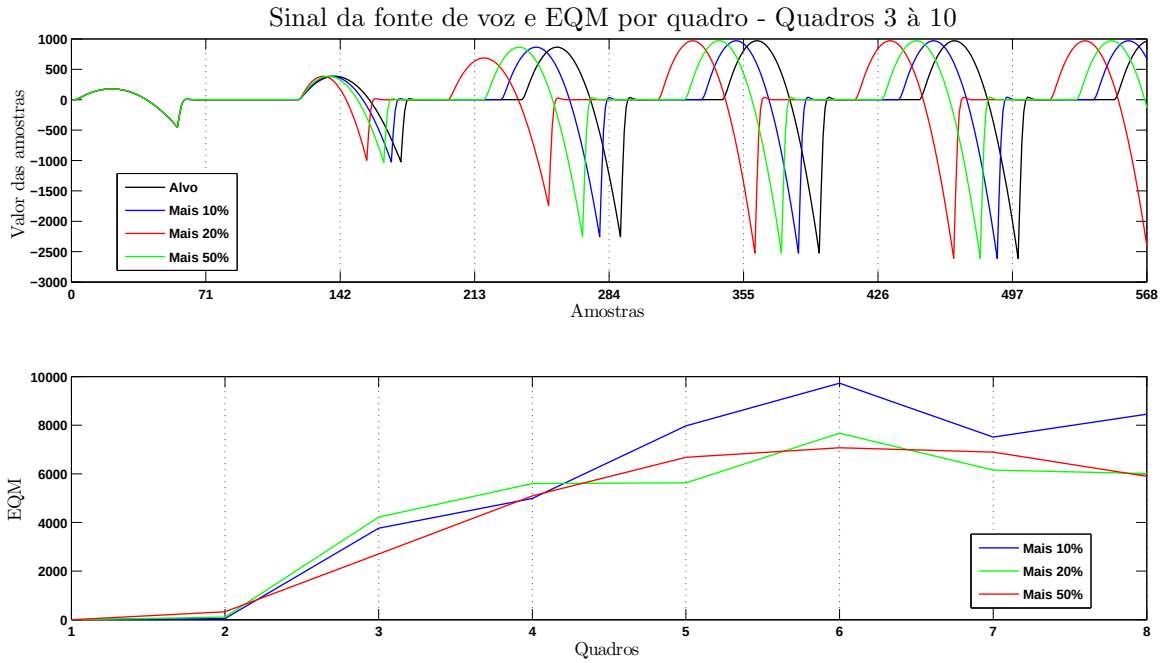

Figura 5.8: Sinal da fonte de voz e do EQM para os quadros de 3 à 10.

foram utilizadas vozes femininas devido à qualidade das vozes sintéticas produzidas pelo AG serem muito semelhantes independente do sexo do falante. As vozes sintéticas produzidas por ambos foram avaliadas através da metodologia abordada na Seção 5.2, incluindo avaliações objetivas e subjetiva.

Nesses experimentos, conforme pode ser observado na Figura 5.9, o *GA-Klatt* gerou vozes sintéticas com médias altas de PESQ (2.8) e RSR (5.5), e baixas médias de EQM (59.0) e D_{LE} (2.3) em relação ao *GA-HLSyn* e Winsnoori. Este resultado é ratificado pela avaliação subjetiva na qual o *GA-Klatt* obteve pontuação próxima à zero indicando assim que as vozes sintéticas geradas foram similares as vozes alvo. Os piores resultados obtidos, de acordo com a avaliação subjetiva, foram os do *GA-HLsyn* com pontuação média de -1.6, indicando assim que a qualidade das vozes sintéticas variaram entre "ligeiramente pior" e "pior", apesar de ter obtido um valor de PESQ 32.2% maior e EQM 4.9% menor do que o software utilizado como baseline. O tempo médio de execução do *GA-Klatt* para cada quadro de voz foi de aproximadamente 44 minutos, ou seja, esse tempo foi 72% maior do que aquele gasto para as simulações com vozes alvo sintéticas. O Winsnoori apresentou os piores valores de PESQ, EQM e D_{LE}. Na avaliação subjetiva, as vozes geradas por esse mesmo software obtiveram qualidade "ligeiramente pior" em relação aos arquivos de voz alvo.

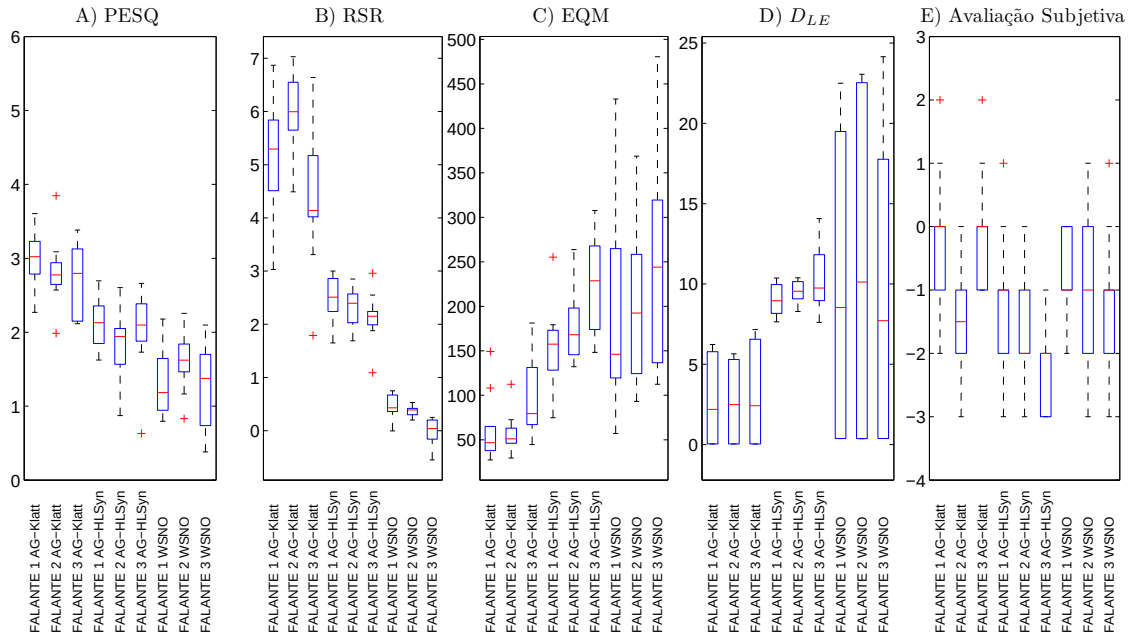

Figura 5.9: A) PESQ, B) RSR C) EQM D) D_{LE} E) Teste Subjetivo para os falantes masculinos.

5.7 Conclusões sobre o capítulo

Neste capítulo foram apresentados os resultados obtidos com o arcabouço baseado em AG para realizar a imitação da voz utilizando dois sintetizadores: Klatt e HLSyn. A qualidade das vozes sintéticas produzidas foram comparadas através de uma metodologia de avaliação com àquelas obtidas pelo software Winsnoori (baseline).

Nos experimentos com voz alvo sintética (Tabela 5.6), o AG obteve melhor desempenho em relação ao Winsnoori se comparado os valores obtidos pela avaliação objetiva (métricas EQM, D_{LE}, PESQ e P.563). Em média, o *newGASpeech* obteve EQM e D_{LE} 12% e 9.5% menores do que o baseline. Já o PESQ e o P.563 foram 53% e 28% maiores, respectivamente. A maior diferença foi na RSR em que o Winsnoori obteve um valor próximo à 1 para os falantes femininos e valor negativo para os masculinos, indicando neste último que nos arquivos gerados pelo baseline existe mais ruído do que sinal sonoro. Em relação ao erro percentual, a maior parte dos parâmetros obteve valor menor do que 50%, com apenas praticamente um terço dos parâmetros com valor acima de 50%. Esses parâmetros com alto erro percentual não são muito sensíveis quando ocorre variação em seus valores, apresentando portanto pouca variação no EQM.

Para os experimentos com voz alvo natural, o *AG-Klatt* superou significativamente o Winsnoori em relação à avaliação objetiva, com EQM e D_{LE} $69,6\%$ e $73,1\%$ menores, respectivamente. Os valores obtidos com o PESQ em conjunto com a avaliação subjetiva, confirmam que o *AG-Klatt* apresentou resultados melhores, o que é importante dado que métricas objetivas que se correlacionam com avaliações subjetivas ainda é um problema de investigação aberto.

Capítulo 6

Conclusão

Este trabalho apresenta a descrição sobre o problema de estimação dos valores dos parâmetros de um sintetizador por formantes, com o objetivo de imitar uma voz alvo. A combinação dos parâmetros de entrada do sintetizador para imitar a voz humana não é tarefa simples, pois existe um número razoável de parâmetros a ser combinado e cada um deles possui um intervalo de valores aceitáveis que deve ser minuciosamente ajustado para produzir uma determinada voz. A dificuldade maior é que a combinação de vários valores diferentes de parâmetros pode levar ao mesmo sinal sintético, ou seja, é um problema de muitos para um.

