Harlei Miguel de Arruda Leite

Proposta de metodologia de avaliação de voz sintética com ênfase no ambiente educacional

Campinas
2014

Universidade Estadual de Campinas
Faculdade de Engenharia Elétrica e de Computação

Harlei Miguel de Arruda Leite

Proposta de metodologia de avaliação de voz sintética com ênfase no ambiente educacional

Tese de Mestrado apresentada à Faculdade de Engenharia Elétrica e de Computação como parte dos requisitos exigidos para a obtenção do título de Mestre em Engenharia Elétrica. Área de concentração: Telecomunicações e Telemática.

Orientador: Dalton Soares Arantes

Este exemplar corresponde à versão final da dissertação defendida pelo aluno Harlei Miguel de Arruda Leite, e orientada pelo Prof. Dr. Dalton Soares Arantes

---

Campinas
2014

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca da Área de Engenharia e Arquitetura
Rose Meire da Silva - CRB 8/5974

Leite, Harlei Miguel de Arruda, 1989-
L536p Proposta de metodologia de avaliação de voz sintética com ênfase no ambiente educacional / Harlei Miguel de Arruda Leite. – Campinas, SP : [s.n.], 2014.

Orientador: Dalton Soares Arantes.
Dissertação (mestrado) – Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação.

1. Sintese da voz. 2. Sistema de processamento de fala. 3. Ambiente educacional. 4. Voz. 5. Fala - Inteligibilidade. I. Arantes, Dalton Soares,1946-. II. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Methodology for evaluation of synthetic speech emphasizing the educational environment
Palavras-chave em inglês:
Voice synthesis
Speech processing system
Educational environment
Voice
Speech - Intelligibility
Área de concentração: Telecomunicações e Telemática
Titulação: Mestre em Engenharia Elétrica
Banca examinadora:
Dalton Soares Arantes [Orientador]
Cecilia Sosa Arias Peixoto
Yuzo Iano
Data de defesa: 09-06-2014
Programa de Pós-Graduação: Engenharia Elétrica

COMISSÃO JULGADORA - TESE DE MESTRADO

Candidato: Harlei Miguel de Arruda Leite

Data da Defesa: 9 de junho de 2014

Título da Tese: "Proposta de Metodologia de Avaliação de Voz Sintética com Ênfase no Ambiente Educacional"

Prof. Dr. Dalton Soares Arantes (Presidente): _____

Profa. Dra. Cecilia Sosa Arias Peixoto: _____

Prof. Dr. Yuzo Iano: _____

Resumo

A principal contribuição desta dissertação é a proposta de uma metodologia de avaliação de voz sintetizada. O método consiste em um conjunto de etapas que buscam auxiliar o avaliador nas etapas de planejamento, aplicação e análise dos dados coletados. O método foi originalmente desenvolvido para avaliar um conjunto de vozes sintetizadas para encontrar a voz que melhor se adapta a ambientes de educação a distância usando avatares. Também foram estudadas as relações entre inteligibilidade, compreensibilidade e naturalidade a fim conhecer os fatores a serem considerados para aprimorar os sintetizadores de fala. Esta dissertação também apresenta os principais métodos de avaliação encontrados na literatura e o princípio de funcionamento dos sistemas TTS (*Text-to-Speech*).

Palavras-chave: Sintese da voz. Sistema de Processamento de Fala. Ambiente Educacional. Voz. Fala - Inteligibilidade.

Abstract

This thesis proposes, as main contribution, a new synthesized voice evaluation methodology. The method consists of a set of steps that seek to assist the assessor in the stages of planning, implementation and analysis of data collected. The method was originally developed to evaluate a set of synthesized voices to find the voice that best fits the environments for distance education using avatars. Relations between intelligibility, comprehensibility and naturalness were studied in order to know the factors to be considered to enhance the speech synthesizers. This thesis also presents the main evaluation methods in the literature and how TTS (Text-to-Speech) systems work.

Key-words: Voice Synthesis. Speech Processing System. Educatoinal Environment. Voice. Speech - Intelligibility.

Índice

À minha avó Maria Tereza

Agradecimentos

Primeiramente agradeço a Deus pela minha vida.

Agradeço à minha avó Maria Tereza, aos meus tios Denise e Cássio, à minha mãe Alessandra e à minha irmã Isabella por todo o apoio dado durante a minha trajetória profissional e pessoal.

Também agradeço ao meu orientador, Prof. Dr. Dalton Soares Arantes, pela sua dedicação e orientação durante estes anos.

Agradeço à Profa. Dra. Cecilia Peixoto por toda ajuda prestada no decorrer do mestrado e ao Tiago Cinto pela troca de experiência proporcionada por nosso trabalho em conjunto.

Também agradeço à minha namorada Sarah Negreiros pelo amor, companheirismo e cumplicidade, e à Cláudia Lunardelli pela ajuda na gravação das falas.