O arcabouço *newGASpeech* desenvolvido empregou o processo de análise-por-síntese em um AG para estimar os parâmetros de um sintetizador de voz por formantes. Neste caso, dois sintetizadores foram acoplados, sendo possível a escolha por qual sintetizador de voz utilizar, Klatt ou HLSyn. Os parâmetros estimados foram melhorados a cada iteração (mecanismo de análise-por-síntese), através da avaliação dos indivíduos feita pela função objetivo mais adequada para o problema, ou seja, o EQM entre os sinais de voz alvo e sintético. Portanto, para medir o erro entre os sinais é necessária previamente a síntese dos indivíduos. Apesar do arcabouço ter a opção de ser executado multiobjetivo, com três métricas implementadas, os melhores resultados foram alcançados com a execução mono-objetivo para estimar os vinte e cinco parâmetros do Klatt, que variam segundo o DECktalk, e os treze parâmetros do HLSyn, se comparados aos resultados obtidos pelo baseline. Foi definida, também, uma metodologia para avaliar os sinais de voz sintéticos gerados pelo AG e pelo Winsnoori, composta por cinco métricas objetivas e uma avaliação subjetiva.

Os experimentos abrangeram como alvo, vozes sintéticas e naturais, e os dois sintetizadores mencionados. Os experimentos com vozes sintéticas (masculinas e femininas) apresentaram resultados semelhantes para estimar os vinte e cinco parâmetros do Klatt e o arcabouço

desenvolvido obteve vozes com qualidade melhor que aquelas obtidas através do baseline, com erro percentual menor do que 10% para parâmetros importantes, como aqueles pertencentes à fonte de voz ($F0$, AV, OQ e DI). Além disso, detectou-se que parâmetros como o AF, $B1$, $B2$ e BNZ são pouco sensíveis a variações em seus valores, apresentando pouca diferença no EQM e convergindo mais lentamente para os valores ótimos.

Os experimentos com voz natural foram avaliados objetivamente e também subjetivamente, pois os falantes utilizados eram desconhecidos. A etapa da avaliação subjetiva presente na metodologia adotada foi bastante importante para ratificar a qualidade das vozes sintéticas produzidas. A Tabela 6.1 apresenta os valores das métricas obtidas nas avaliações realizadas nas vozes sintéticas masculinas produzidas a partir de vozes alvo naturais. O AG-Klatt obteve melhor desempenho com médias de EQM e D_{LE} 70% e 75% menores que àquelas obtidas pelo AG-HLSyn e pelo baseline. Os valores do PESQ foram semelhantes para o AG com ambos sintetizadores e o Winsnoori obteve média 52% menor. Na avaliação subjetiva, o AG-Klatt foi o que gerou vozes sintéticas mais próximas das vozes alvo naturais, com média -0.5. Em contrapartida, os resultados mais ruins foram obtidos através do AG-HLSyn, com qualidade das vozes variando entre "pior" e "ligeiramente pior". A RSR indica que a potência das vozes sintéticas do AG-Klatt é maior que 5.14 em relação ao ruído. Nesta métrica, o pior resultado foi com as vozes sintéticas do baseline, com média de 0.28.

Tabela 6.1: Média das métricas para falantes masculinos (vozes alvo naturais).

Média das métricas	*AG-Klatt*	*AG-HLSyn*	Wsno
EQM	59	184.7	194.1
D_{LE}	2.4	9.4	8.8
RSR	5.1	2.3	0.3
PESQ	2.8	2	1.4
Avaliação subjetiva	-0.5	-1.7	-1

Uma das maiores dificuldades encontradas é o fato da dimensão do espaço de busca do AG ser muito grande em virtude da quantidade de parâmetros que devem ser combinados para compor o arquivo de entrada do sintetizador de voz. Aliado a isso, ainda existe a necessidade do arcabouço estimar os parâmetros para o quadro atual e mais alguns quadros adiante (Look-aheads). Isto faz com que o espaço de busca tenha uma dimensão bastante extensa, dificultando a busca por valores ótimos. Ainda em relação à convergência, parâmetros com alto valor de erro percentual indicam que estes são menos sensíveis à variações, portanto,

dificultando a convergência deles para os valores corretos.

O *newGASpeech* [estimation of Klatt parameters, 2015], está disponibilizado gratuitamente, oferecendo, assim, uma opção de código aberto, e melhor que o baseline, para usuários que queiram estimar parâmetros de um sintetizador por formantes, com o objetivo de imitar voz humana. Este arcabouço é uma solução fácil de usar, podendo ser acoplada e/ou ajustada para funcionar em qualquer tipo de aplicação que se deseje fazer imitação da voz, pois, apesar de ter mais de três décadas, os sintetizadores por formantes são bastante populares entre foneticistas e profissionais da fala.

6.1 Trabalhos Futuros

Apesar do arcabouço desenvolvido suprir a tarefa de imitação automática da voz, algumas melhorias ainda são necessárias em sua implementação como aprimorar o AG para que a convergência para os valores corretos dos parâmetros do sintetizador seja mais rápida, utilizando para isso a mineração de dados para a descoberta de padrões, por exemplo. Uma outra opção para trabalhos futuros é estruturar o código de maneira modular para que seja possível o acoplamento de outros sintetizadores de voz, incluindo aqueles para a síntese em português brasileiro, possibilitando dessa maneira avaliar a eficiência do arcabouço desenvolvido para imitar vozes em outro idioma.

Outras possibilidades mais simples como continuação deste trabalho abrangem: investigar métricas objetivas que contrastem entre si para avaliar vozes (funções de minimização versus maximização) e que possam ser empregadas como funções objetivo, permitindo assim a utilização efetiva do *newGASpeech* multiobjetivo, avaliar as vozes sintéticas produzidas pelo arcabouço através de sistemas de reconhecimento de voz e de locutor que utilizem MOM, verificar a eficiência do AG em imitar vozes que apresentem entonação e emotividade.

Referências Bibliográficas

[P80, 1996] (1996). ITU-T P.800. Methods for subjective determination of transmission quality.

[Pes, 2001] (2001). ITU-T recommendation P.862. Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs.

[P56, 2004] (2004). ITU-T recommendation P.563. Single-ended method for objective speech quality assessment in narrow-band telephony applications.

[HTS, 2010] (Visited on August, 2010.). http://hts.sp.nitech.ac.jp/.

[HTK, 2010] (Visited on March, 2010.). http://htk.eng.cam.ac.uk/.

[Agbele et al., 2012] Agbele, K. K., Adesina, A. O., Ekong, D. O., and Ayangbekun, O. J. (2012). State-of-the-art review on relevance of genetic algorithm to internet web search. *Applied Computational Intelligence and Soft Computing*, 2012:152385:1–152385:7.

[Anumanchipalli et al., 2010] Anumanchipalli, G. K., Cheng, Y. C., Fernandez, J., Huang, X., Mao, Q., and Black, A. W. (2010). KlaTTStat: Knowledge-based Statistical Parametric Speech Synthesis. In *7th ISCA Workshop on Speech Synthesis*.

[Aylett and Yamagishi, 2008] Aylett, M. P. and Yamagishi, J. (2008). Combining statistical parameteric speech synthesis and unit-selection for automatic voice cloning. In *Langtech*.

[Bangayan et al., 1997] Bangayan, P., Long, C., Alwan, A. A., Kreiman, J., and Gerratt, B. R. (1997). Analysis by synthesis of pathological voices using the Klatt synthesizer. *Speech Communication*, 22(4):343–368.

[Bansal et al., 2015] Bansal, A., Kaiwartya, O., Singh, R., and Prakash, S. (2015). Maximizing fault tolerance and minimizing delay in virtual network embedding using nsga-ii. In *Pro-*

ceedings of the Third International Symposium on Women in Computing and Informatics, WCI '15, pages 124–130. ACM.

[Black, 2006] Black, A. W. (2006). CLUSTERGEN: A statistical parametric synthesizer using trajectory modeling. *Interspeech 2006*, pages 1762–1765.

[Blickle, 1996] Blickle, T. (1996). *Theory of Evolutionary Algorithms and Application to System Synthesis*. PhD thesis, Swiss Federal Institute of Technology.

[Borges et al., 2008] Borges, J., Couto, I., Oliveira, F., Imbiriba, T., and Klautau, A. (2008). GASpeech: A framework for automatically estimating input parameters of Klatt's speech synthesizer. *Neural Networks*, pages 81–86.

[Brito, 2007] Brito, J. (2007). Genetic learning of vocal tract area functions for articulatory synthesis of spanish vowels. *Applied Soft Computing*, 7(3):1035–1043.

[Bulut et al., 2002] Bulut, M., Narayanan, S. S., and Syrdal, A. K. (2002). Expressive speech synthesis using a concatenative synthesizer. In *International Conference on Spoken Language Processing*, pages 1265–1268.

[Carvalho and Araújo, 2009] Carvalho, A. G. and Araújo, A. F. R. (2009). Improving nsga-ii with an adaptive mutation operator. In *GECCO (Companion)*, pages 2697–2700. ACM.

[Couto and Borges, 2008] Couto, I. C. and Borges, J. V. M. (2008). Otimização multi-objetivo aplicada à síntese de voz. Trabalho de Conclusão de Curso apresentado para obtenção do grau de Engenheiro em Engenharia da Computação, do Instituto de Tecnologia, da Faculdade de Engenharia da Computação da Universidade Federal do Pará.

[Deb, 2001] Deb, K. (2001). *Multi-Objective Optimization using Evolutionary Algorithms*. Wiley.

[Deb et al., 2000] Deb, K., Agrawal, S., and Pratap, A. (2000). A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II. In *Proceedings of the Parallel Problem Solving from Nature VI*, pages 849–858.

[estimation of Klatt parameters, 2015] estimation of Klatt parameters, A. A. (Visited 10-February-2015). http://www.laps.ufpa.br/autoklatt.

[Fraj et al., 2012] Fraj, S., Schoentgen, J., and Grenez, F. (2012). Development and perceptual assessment of a synthesizer of disordered voices. *The Journal of the Acoustical Society of America*, 132(4):2603–2615.

[Gamal et al., 2015] Gamal, M., Morsy, E., and Fathy, A. (2015). Multi-objective transmitters placement problem in wireless networks. In *Proceedings of the Sixth International Symposium on Information and Communication Technology*, SoICT 2015, pages 156–162. ACM.

[Hallahan, 1995] Hallahan, W. I. (1995). DECtalk software: Text-to-speech technology and implementation. *Digital Technical Journal*, 7(4):5–19.

[Hamdan, 2012] Hamdan, M. (2012). On the disruption-level of polynomial mutation for evolutionary multi-objective optimisation algorithms. *Computing and Informatics*, 29(5):783–800.

[Hanson et al., 1999] Hanson, H. M., McGowan, R. S., Stevens, K. N., and Beaudoin, R. E. (1999). Development of rules for controlling the HLsyn speech synthesizer. In *Proceedings of the Acoustics, Speech, and Signal Processing, 1999. On 1999 IEEE International Conference - Volume 01*, ICASSP '99, pages 85–88, Washington, DC, USA. IEEE Computer Society.

[Ho et al., 1999] Ho, C. W., Lee, K. H., and Leung, K. S. (1999). A genetic algorithm based on mutation and crossover with adaptive probabilities. In *Proceedings of the 1999 Congress on Evolutionary Computation*, volume 1, page 775 Vol. 1.

[Howard and Huckvale, 2005] Howard, I. S. and Huckvale, M. A. (2005). Training a vocal tract synthesizer to imitate speech using distal learning. In *Proceedings of InterSpeech 2005*.

[Imbens et al., 2005] Imbens, G. W., Newey, W. K., and Ridder, G. (2005). Mean-square-error calculations for average treatment effects. *IEPR Working Paper No. 05.34*.

[Jinachitra and Smith III, 2005] Jinachitra, P. and Smith III, J. O. (2005). Joint estimation of glottal source and vocal tract for vocal synthesis using Kalman smoothing and EM algorithm. In *Applications of Signal Processing to Audio and Acoustics*, pages 327–330.

[Kain et al., 2004] Kain, E., Niu, X., Hosom, J. P., Miao, Q., and Santen, J. V. (2004). Formant resynthesis of dysarthric speech. In *In: IEEE Workshop on Speech Synthesis*, pages 25–30.

[Keller, 1994] Keller, E., editor (1994). *Fundamentals of speech synthesis and speech recognition: basic concepts, state-of-the-art and future challenges*. John Wiley and Sons Ltd., Chichester, UK.

[Khatun et al., 2015] Khatun, S., Alam, H. U., and Shatabda, S. (2015). An efficient genetic algorithm for discovering diverse-frequent patterns. *Computing Research Repository (CoRR)*, abs/1507.05275.

[King, 2011] King, S. (2011). An introduction to statistical parametric speech synthesis. *Sadhana-Academy proceedings in engineering sciences*, 36(5):837–852.

[Klatt, 1980] Klatt, D. (1980). Software for a cascade / parallel formant synthesizer. *Journal of the Acoustical Society of America*, 67:971–95.

[Klatt and Klatt, 1990] Klatt, D. and Klatt, L. (1990). Analysis, synthesis, and perception of voice quality variations among female and male speakers. *Journal of the Acoustical Society of America*, 87:820–57.

[Kondo, 2012] Kondo, K. (2012). *Subjective Quality Measurement of Speech - Its Evaluation, Estimation and Applications*. Springer.

[Laprie and Bonneau, 2002] Laprie, Y. and Bonneau, A. (2002). A copy synthesis method to pilot the Klatt synthesizer. In *International Conference on Speech and Language Processing*, page 4 p, Denver, USA.

[Laprie and Bonneau, 2007] Laprie, Y. and Bonneau, A. (2007). Construction of perception stimuli with copy synthesis. In des Saarlandes, U., editor, *16th International Congress of Phonetic Sciences - ICPhS 2007*.

[Lemmetty, 1999] Lemmetty, S. (1999). *Review of Speech Synthesis Technology*. PhD thesis, Department Electrical and Communication Engineering - Helsinki University of Technology.

[Li et al., 2015] Li, J., Chen, J., Xin, B., and Dou, L. (2015). Solving multi-objective multi-stage weapon target assignment problem via adaptive nsga-ii and adaptive moea/d: A comparison study. In *Evolutionary Computation (CEC), 2015 IEEE Congress on*, pages 3132–3139.

[Liu and Kewley-Port, 2004] Liu, C. and Kewley-Port, D. (2004). Straight: A new speech synthesizer for vowel formant discrimination. *Acoustic Research Letters Online*, pages 31–36.

[Ly et al., 2015] Ly, D. T. H., Hanh, N. T., Binh, H. T. T., and Nghia, N. D. (2015). An improved genetic algorithm for maximizing area coverage in wireless sensor networks. In *Proceedings of the Sixth International Symposium on Information and Communication Technology*, SoICT 2015, pages 61–66. ACM.

[Ma et al., 2015] Ma, X., Liu, F., Qi, Y., Wang, X., Li, L., Jiao, L., Yin, M., and Gong, M. (2015). A multiobjective evolutionary algorithm based on decision variable analyses for multi-objective optimization problems with large scale variables. *Evolutionary Computation, IEEE Transactions on*, PP(99):1–1.

[Malfait et al., 2006] Malfait, L., Berger, J., and Kastner, M. (2006). P.563—The ITU-T Standard for Single-Ended Speech Quality Assessment. *Audio, Speech and Language Processing, IEEE Transactions on [see also Speech and Audio Processing, IEEE Transactions on]*, 14(6):1924–1934.

[Martin, 1981] Martin, P. (1981). Extraction de la fréquence fondamentale par intercorrélation avec une fonction peigne. *Actes des 12èmes Journées d'Etudes sur la Parole*, pages 223–232.

[Neto et al., 2013] Neto, M. U., Silva, J. E., Silva, D. A., Gomes, L. C. T., Campolina, T. A. M., Yehia, H. C., N., V. M., and Sansão, J. P. H. (2013). Parametric analysis of speech signals based on estimation of joint source-filter model using evolutionary computation. In *CDS 2013 : The Seventh International Conference on Digital Society*, pages 32–36.

[Nunez and Attoh-Okine, 2014] Nunez, S. G. and Attoh-Okine, N. (2014). Metaheuristics in big data: An approach to railway engineering. In *Big Data (Big Data), 2014 IEEE International Conference on*, pages 42–47.

[Padhye, 2012] Padhye, N. (2012). Evolutionary approaches for real world applications in 21st century. In *Proceedings of the 14th Annual Conference Companion on Genetic and Evolutionary Computation*, GECCO '12, pages 43–48.

[Philippsen et al., 2014] Philippsen, A., Reinhart, F. R., and Wrede, B. (2014). Learning how to speak: Imitation-based refinement of syllable production in an articulatory-acoustic model. In *IEEE Int. Conf. on Development and Learning and on Epigenetic Robotics (ICDL)*, pages 187–192.

[Rabiner, 1989] Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

[Raghuwanshi and Kakde, 2004] Raghuwanshi, M. M. and Kakde, O. G. (2004). Survey on multiobjective evolutionary and real coded genetic algorithms. In *Proceedings of the 8th Asia Pacific symposium on intelligent and evolutionary systems*, pages 150–161.

[Reeves and Rowe, 2003] Reeves, C. R. and Rowe, J. E. (2003). *Genetic algorithms: principles and perspectives. A guide to GA theory*. Kluwer Academic Publishers.

[Rothauser et al., 1969] Rothauser, E. H., Chapman, W. D., Guttman, N., Hecker, M. H. L., Nordby, K. S., Silbiger, H. R., Urbanek, G. E., and Weinstock, M. (1969). IEEE recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics*, 17:225–246.

[Rubin et al., 1981] Rubin, P., Baer, T., and Mermelstein, P. (1981). An articulatory synthesizer for perceptual research. *Journal of the Acoustical Society of America 70*, 2:321–328.

[Russel and Norvig, 2013] Russel, S. J. and Norvig, P. (2013). *Inteligência Artificial*. Elsevier, 3 edition.

[Sahil et al., 2015] Sahil, Sood, S., Mehmi, S., and Dogra, S. (2015). Artificial intelligence for designing user profiling system for cloud computing security: Experiment. In *Computer Engineering and Applications (ICACEA), 2015 International Conference on Advances in*, pages 51–58.

[Schafer et al., 2011] Schafer, M., Schnelling, C., Geiser, B., and Vary, P. (2011). *Konferenz Elektronische Sprachsignalverarbeitung (ESSV)*, volume 61. TUDpress Verlag der Wissenschaften GmbH, 1 edition.

[Shrivastav and Sapienza, 2006] Shrivastav, R. and Sapienza, C. M. (2006). Some difference limens for the perception of breathiness. *The Journal of the Acoustical Society of America*, 120(1):416–423.

[Singh and Kalra, 2014] Singh, S. and Kalra, M. (2014). Task scheduling optimization of independent tasks in cloud computing using enhanced genetic algorithm. *International Journal of Application or Innovation in Engineering and Management (IJAIEM)*, 3:286–291.