Agradeços a todos os meus amigos, em especial aos do ComLab: Veruska, Fábio, Carlos, Cláudio e André.

Agradeço a todos os professores e funcionários da FEEC e do IC, em especial: Prof. Dr. Yuzo Iano, Prof. Dr. José Mário de Martino, Prof. Dr. Luíz Geraldo P. Meloni, Prof. Dr. Ivan Luiz M. Ricarte, Profa. Dra. Maria Cecilia C. Baranauskas, Noêmia Benatti, Alaide Ramos e Edson Sanches.

Um agradecimento especial a todos os amigos que tive a oportunidade de conhecer nesses últimos anos.

E por fim, à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pela concessão da bolsa de Mestrado. Também agradeço à FAPESP e Padtec S. A., pelo apoio material e financeiro ao ComLab.

"*Any sufficiently advanced technology is indistinguishable from magic.*"

Arthur C. Clarke

Lista de Figuras

Lista de Tabelas

Lista de Acrônimos e Notação

DRT	*Diagnostic Rhyme Test*
HMM	*Hidden Markov Model*
HTS	*H triple S*
MLSA	*Mel Log Spectrum Approximation*
MOS	*Mean Opinion Score*
MRT	*Modified Rhyme Test*
PSOLA	*Pitch Synchronous Overlap Add*
STT	*Speech-to-Text*
TD-PSOLA	*Time-Domain Pitch-Synchronous Overlap-Add*
TTS	*Text-to-Speech*
WER	*Word Error Rate*

$F0$	Frequência fundamental da voz
OQ	Quociente de abertura
$V0$	Grau de excitação da voz
$F1...Fn$	Frequência dos formantes
$A1...An$	Amplitude dos formantes
FN	Frequência de um ressnonador de baixa frequência
ALF	Intensidade da região de baixa frequência
AHF	Intensidade da região de alta frequência
$BW1...BWn$	*Bandwidth*

Capítulo 1

Introdução

1.1 Contextualização e Motivação

Em 1965, Gordon E. Moore escreveu o célebre artigo intitulado "*Cramming more components onto integrated circuits*" afirmando que a densidade de transistores nos circuitos integrados teriam um aumento de 60%, pelo mesmo custo, a cada período de 18 meses [9]. Esta afirmação se provou verdadeira com o tempo e o cenário tecnológico atual reflete isso. Hoje dispõe-se de tecnologias com grande capacidade de processamento e armazenamento, que nos dão inúmeras possibilidades que não tínhamos no passado, principalmente em relação a funcionalidades multimídia e interfaces homem-computador.

No passado, a interação com computadores se dava por meio de uma interface rudimentar baseada em texto, acessível somente a pessoas com conhecimento técnico. Nos dias atuais, as interfaces evoluíram, possibilitando o seu uso por um grande número de pessoas com pouco conhecimento técnico [10].

Esta evolução tornou viável a criação de aplicações antes inimagináveis, tais como *softwares* de editoração gráfica, aplicações para modelagem tridimensional, simulação de fenômenos da natureza, editoração de vídeos em alta definição, jogos tridimensionais, plataformas *online* de educação a distância, dentre outras atividades que exigem o uso de recursos computacionais.

Um meio de interface homem-máquina que ganhou destaque nas últimas décadas foi a síntese de fala. Por meio dela, a interface se tornou acessível não somente para pessoas leigas, mas também para deficientes visuais. Além do mais, a troca de informações feita de modo verbal em situações de interação homem-máquina é cerca de duas vezes mais eficiente do que qualquer outra forma de comunicação [11].

No ano de 2012, o Laboratório de Comunicações (ComLab), da Faculdade de Engenharia Elétrica e de Computação (FEEC) da Universidade Estadual de Campinas (UNICAMP) passou a pesquisar meios alternativos de ensino baseado no uso de tecnologia, em especial AVAs (Ambiente Virtual de Aprendizagem). Com a pesquisa, começou a ser desenvolvido um ambiente de ensino a distância que faz uso de avatares [12] [13] [14] [15] [16].

Conforme a plataforma evoluía, observou-se que integrar voz humana previamente gravada nos avatares não era a melhor solução por três motivos: (1) Para uma boa gravação é necessário um estúdio; (2) É difícil encontrar alguém disposto a fazer as gravações e (3) Uma vez gravada,

a modificação do roteiro de aula é trabalhosa, necessitando de regravação.

Por conta dessas dificuldades, chegou-se à conclusão que o uso de sintetizadores de voz é a melhor maneira de gerar áudios de fala para os avatares por conta de três fatores: (1) Simplicidade, pois basta entrar com o texto a ser falado; (2) Baixo custo, pois não é necessário um estúdio de gravação e (3) Facilidade de alterar o roteiro de aula.