[Sivanandam and Deepa, 2008] Sivanandam, S. N. and Deepa, S. N. (2008). *Introduction to Genetic Algorithm*. Springer.

[Soares, 1997] Soares, G. L. (1997). Algoritmos genéticos: Estudos, novas técnicas e aplicações. Master's thesis, Universidade Federal de Minhas Gerais.

[Srinivas and Deb, 1994] Srinivas, N. and Deb, K. (1994). Multiobjective optimization using nondominated sorting in genetic algorithms. *Evolutionary Computation*, 2(3):221–248.

[TIDIGITS, 2014] TIDIGITS (Visited 20-June-2014). https://catalog.ldc.upenn.edu/ldc93s10.

[Tokuda et al., 2013] Tokuda, K., Nankaku, Y., Toda, T., Zen, H., Yamagishi, J., and Oura, K. (2013). Speech synthesis based on Hidden Markov Models. *Proceedings of the IEEE*, 101(5):1234–1252.

[Tokuda et al., 2000] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., and Kitamura, T. (2000). Speech parameter generation algorithms for HMM-based speech synthesis. *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, 3:1315–1318.

[Triantafyllidis et al., 2015] Triantafyllidis, K., Bondarev, E., and de With, P. H. N. (2015). Guided rule-based multi-objective optimization for real-time distributed systems. In *Software Engineering and Advanced Applications (SEAA), 2015 41st Euromicro Conference on*, pages 224–232.

[Trindade et al., 2013] Trindade, J., Araujo, F., Klautau, A., and Batista, P. (2013). A genetic algorithm with look-ahead mechanism to estimate formant synthesizer input parameters. In *IEEE Congress on Evolutionary Computation*, pages 3035–3042. IEEE.

[Verma and Kumar, 2014] Verma, V. K. and Kumar, B. (2014). Genetic algorithm: an overview and its application. *International Journal of advanced studies in Computer Science and Engineering*, 3:21–27.

[Wang et al., 2015] Wang, X., Li, X., and Leung, V. C. M. (2015). Artificial intelligence-based techniques for emerging heterogeneous network: State of the arts, opportunities, and challenges. *IEEE Access*, 3:1379–1391.

[Yarrington et al., 2005] Yarrington, D., Pennington, C., Gray, J., and Bunnell, H. T. (2005). A system for creating personalized synthetic voices. In *Proceedings of the 7th International ACM SIGACCESS Conference on Computers and Accessibility*, Assets '05, pages 196–197, New York, NY, USA. ACM.

[Young, 2005] Young, S. (2005). *The HTK Book*. Cambridge University Engineering Department, version 3.3 edition.

[Yuan et al., 2015] Yuan, Y., Xu, H., Wang, B., and Yao, X. (2015). A new dominance relation based evolutionary algorithm for many-objective optimization. *Evolutionary Computation, IEEE Transactions on*, PP(99):1–1.

[Zen et al., 2007] Zen, H., Nose, T., Yamagish, J., and Sako, S. (2007). The HMM-based speech synthesis system (HTS) version 2.0. *Sixth ISCA Workshop on Speech Synthesis*, pages 294–299.

[Zhan et al., 2015] Zhan, Z., Liu, X., Gong, Y., Zhang, J., Chung, H. S., and Li, Y. (2015). Cloud computing resource scheduling and a survey of its evolutionary approaches. *ACM Computing Surveys*, 47(4):63:1–63:33.

Apêndice A

Parâmetros do Sintetizador de Klatt (Versão KLSYN88)

Parâmetro	**Valor mínimo**	**Valor máximo**	**Quantidade de valores possíveis**
F0	0	5000	5001
AV	0	80	81
OQ	10	99	90
SQ	100	500	401
TL	0	41	42
FL	0	100	101
DI	0	100	101
AH	0	80	81
AF	0	80	81
F1	180	1300	1121
B1	30	1000	971
DF1	0	100	101
DBI	0	400	401
F2	550	3000	2451
B2	40	1000	961
F3	1200	4800	3601
B3	60	1000	941
F4	2400	4990	2591
B4	100	1000	901
F5	3000	4990	1991
B5	100	1500	1401
F6	3000	4990	1991
B6	100	4000	3901
FNP	180	500	321
BNP	40	1000	961

Parâmetro	**Valor mínimo**	**Valor máximo**	**Quantidade de valores possíveis**
FNZ	180	800	621
BNZ	40	1000	961
FTP	300	3000	2701
BTP	40	1000	961
FTZ	300	3000	2701
BTZ	40	2000	1961
A2F	0	80	81
A3F	0	80	81
A4F	0	80	81
A5F	0	80	81
A6F	0	80	81
AB	0	80	81
B2F	40	1000	961
B3F	60	1000	941
B4F	100	1000	901
B5F	100	1500	1401
B6F	100	4000	3901
ANV	0	80	81
A1V	0	80	81
A2V	0	80	81
A3V	0	80	81
A4V	0	80	81
ATV	0	80	81