Um problema recorrente durante o desenvolvimento da plataforma de ensino foi encontrar a voz sintética ideal para os avatares, de forma a tornar as aulas atrativas e inteligíveis. Apesar de existir uma grande quantidade de sintetizadores, uma grande parte destes não apresentam um bom nível de inteligibilidade e naturalidade [1].

Um dos motivos que levou ao desenvolvimento deste trabalho foi ver que não existem muitos trabalhos que apresentem uma metodologia formal de avaliação de voz. A grande maioria dos trabalhos se restringe a avaliar características de um determinado sintetizador ou apresentar um determinado método de avaliação.

Um segundo motivo que levou ao desenvolvimento deste trabalhou foi compreender a relação entre inteligibilidade, compreensibilidade e naturalidade e o impacto que o uso de voz sintetizada traz em ambientes virtuais de educação a distância.

1.2 Objetivos do Trabalho

Este trabalho surgiu da necessidade de encontrar uma voz sintetizada de alta qualidade para a plataforma de ensino a distância usando avatares com vozes sintetizadas. O processo de encontrar uma voz sintetizada adequada envolve a aplicação de diversos métodos de avaliação de voz. Desta forma, este trabalho tem por objetivo apresentar uma metodologia de avaliação de voz para encontrar a voz ideal para a plataforma de ensino desenvolvida, e que também seja flexível o suficiente para que seja aplicada em outros projetos, não necessariamente na área de educação.

Adicionalmente, foi realizado em paralelo um estudo das relações entre inteligibilidade, compreensibilidade e naturalidade. O que levou ao estudo dessas relações foi o trabalho de Chang [8], que supõe que existem fatores secundários que influenciam na compreensibilidade, além da inteligibilidade. Desta forma, o estudo buscou encontrar a influência da naturalidade na inteligibilidade e na compreensibilidade. Os resultados deste estudo podem servir no desenvolvimento de novas vozes artificiais.

Outro objetivo deste trabalho foi apresentar uma proposta de arquitetura para a metodologia de avaliação desenvolvida. A arquitetura é importante pois, uma vez que seja implementada, o processo de desenvolvimento dos formulários, a aquisição dos dados e a avaliação se tornam triviais, diminuindo o tempo gasto do processo de encontrar a voz ideal.

Finalmente, objetivos menores integram este trabalho, tais como apresentar as diversas metodologias de voz, o funcionamento das principais técnicas de geração de fala artificial, os principais problemas dos sintetizadores de voz atuais e a influência de vozes sintetizadas em ambientes de educação. Apesar de menores, todos eles foram fundamentais para a construção da metodologia de avaliação e para o processo de busca da voz ideal para a plataforma.

1.3 Estrutura da Dissertação

Este trabalho está organizado da seguinte maneira:

- No Capítulo 2, é apresentado o histórico dos sistemas TTS.
- No Capítulo 3, é apresentada a estrutura de um sistema TTS, assim como os principais métodos de geração de fala.
- No Capítulo 4, é apresentado os principais métodos de avaliação de voz.
- No Capítulo 5, é apresentada as relações entre inteligibilidade, compreensibilidade e naturalidade.
- No Capítulo 6, a metodologia de avaliação de voz é apresentada, assim como todo o processo de aplicação da metodologia para encontrar a voz ideal para a plataforma. Também é apresentada uma proposta de arquitetura de automatização para a metodologia.
- Finalmente, o Capítulo 7 contém as conclusões obtidas e trabalhos futuros.

Capítulo 2

Histórico dos Sistemas TTS

O processo de geração de fala artificial sempre foi visto com grande entusiasmo, tanto pelas crianças quanto pelos adultos. Os filmes de ficção científica exploram a síntese de voz a exaustão para criar vozes de robôs e humanoides. Nos dias atuais, o processo de geração de fala artificial é uma realidade, mas para chegar ao nível atual de qualidade, inúmeras pesquisas foram feitas nos últimos séculos. Neste capítulo, apresentamos o histórico dos sistemas TTS.

2.1 Sintetizadores Mecânicos

Os primeiros relatos sobre a tentativa de produzir fala artificial nos remetem ao ano de 1779 [2] [17] [1] em São Petersburgo com o professor Christian Kratzenstein desenvolvendo aparatos mecânicos para atuarem como ressonadores, como mostra a Figura 2.1. O ressonador foi modelado de forma a ser similar com o trato vocal humano e funciona de maneira semelhante a um instrumento.

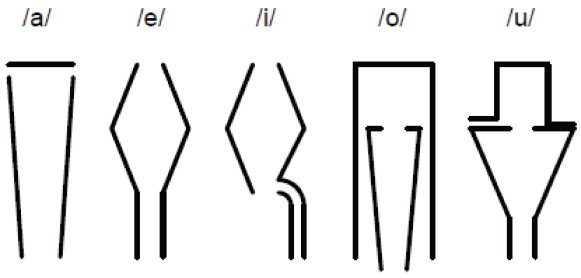


Figura 2.1: Ressonadores de Kratzenstein. Fonte: [1].

Alguns anos mais tarde, em 1791, Wolfgang von Kempelen apresentou em Vienna sua máquina mecânica de produção de fala artificial [4] [17]. A máquina era composta de uma câmera de pressão que simulava os pulmões, uma palheta vibratória para servir como cordas vocais e um tubo de couro para servir como trato vocal. A máquina podia reproduzir sons de vogais e consoantes por meio da manipulação do tubo de couro. Apesar de ter apresentado sua máquina em 1791, Kempelen iniciou seu projeto 20 anos antes, em 1769, antes mesmo de Kratzenstein.

Em 1800, Charles Wheatstone construiu uma versão aprimorada da máquina mecânica de produção de voz de Kempelen. A máquina é apresentada na Figura 2.2. Com os aprimoramentos, a máquina passou a ser capaz de pronunciar palavras completas, diferentemente da versão original que produzia somente vogais e algumas consoantes.

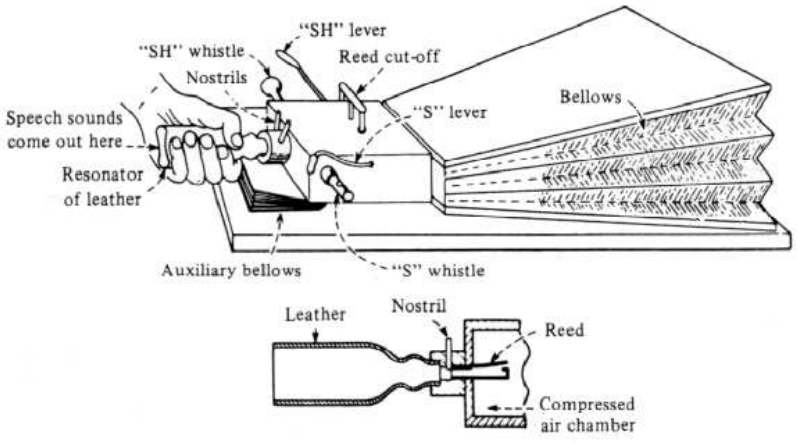


Figura 2.2: Reconstrução da Máquina Falante de Kempelen por Wheatstone. Fonte: [2].

Em 1838, Willis [18] encontrou a relação entre sons de vogais e a geometria correspondente do trato vocal. Em sua pesquisa, ele sintetizou diferentes vogais com tubos ressonadores semelhantes ao trato vocal. Durante o processo de pesquisa, ele descobriu que a qualidade dos sons das vogais depende somente do comprimento do tubo ressonador e não de seu diâmetro [17].

Os experimentos com aparatos mecânicos para a produção da fala continuaram até 1960. Apesar das várias tentativas de construir uma máquina falante, nenhum dos aparatos obteve sucesso. Na sua grande maioria, somente vogais podiam ser reproduzidas isoladamente e quando possível, a pronúncia de palavras era extremamente precária. Informações detalhadas sobre o funcionamento dos principais artefatos mecânicos para geração de síntese de fala podem ser encontradas em [17], [18], [2], [4] e [1].

2.2 Sintetizadores Elétricos

O primeiro sintetizador 100% elétrico foi apresentado por Stewart em 1922 [4]. O sintetizador tinha uma fonte de sinal sonoro como excitação e dois circuitos ressonadores para modelar a ressonância acústica do trato vocal. A máquina era capaz de reproduzir sons de vogais com baixa qualidade. Na mesma época, Wagner desenvolveu um dispositivo semelhante ao de Stewart. Seu dispositivo consistia de quatro ressonadores elétricos conectados em paralelo e uma fonte de sinal sonoro como excitação. As saídas dos quatro ressonadores são combinadas para gerar um espectro vocal. Diferentemente do modelo proposto por Stewart, o dispositivo de Wagner propiciava uma taxa de inteligibilidade aceitável [17].

O primeiro dispositivo considerado como um sintetizador de voz foi o VODER (*Voice Operating Demonstrator*) desenvolvido por Homer Dudley em 1939 [3]. O VODER foi baseado

no VOCODER (*Voice Coder*) desenvolvido pela Bell Laboratories em meados dos anos 30. A qualidade do VODER, apesar de sua baixa inteligibilidade, serviu de motivação para que outros sistemas de síntese de fala fossem construídos. Atualmente, os sintetizadores de voz ainda são similares em relação a sua arquitetura com o sistema VODER. A Figura 2.3 apresenta uma mulher operando o sistema VODER. A forma de funcionamento do dispositivo pode ser encontrado em [3].