Apêndice B

Exemplo de arquivo de entrada do KLSYN88

Devido ao extenso número de parâmetros do Sintetizador de Klatt, o arquivo será exibido na próxima página em orientação paisagem.

```
F0 AV OQ SQ TL FL DI AH AF F1 B1 DF1 DB1 F2 B2 F3 B3 F4 B4 F5 B5 F6 B6 FNP BNP FNZ BNZ FTP BTP FTZ BTZ A2F A3F A4F A5F A6F AB B2F B3F B4F B5F B6F ANV A1V A2V A3V A4V ATV
0 0 99 0 41 0 0 0 0 500 200 0 0 1400 90 3100 254 4200 402 4500 552 4990 1000 570 80 500 200 1000 200 1000 200 0 0 0 0 0 0 250 320 350 500 1500 0 0 0 0 0 0
0 0 99 0 41 0 0 0 0 500 200 0 0 1400 90 3100 254 4200 402 4500 552 4990 1000 570 80 500 200 1000 200 1000 200 0 0 0 0 0 0 250 320 350 500 1500 0 0 0 0 0 0
0 0 99 0 41 0 0 0 0 500 200 0 0 1400 90 3100 254 4200 402 4500 552 4990 1000 570 80 500 200 1000 200 1000 200 0 0 0 0 0 0 250 320 350 500 1500 0 0 0 0 0 0
0 0 99 0 41 0 0 0 0 500 200 0 0 1400 90 3100 254 4200 402 4500 552 4990 1000 570 80 500 200 1000 200 1000 200 0 0 0 0 0 0 250 320 350 500 1500 0 0 0 0 0 0
0 0 99 0 41 0 0 0 0 500 200 0 0 1400 90 3100 254 4200 402 4500 552 4990 1000 570 80 500 200 1000 200 1000 200 0 0 0 0 0 0 250 320 350 500 1500 0 0 0 0 0 0
0 0 99 0 41 0 0 0 0 500 200 0 0 1400 90 3100 254 4200 402 4500 552 4990 1000 570 80 500 200 1000 200 1000 200 0 0 0 0 0 0 250 320 350 500 1500 0 0 0 0 0 0
0 0 99 0 41 0 0 0 0 500 200 0 0 1400 90 3100 254 4200 402 4500 552 4990 1000 570 80 500 200 1000 200 1000 200 0 0 0 0 0 0 250 320 350 500 1500 0 0 0 0 0 0
0 0 99 0 41 0 0 0 0 500 200 0 0 1400 90 3100 254 4200 402 4500 552 4990 1000 570 80 500 200 1000 200 1000 200 0 0 0 0 0 0 250 320 350 500 1500 0 0 0 0 0 0
0 0 99 0 41 0 0 0 0 500 200 0 0 1400 90 3100 254 4200 402 4500 552 4990 1000 570 80 500 200 1000 200 1000 200 0 0 0 0 0 0 250 320 350 500 1500 0 0 0 0 0 0
0 0 99 0 41 0 0 0 0 500 200 0 0 1400 90 3100 254 4200 402 4500 552 4990 1000 570 80 500 200 1000 200 1000 200 0 0 0 0 0 0 250 320 350 500 1500 0 0 0 0 0 0
0 0 99 0 41 0 0 0 0 500 200 0 0 1400 90 3100 254 4200 402 4500 552 4990 1000 570 80 500 200 1000 200 1000 200 0 0 0 0 0 0 250 320 350 500 1500 0 0 0 0 0 0
0 0 99 0 41 0 0 0 0 500 200 0 0 1400 90 3100 254 4200 402 4500 552 4990 1000 570 80 500 200 1000 200 1000 200 0 0 0 0 0 0 250 320 350 500 1500 0 0 0 0 0 0
0 0 99 0 41 0 0 0 0 500 200 0 0 1400 90 3100 254 4200 402 4500 552 4990 1000 570 80 500 200 1000 200 1000 200 0 0 0 0 0 0 250 320 350 500 1500 0 0 0 0 0 0
0 0 99 0 41 0 0 0 0 500 200 0 0 1400 90 3100 254 4200 402 4500 552 4990 1000 570 80 500 200 1000 200 1000 200 0 0 0 0 0 0 250 320 350 500 1500 0 0 0 0 0 0
0 0 99 0 41 0 0 0 0 500 200 0 0 1400 90 3100 254 4200 402 4500 552 4990 1000 570 80 500 200 1000 200 1000 200 0 0 0 0 0 0 250 320 350 500 1500 0 0 0 0 0 0
0 0 99 0 41 0 0 0 0 500 200 0 0 1400 90 3100 254 4200 402 4500 552 4990 1000 570 80 500 200 1000 200 1000 200 0 0 0 0 0 0 250 320 350 500 1500 0 0 0 0 0 0
0 0 99 0 41 0 0 0 0 500 200 0 0 1400 90 3100 254 4200 402 4500 552 4990 1000 570 80 500 200 1000 200 1000 200 0 0 0 0 0 0 250 320 350 500 1500 0 0 0 0 0 0
0 0 99 0 41 0 0 0 0 500 200 0 0 1400 90 3100 254 4200 402 4500 552 4990 1000 570 80 500 200 1000 200 1000 200 0 0 0 0 0 0 250 320 350 500 1500 0 0 0 0 0 0
0 0 99 0 41 0 0 34 52 500 910 0 0 1400 343 3100 271 4200 410 4500 560 4990 1000 602 80 500 200 1000 200 1000 200 0 0 0 0 0 50 250 320 350 500 1500 0 0 0 0 0 0
0 0 99 0 41 0 0 30 53 500 970 0 0 1400 328 3100 273 4200 411 4500 561 4990 1000 606 80 500 200 1000 200 1000 200 0 0 0 0 0 50 250 320 350 500 1500 0 0 0 0 0 0
0 0 99 0 41 0 0 27 54 500 1009 0 0 1400 311 3100 274 4200 412 4500 562 4990 1000 609 80 500 200 1000 200 1000 200 0 0 0 0 0 50 250 320 350 500 1500 0 0 0 0 0 0
0 0 99 0 41 0 0 27 54 500 805 0 0 1400 262 3100 244 4205 397 4500 547 4990 1000 551 80 500 200 1000 200 1000 200 0 0 0 0 0 50 250 320 350 500 1500 0 0 0 0 0 0
0 0 99 0 41 0 0 28 54 491 474 0 0 1400 213 3100 212 4210 381 4500 531 4990 1000 500 200 500 200 1000 200 1000 200 0 0 0 0 0 50 250 320 350 500 1500 0 0 0 0 0 0
2137 57 48 0 13 0 1 38 53 500 200 0 0 1400 90 3100 150 4215 350 4500 500 4990 1000 700 80 500 200 1000 200 1000 200 60 0 0 40 0 0 250 320 350 500 1500 0 0 0 0 0 0
2150 58 47 0 12 0 2 38 31 500 200 0 0 1412 90 3130 150 4220 350 4500 500 4990 1000 707 80 500 200 1000 200 1000 200 0 0 0 0 0 0 250 320 350 500 1500 0 0 0 0 0 0
2159 58 47 0 12 0 3 38 24 500 200 0 0 1423 90 3160 150 4225 350 4500 500 4990 1000 713 80 500 200 1000 200 1000 200 0 0 0 0 0 0 250 320 350 500 1500 0 0 0 0 0 0
2166 58 47 0 12 0 4 38 18 500 200 0 0 1435 90 3190 150 4230 350 4500 500 4990 1000 720 80 500 200 1000 200 1000 200 0 0 0 0 0 0 250 320 350 500 1500 0 0 0 0 0 0
2172 58 47 0 12 0 4 38 14 500 200 0 0 1446 90 3220 150 4235 350 4500 500 4990 1000 726 80 500 200 1000 200 1000 200 0 0 0 0 0 0 250 320 350 500 1500 0 0 0 0 0 0
2176 58 47 0 12 0 4 38 10 500 200 0 0 1458 90 3250 150 4240 350 4500 500 4990 1000 733 80 500 200 1000 200 1000 200 0 0 0 0 0 0 250 320 350 500 1500 0 0 0 0 0 0
2177 58 46 0 12 0 4 38 7 500 200 0 0 1469 90 3242 150 4245 350 4500 500 4990 1000 739 80 500 200 1000 200 1000 200 0 0 0 0 0 0 250 320 350 500 1500 0 0 0 0 0 0
2177 58 46 0 12 0 4 38 6 500 200 0 0 1481 90 3233 150 4250 350 4500 500 4990 1000 746 80 500 200 1000 200 1000 200 0 0 0 0 0 0 250 320 350 500 1500 0 0 0 0 0 0
```

```
2177 58 46 0 12 0 4 38 5 500 200 0 0 1493 90 3225 150 4255 350 4500 500 4990 1000 749 80 500 200 1000 200 1000 200 0 0 0 0 0 0 250 320 350 500 1500 0 0 0 0 0 0
2187 58 46 0 13 0 4 38 5 500 200 0 0 1504 90 3217 150 4260 350 4500 500 4990 1000 749 80 500 200 1000 200 1000 200 0 0 0 0 0 0 250 320 350 500 1500 0 0 0 0 0 0
2196 58 47 0 13 0 4 38 6 500 200 0 0 1516 90 3208 150 4265 350 4500 500 4990 1000 749 80 500 200 1000 200 1000 200 0 0 0 0 0 0 250 320 350 500 1500 0 0 0 0 0 0
2206 58 47 0 13 0 4 38 7 500 200 0 0 1527 90 3200 150 4270 350 4500 500 4990 1000 749 80 500 200 1000 200 1000 200 0 0 0 0 0 0 250 320 350 500 1500 0 0 0 0 0 0
2216 58 47 0 13 0 4 38 8 500 200 0 0 1539 90 3192 150 4275 350 4500 500 4990 1000 749 80 500 200 1000 200 1000 200 0 0 0 0 0 0 250 320 350 500 1500 0 0 0 0 0 0
2225 58 47 0 13 0 4 38 9 500 200 0 0 1551 90 3183 150 4280 350 4500 500 4990 1000 749 80 500 200 1000 200 1000 200 0 0 0 0 0 0 250 320 350 500 1500 0 0 0 0 0 0
```

Apêndice C

Harvard Sentences

Subconjunto de 240 sentenças, compreendendo 24 agrupamentos ($H1$ à $H24$) de frases sintetizadas através do Klatt88 para os 6 falantes do DECtalk.

H1 Harvard Sentences

1. The birch canoe slid on the smooth planks.
2. Glue the sheet to the dark blue background.
3. It's easy to tell the depth of a well.
4. These days a chicken leg is a rare dish.
5. Rice is often served in round bowls.
6. The juice of lemons makes fine punch.
7. The box was thrown beside the parked truck.
8. The hogs were fed chopped corn and garbage.
9. Four hours of steady work faced us.
10. A large size in stockings is hard to sell.

H2 Harvard Sentences

1. The boy was there when the sun rose.
2. A rod is used to catch pink salmon.
3. The source of the huge river is the clear spring.
4. Kick the ball straight and follow through.
5. Help the woman get back to her feet.
6. A pot of tea helps to pass the evening.
7. Smoky fires lack flame and heat.
8. The soft cushion broke the man's fall.
9. The salt breeze came across from the sea.
10. The girl at the booth sold fifty bonds.

H3 Harvard Sentences

1. The small pup gnawed a hole in the sock.
2. The fish twisted and turned on the bent hook.

3. Press the pants and sew a button on the vest.
4. The swan dive was far short of perfect.
5. The beauty of the view stunned the young boy.
6. Two blue fish swam in the tank.
7. Her purse was full of useless trash.
8. The colt reared and threw the tall rider.
9. It snowed, rained, and hailed the same morning.
10. Read verse out loud for pleasure.

H4 Harvard Sentences

1. Hoist the load to your left shoulder.
2. Take the winding path to reach the lake.
3. Note closely the size of the gas tank.
4. Wipe the grease off his dirty face.
5. Mend the coat before you go out.
6. The wrist was badly strained and hung limp.
7. The stray cat gave birth to kittens.
8. The young girl gave no clear response.
9. The meal was cooked before the bell rang.
10. What joy there is in living.

H5 Harvard Sentences

1. A king ruled the state in the early days.
2. The ship was torn apart on the sharp reef.
3. Sickness kept him home the third week.
4. The wide road shimmered in the hot sun.
5. The lazy cow lay in the cool grass.
6. Lift the square stone over the fence.
7. The rope will bind the seven books at once.
8. Hop over the fence and plunge in.
9. The friendly gang left the drug store.
10. Mesh wire keeps chicks inside.

H6 Harvard Sentences

1. The frosty air passed through the coat.
2. The crooked maze failed to fool the mouse.
3. Adding fast leads to wrong sums.
4. The show was a flop from the very start.
5. A saw is a tool used for making boards.
6. The wagon moved on well oiled wheels.
7. March the soldiers past the next hill.
8. A cup of sugar makes sweet fudge.
9. Place a rosebush near the porch steps.
10. Both lost their lives in the raging storm.

H7 Harvard Sentences

1. We talked of the side show in the circus.
2. Use a pencil to write the first draft.
3. He ran half way to the hardware store.
4. The clock struck to mark the third period.
5. A small creek cut across the field.
6. Cars and busses stalled in snow drifts.
7. The set of china hit the floor with a crash.
8. This is a grand season for hikes on the road.
9. The dune rose from the edge of the water.
10. Those words were the cue for the actor to leave.

H8 Harvard Sentences

1. A yacht slid around the point into the bay.
2. The two met while playing on the sand.
3. The ink stain dried on the finished page.
4. The walled town was seized without a fight.
5. The lease ran out in sixteen weeks.
6. A tame squirrel makes a nice pet.
7. The horn of the car woke the sleeping cop.
8. The heart beat strongly and with firm strokes.
9. The pearl was worn in a thin silver ring.
10. The fruit peel was cut in thick slices.