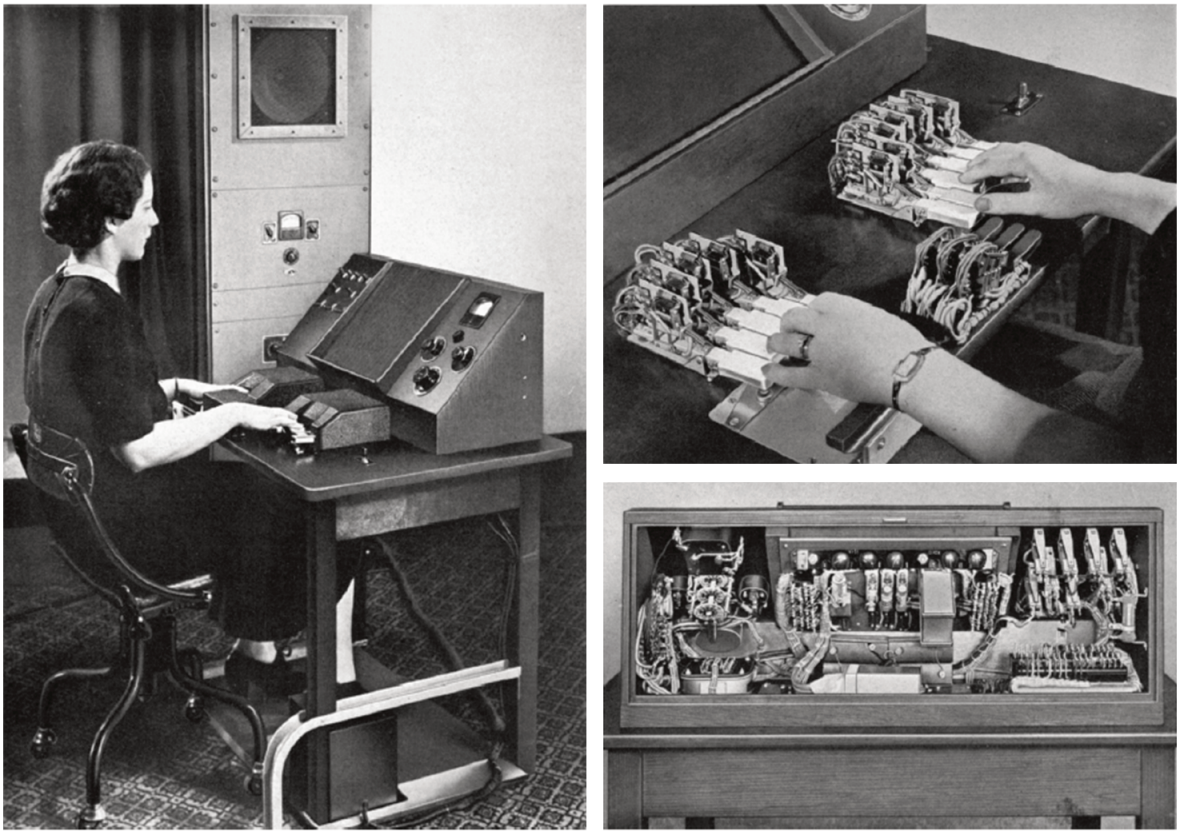


Figura 2.3: Demonstração do VODER. Fonte: [3].

Entre 1950 e 1960, diversas pesquisas na área de síntese de fala surgiram. Em 1953, o primeiro sintetizador por formantes, chamado PAT (*Parametric Artificial Talker*) foi desenvolvido por Walter Lawrence. Em 1958, o primeiro sintetizador articulatório foi desenvolvido por George Rosen no *Massachusetts Institute of Technology* (MIT). Em 1960, o primeiro sintetizador LPC (*Linear Predictive Coding*) foi construído.

O primeiro sintetizador completo para a língua inglesa foi desenvolvido por Noriko Umeda e sua equipe no *Electrotehnical Laboratory* no Japão [4] em 1968. O sintetizador era baseado no modelo articulatório e incluía um módulo de análise sintática com heurísticas sofisticadas. O sintetizador produzia fala com um nível de inteligibilidade aceitável para a época, porém monótono.

Em 1979 Allen, Hunnicutt e Klatt apresentaram o MITalk *laboratory text-to-speech system* desenvolvido no MIT [1]. O sistema teve boa aceitação e foi comercializado pela Telesensory Systems Inc. Em 1981, Klatt apresentou o sistema Klattalk que utiliza uma sofisticada fonte

de voz, descrito com maiores detalhes em [4]. As tecnologias usadas no MITalk e no Klattalk foram a base de muitos dos sintetizadores atuais.

Até a década de 70, as pesquisas com síntese de voz se restringiam aos laboratórios de pesquisa. A partir da década de 70, uma grande quantidade de sintetizadores de voz comerciais começou a aparecer, tais como a *Kurzwel Reading Machine*, *Votrax Type-N-Talk*, *Speech Plus Prose-2000*, *Digital DECTalk*, *INFOVOX* e *Street Elect Echo*. A Figura 2.4 mostra a evolução das tecnologias até chegar aos primeiros sintetizadores de voz comerciais.

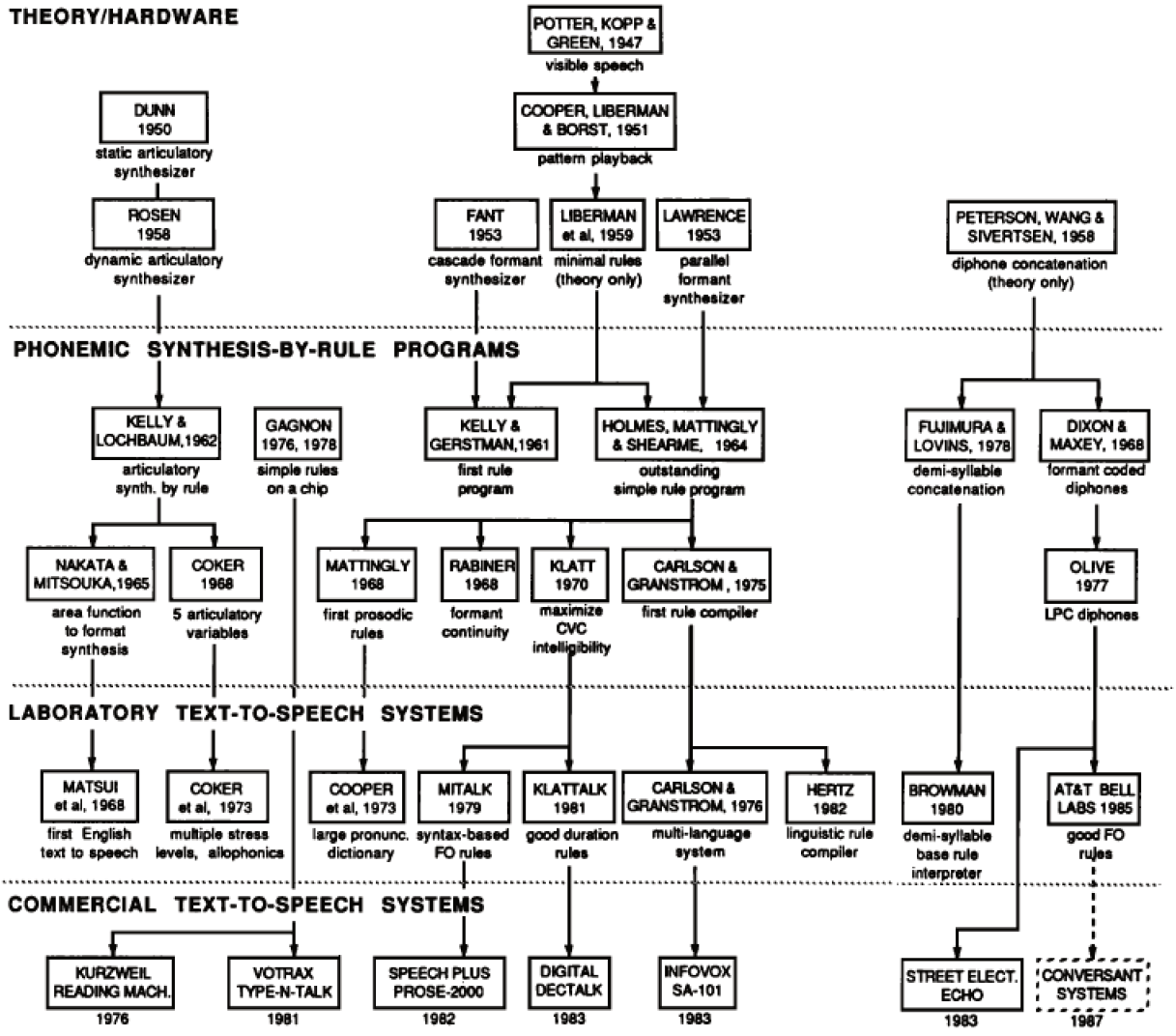


Figura 2.4: Evolução das Tecnologias de Síntese de Fala. Fonte: [4].

Em 1976, um leitor ótico integrado a um sistema TTS para auxiliar deficientes visuais foi apresentada por Kurzweil [4]. Possivelmente se trata da primeira aplicação integrada a um sistema TTS. A máquina leitora funcionava relativamente bem, mas por conta do alto preço na época, ela teve uma baixa aceitação por parte dos consumidores. No entanto, ela foi amplamente utilizada em bibliotecas para auxiliar deficientes visuais [1].