H9 Harvard Sentences

1. The Navy attacked the big task force.
2. See the cat glaring at the scared mouse.
3. There are more than two factors here.
4. The hat brim was wide and too droopy.
5. The lawyer tried to lose his case.
6. The grass curled around the fence post.
7. Cut the pie into large parts.
8. Men strive but seldom get rich.
9. Always close the barn door tight.
10. He lay prone and hardly moved a limb.

H10 Harvard Sentences

1. The slush lay deep along the street.
2. A wisp of cloud hung in the blue air.
3. A pound of sugar costs more than eggs.
4. The fin was sharp and cut the clear water.
5. The play seems dull and quite stupid.
6. Bail the boat to stop it from sinking.
7. The term ended in late june that year.
8. A Tusk is used to make costly gifts.
9. Ten pins were set in order.
10. The bill was paid every third week.

H11 Harvard Sentences

1. Oak is strong and also gives shade.
2. Cats and Dogs each hate the other.
3. The pipe began to rust while new.
4. Open the crate but don't break the glass.
5. Add the sum to the product of these three.
6. Thieves who rob friends deserve jail.
7. The ripe taste of cheese improves with age.
8. Act on these orders with great speed.
9. The hog crawled under the high fence.
10. Move the vat over the hot fire.

H12 Harvard Sentences

1. The bark of the pine tree was shiny and dark.
2. Leaves turn brown and yellow in the fall.
3. The pennant waved when the wind blew.
4. Split the log with a quick, sharp blow.
5. Burn peat after the logs give out.
6. He ordered peach pie with ice cream.
7. Weave the carpet on the right hand side.
8. Hemp is a weed found in parts of the tropics.
9. A lame back kept his score low.
10. We find joy in the simplest things.

H13 Harvard Sentences

1. Type out three lists of orders.
2. The harder he tried the less he got done.
3. The boss ran the show with a watchful eye.
4. The cup cracked and spilled its contents.
5. Paste can cleanse the most dirty brass.
6. The slang word for raw whiskey is booze.
7. It caught its hind paw in a rusty trap.
8. The wharf could be seen at the farther shore.
9. Feel the heat of the weak dying flame.
10. The tiny girl took off her hat.

H14 Harvard Sentences

1. A cramp is no small danger on a swim.
2. He said the same phrase thirty times.
3. Pluck the bright rose without leaves.
4. Two plus seven is less than ten.
5. The glow deepened in the eyes of the sweet girl.
6. Bring your problems to the wise chief.
7. Write a fond note to the friend you cherish.
8. Clothes and lodging are free to new men.

9. We frown when events take a bad turn.
10. Port is a strong wine with a smoky taste.

H15 Harvard Sentences

1. The young kid jumped the rusty gate.
2. Guess the result from the first scores.
3. A salt pickle tastes fine with ham.
4. The just claim got the right verdict.
5. Those thistles bend in a high wind.
6. Pure bred poodles have curls.
7. The tree top waved in a graceful way.
8. The spot on the blotter was made by green ink.
9. Mud was spattered on the front of his white shirt.
10. The cigar burned a hole in the desk top.

H16 Harvard Sentences

1. The empty flask stood on the tin tray.
2. A speedy man can beat this track mark.
3. He broke a new shoelace that day.
4. The coffee stand is too high for the couch.
5. The urge to write short stories is rare.
6. The pencils have all been used.
7. The pirates seized the crew of the lost ship.
8. We tried to replace the coin but failed.
9. She sewed the torn coat quite neatly.
10. The sofa cushion is red and of light weight.

H17 Harvard Sentences

1. The jacket hung on the back of the wide chair.
2. At that high level the air is pure.
3. Drop the two when you add the figures.
4. A filing case is now hard to buy.
5. An abrupt start does not win the prize.
6. Wood is best for making toys and blocks.
7. The office paint was a dull, sad tan.
8. He knew the skill of the great young actress.
9. A rag will soak up spilled water.
10. A shower of dirt fell from the hot pipes.

H18 Harvard Sentences

1. Steam hissed from the broken valve.
2. The child almost hurt the small dog.
3. There was a sound of dry leaves outside.
4. The sky that morning was clear and bright blue.
5. Torn scraps littered the stone floor.
6. Sunday is the best part of the week.

7. The doctor cured him with these pills.
8. The new girl was fired today at noon.
9. They felt gay when the ship arrived in port.
10. Add the store's account to the last cent.

H19 Harvard Sentences

1. Acid burns holes in wool cloth.
2. Fairy tales should be fun to write.
3. Eight miles of woodland burned to waste.
4. The third act was dull and tired the players.
5. A young child should not suffer fright.
6. Add the column and put the sum here.
7. We admire and love a good cook.
8. There the flood mark is ten inches.
9. He carved a head from the round block of marble.
10. She has a smart way of wearing clothes.

H20 Harvard Sentences

1. The fruit of a fig tree is apple shaped.
2. Corn cobs can be used to kindle a fire.
3. Where were they when the noise started.
4. The paper box is full of thumb tacks.
5. Sell your gift to a buyer at a good gain.
6. The tongs lay beside the ice pail.
7. The petals fall with the next puff of wind.
8. Bring your best compass to the third class.
9. They could laugh although they were sad.
10. Farmers came in to thresh the oat crop.

H21 Harvard Sentences

1. The brown house was on fire to the attic.
2. The lure is used to catch trout and flounder.
3. Float the soap on top of the bath water.
4. A blue crane is a tall wading bird.
5. A fresh start will work such wonders.
6. The club rented the rink for the fifth night.
7. After the dance, they went straight home.
8. The hostess taught the new maid to serve.
9. He wrote his last novel there at the inn.
10. Even the worst will beat his low score.

H22 Harvard Sentences

1. The cement had dried when he moved it.
2. The loss of the second ship was hard to take.
3. The fly made its way along the wall.
4. Do that with a wooden stick.

5. Live wires should be kept covered.
6. The large house had hot water taps.
7. It is hard to erase blue or red ink.
8. Write at once or you may forget it.
9. The doorknob was made of bright clean brass.
10. The wreck occurred by the bank on Main Street.

H23 Harvard Sentences

1. A pencil with black lead writes best.
2. Coax a young calf to drink from a bucket.
3. Schools for ladies teach charm and grace.
4. The lamp shone with a steady green flame.
5. They took the axe and the saw to the forest.
6. The ancient coin was quite dull and worn.
7. The shaky barn fell with a loud crash.
8. Jazz and swing fans like fast music.
9. Rake the rubbish up and then burn it.
10. Slash the gold cloth into fine ribbons.

H24 Harvard Sentences

1. Try to have the court decide the case.
2. They are pushed back each time they attack.
3. He broke his ties with groups of former friends.
4. They floated on the raft to sun their white backs.
5. The map had an X that meant nothing.
6. Whitings are small fish caught in nets.
7. Some ads serve to cheat buyers.
8. Jerk the rope and the bell rings weakly.
9. A waxed floor makes us lose balance.
10. Madam, this is the best brand of corn.

Apêndice D

Histogramas

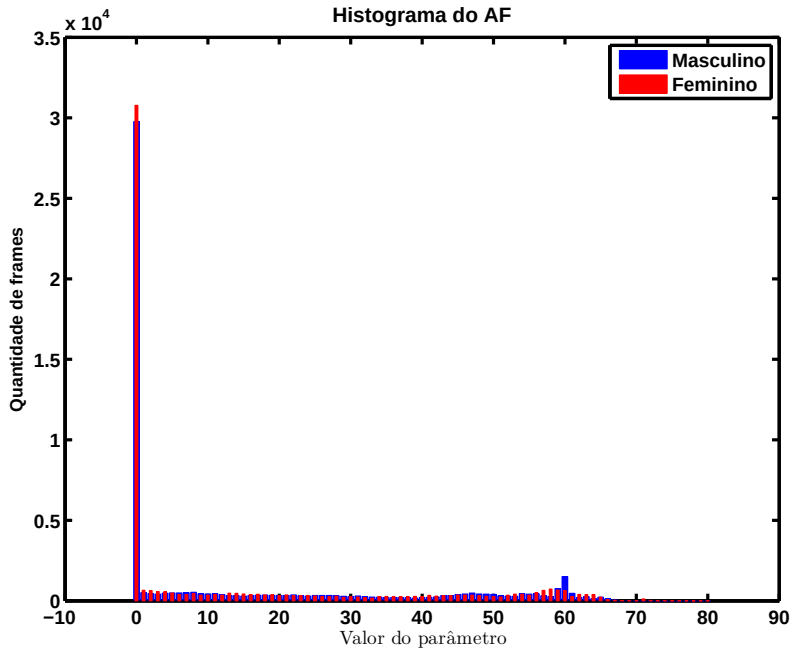

Figura D.1: Histograma do parâmetro AF para falantes masculinos e femininos.