Atualmente existem inúmeros sintetizadores *open source* e comerciais. Os atuais sintetizadores fazem uso de sofisticados métodos e algoritmos e proporcionam um alto nível de inteligi-

bilidade. Com a maturidade dos sistemas TTS, estes passaram a integrar centrais telefônicas, caixas ATM, celulares, *tablets*, *videogames*, computadores e até mesmo brinquedos. Dentre eles, os mais populares são:

Ivona

O sintetizador da empresa Ivona (adquirida pela empresa Amazon) oferece um conjunto de 49 vozes em 22 línguas incluindo o Português do Brasil. O sintetizador se comunica por meio da API SAPI5 que permite a sua integração nas mais diversas aplicações. A empresa Ivona também oferece aplicações de acessibilidade, leitor de texto, soluções para telecomunicações, sistema STT, entre outras.

Loquendo

O sintetizador da empresa Loquendo (adquirida pela empresa Nuance) oferece um conjunto de 68 vozes em 18 línguas incluindo o Português do Brasil. O sintetizador se comunica por meio da API SAPI5 que permite a sua integração nas mais diversas aplicações. A empresa Loquendo também oferece soluções STT e aplicações que fazem uso de sintetizadores de voz, assim como a empresa Ivona.

Vocalize

O sintetizador da empresa Vocalize oferece um conjunto de 4 vozes em Português do Brasil. O sintetizador se comunica por meio de uma API proprietária da empresa. A empresa também oferece soluções STT e aplicações que fazem uso de sintetizadores de voz.

Microsoft Speech

O sistema operacional Windows por padrão possui integrado um sintetizador desenvolvido pela Microsoft. No entanto, a solução conta com somente uma voz em inglês. O sintetizador se comunica por meio da API SAPI5 que permite a sua integração nas mais diversas aplicações. Junto com o Windows, aplicações de acessibilidade fazem uso do sintetizador. O Windows também possuí por padrão um sistema STT para a língua portuguesa.

Festival

Festival é um sintetizador desenvolvido por Alan W. Black no *Centre for Speech Technology Research* (CSTR) na *University of Edinburgh*. O Festival oferece suporte para inúmeras línguas incluindo o Português do Brasil e métodos de geração de fala. Atualmente existem inúmeras API para a plataforma e é mantida por um grande número de colaboradores.

Cepstral

O sintetizador da empresa Cepstral oferece um conjunto de 27 vozes em 6 idiomas (não inclui o Português). O principal atrativo do sintetizador é a possibilidade de alterar características da voz, como a velocidade da pronúncia, a entonação e a possibilidade de inserir efeitos. A empresa também permite a geração de voz em 8-kHz para uso em telefonia. A empresa também oferece soluções prontas usando sistemas TTS.

2.3 Considerações Finais

Muito esforço foi empregado no desenvolvimento de sintetizadores de voz desde 1779 até os dias de hoje. Os sintetizadores de voz começaram como pequenos experimentos dentro de laboratórios de pesquisa até chegarem a produtos comerciais amplamente utilizados nas mais diversas áreas. Com o passar dos anos, diversos algoritmos e métodos de produção de fala artificial foram desenvolvidos, tornando a fala sintética inteligível e em alguns casos, com uma naturalidade próxima da fala humana. No capítulo seguinte, a estrutura de um sistema TTS é apresentada, assim como os principais métodos de produção de fala artificial.

Capítulo 3

Sistemas TTS

A fala é o principal meio de comunicação entre pessoas. Nas últimas décadas, surgiram diversos estudos visando o desenvolvimento de métodos e sistemas de geração de fala artificial [19] [20] [21]. Progressos recentes em síntese de fala permitiram o desenvolvimento de sintetizadores com alto nível de inteligibilidade, compreensibilidade e naturalidade. Neste capítulo é dada uma definição sobre sistemas TTS (*Text-to-Speech*) e os principais métodos de geração de fala são apresentados.

3.1 Estrutura de um Sistema TTS

Um sistema TTS recebe como entrada um texto escrito em linguagem natural e o sintetiza, gerando um sinal de fala correspondente ao texto de entrada. Um sistema TTS é dividido em dois estágios: *Front-End* e *Back-End*, conforme mostra a Figura 3.1.

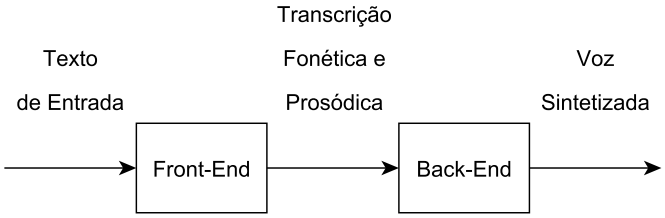


Figura 3.1: Arquitetura de um Sistema TTS.