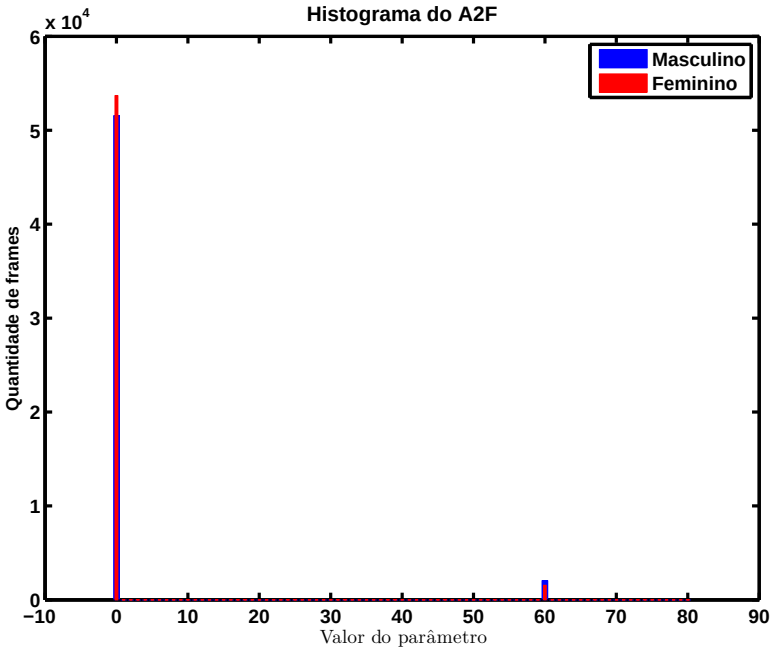

Figura D.2: Histograma do parâmetro A2F para falantes masculinos e femininos.

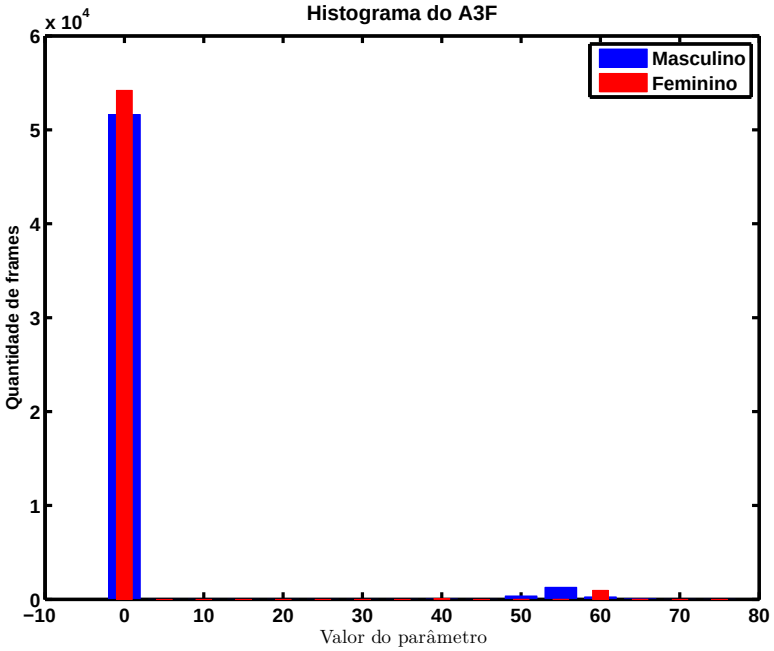

Figura D.3: Histograma do parâmetro A3F para falantes masculinos e femininos.

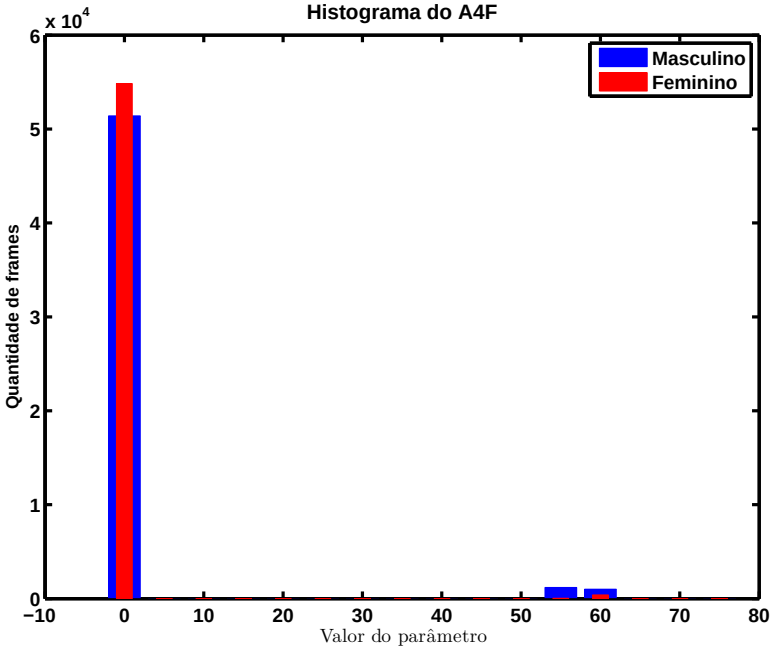

Figura D.4: Histograma do parâmetro A4F para falantes masculinos e femininos.

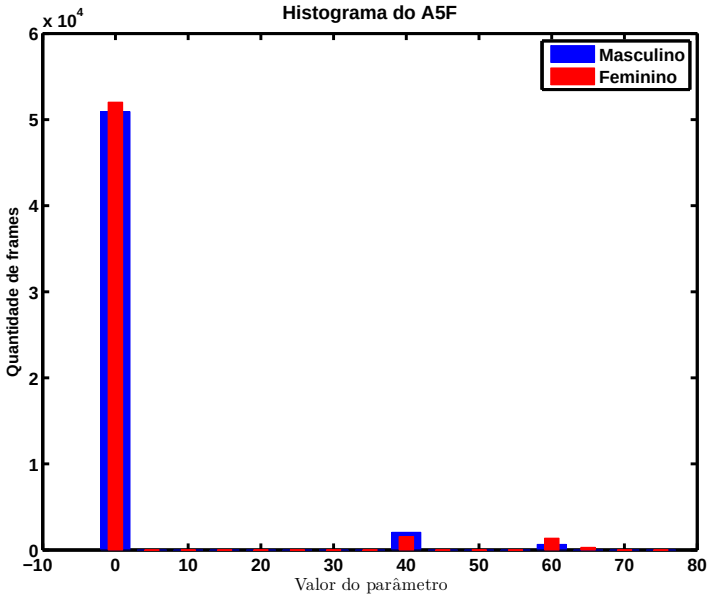

Figura D.5: Histograma do parâmetro A5F para falantes masculinos e femininos.

Apêndice E

Não convergência dos parâmetros AF, B1, B2 e BNZ

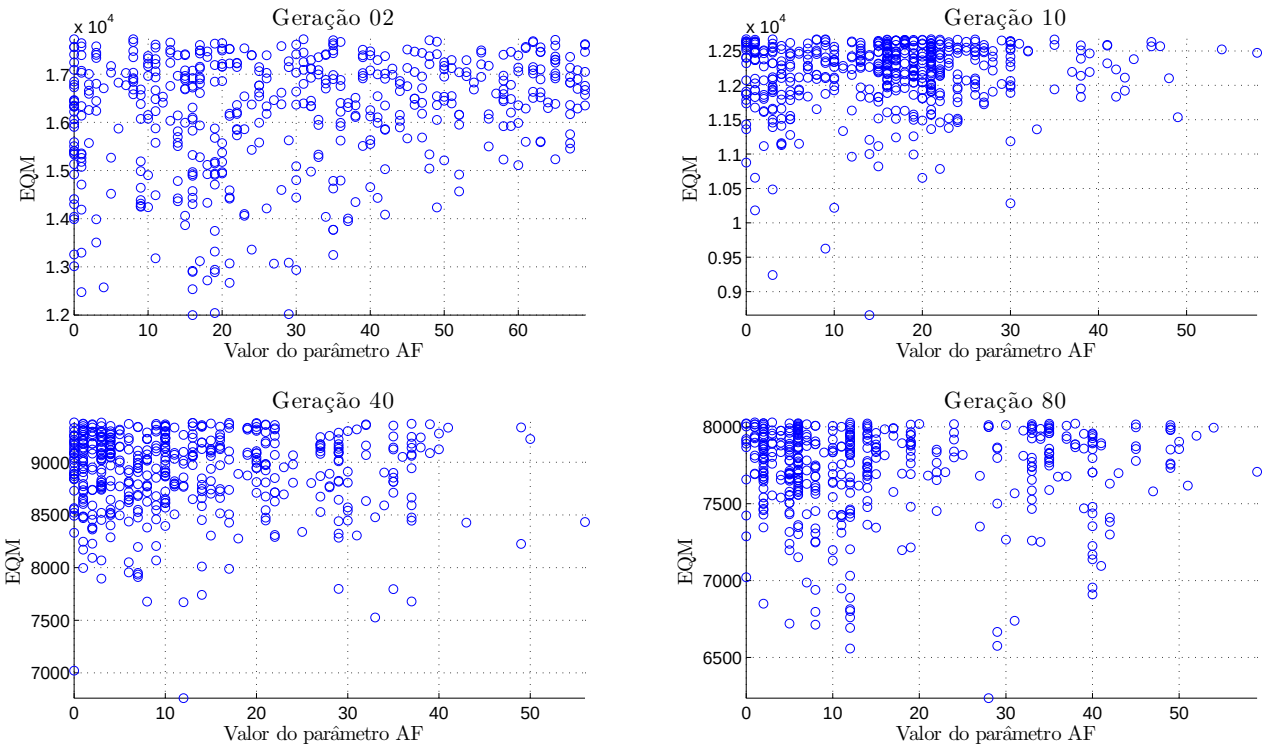

Figura E.1: Valores do parâmetro AF nas gerações 2, 10, 40 e 80.

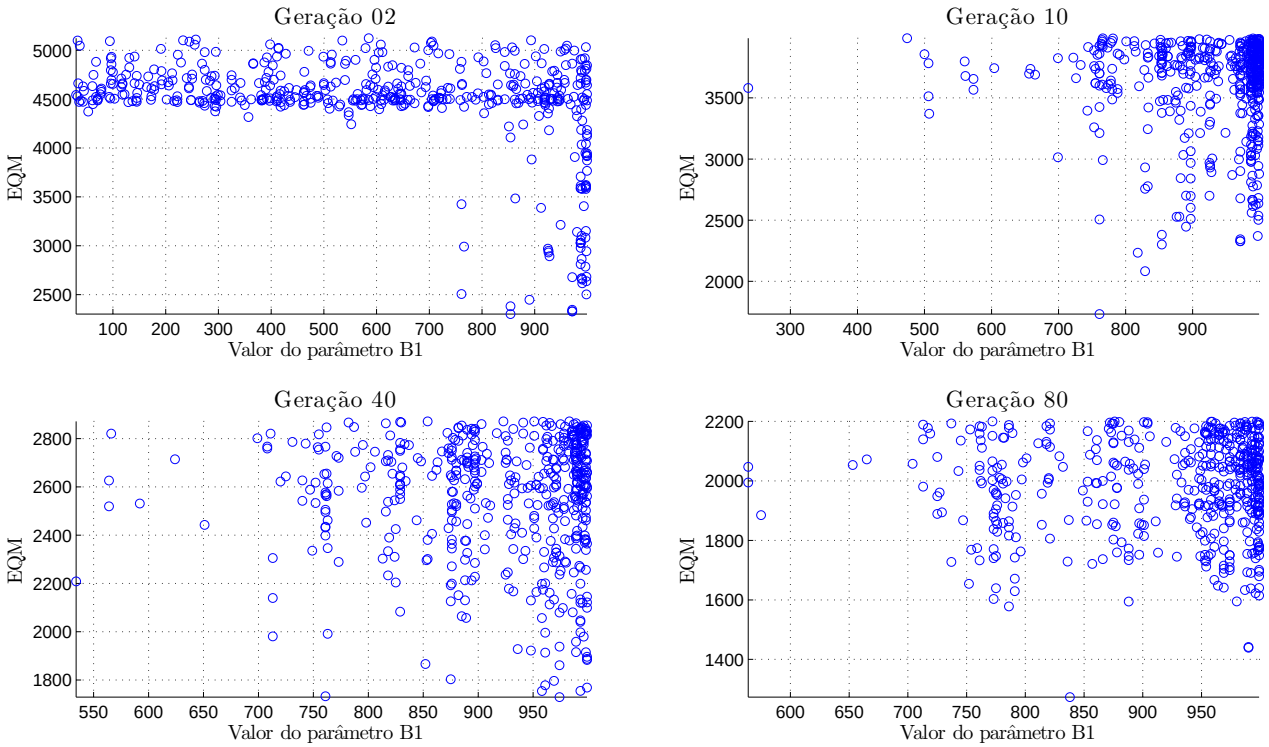

Figura E.2: Valores do parâmetro B1 nas gerações 2, 10, 40 e 80.

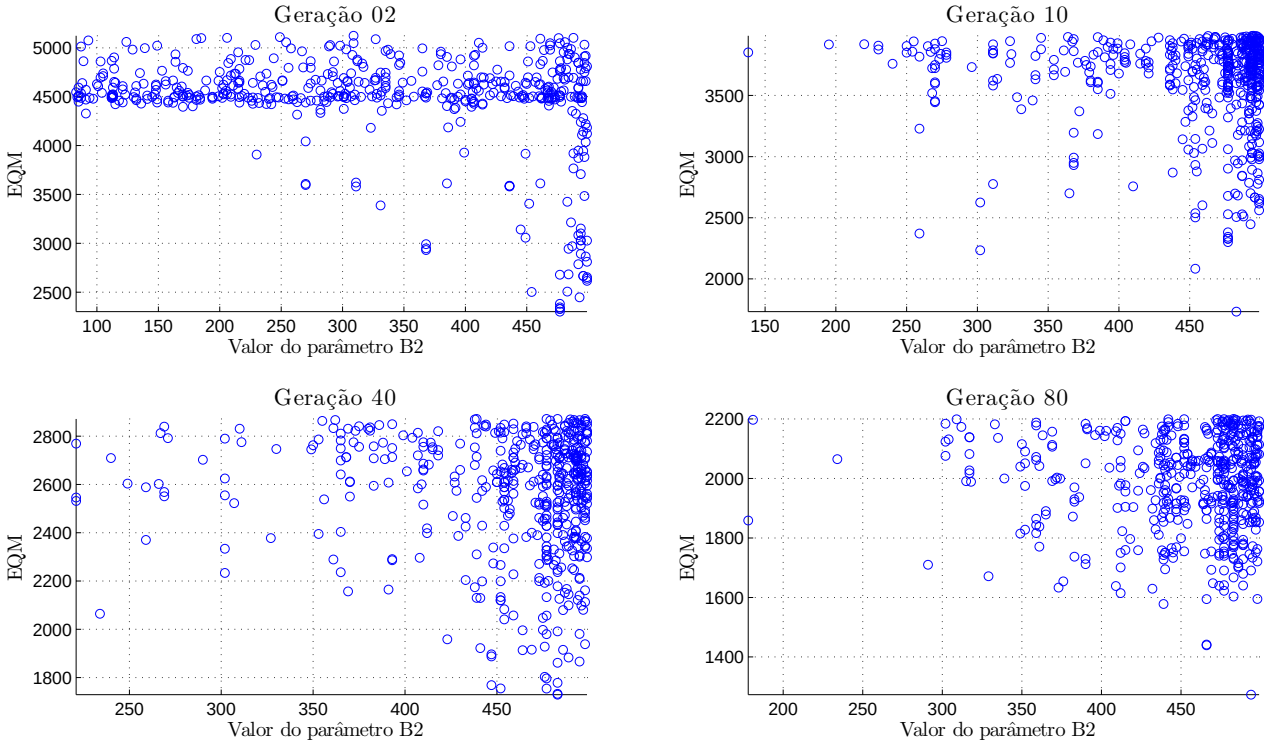

Figura E.3: Valores do parâmetro B2 nas gerações 2, 10, 40 e 80.

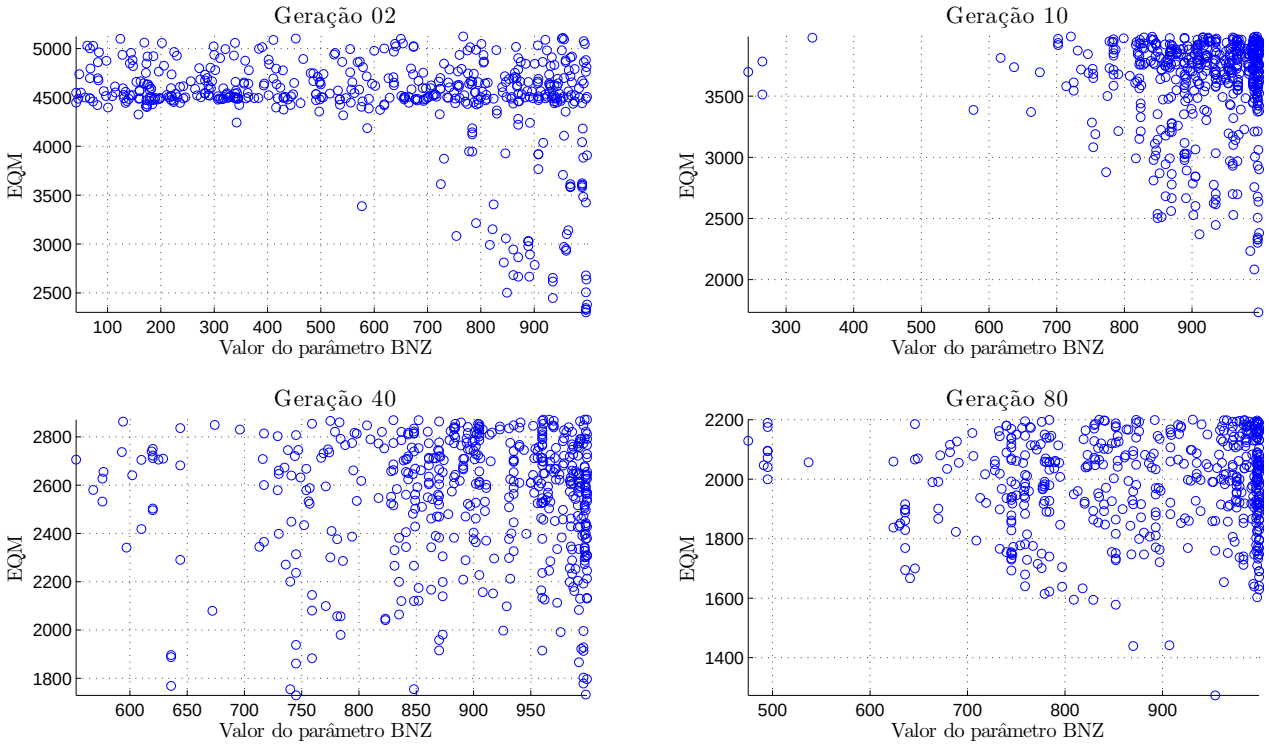

Figura E.4: Valores do parâmetro BNZ nas gerações 2, 10, 40 e 80.