Na etapa de *Front-End* é realizado o processamento do texto, que consiste em obter a transcrição fonética e a descrição prosódica. Entende-se por transcrição fonética uma transcrição dos sons da fala e por descrição prosódica as informações referentes a duração dos fonemas, ênfase no discurso, mudanças nos valores de *pitch*, entre outras características.

Na etapa de *Back-End* é realizado o processamento digital de sinais de fala, que consiste em utilizar métodos para produzir fala sintética. Existem diversos métodos, dentre eles a síntese articulatória, a síntese por formantes, a síntese concatenativa e a síntese baseada em modelos ocultos de Markov (HMM). Os métodos serão abordados com maiores detalhes no decorrer do capítulo.

3.2 *Front-End*

Na etapa de *Front-End* é realizado o processamento do texto. O objetivo desta etapa é obter a transcrição fonética e prosódica a partir da análise linguística do texto de entrada. A etapa de *Front-End* é fortemente dependente das características da língua com a qual se está trabalhando, de forma que ela deve ser projetada para analisar um único idioma.

Além dos dados relativos ao conteúdo semântico do texto, a etapa de *Front-End* deve fornecer informações prosódicas, como a duração dos fonemas, ênfases específicas no discurso, mudanças nos valores de *pitch*, entre outras. Para extrair essas características do texto, diversas análises devem ser feitas, tais como a análise lexical, gramatical, sintática e semântica.

A primeira tarefa da etapa de *Front-End* é processar o texto de entrada e transcrevê-lo de forma conveniente para a etapa de *Back-End*. Neste processo, todos os caracteres não alfabéticos e siglas são transcritos por extenso, na forma que devem ser pronunciados. Por exemplo, a frase "Eu estacionei o meu carro na Av. Albert Einstein nº 400 para uma reunião com o Prof. Dr. Dalton" seria transcrita como "Eu estacionei o meu carro na avenida Albert Einstein número quatrocentos para uma reunião com o professor doutor Dalton". A Tabela 3.1 apresenta alguns exemplos de transcrição.

Tabela 3.1: Transcrição de Símbolos e Siglas

Texto Original	Texto Transcrito
$20°C$	Vinte graus centígrados
05/04/1989	Cinco de abril de mil novecentos e oitenta e nove
18:30	Dezoito horas e trinta minutos
UNICAMP	Unicampi
R. Limoeiro nº 4	Rua Limoeiro número 4
R$ 1,99	Um real e noventa e nove centavos

Após o texto ser processado de maneira conveniente, as palavras são transcritas foneticamente, considerando a acentuação das palavras e sinais de pontuação, além das pausas necessárias na leitura, para que o computador saiba como pronunciar cada fonema considerando as características prosódicas do texto [22].

Uma tarefa importante da etapa de *Front-End* é tratar os problemas inerentes a língua como a ambiguidade de homógrafos. Homógrafos são palavras que possuem a mesma grafia, mas que possuem uma pronúncia e significado diferentes, dependendo do contexto. Um exemplo é a frase: "Eu gosto de você" e "Meu gosto é diferente"; a palavra gosto possui a mesma grafia, mas possuem pronúncia e significados diferentes. A Tabela 3.2 apresenta algumas expressões com homógrafos.

Tabela 3.2: Palavras Homógrafas

Contexto 1	Contexto 2
Eu almoço ao meio-dia	Hoje tem carne no almoço
Eu troco a nota em moedas	Ainda não recebi o troco
Estou com muita sede	Vou até a sede do clube

Um dos maiores desafios da etapa de *Front-End* é identificar a pronúncia correta a um mesmo símbolo fonético [1]. Na língua portuguesa, a letra "x" corresponde a diversos fonemas em cada uma das seguintes palavras: "xícara" - som de ch, "máximo" - som de ss, "táxi" - som de ks e "exemplo" - som de z. Todas as particularidades de cada língua devem ser tratadas para que não ocorram erros de pronúncia na fala artificial.

Para que um sistema TTS seja autônomo, é necessário que toda a etapa de *Front-End* seja realizada de maneira automática, robusta e rápida, com um mínimo de interferência humana. Atualmente, a língua inglesa conta com um módulo *Front-End open source* tendo como base o sistema Festival. No entanto, ainda não existe um módulo *Front-End open source* para a língua portuguesa.

3.3 *Back-End*

A fala sintética pode ser produzida por vários métodos diferentes. Os métodos são usualmente classificados em quatro grupos: (1) Síntese Articulatória; (2) Síntese por Formantes; (3) Síntese Concatenativa e (4) Síntese HMM (*Hidden Markov Model*). A escolha do método de síntese de voz ideal depende exclusivamente dos requisitos de plataforma (*hardware*) e características da voz desejada. A síntese articulatória ainda não é utilizada em sintetizadores comerciais por ser um método de difícil implementação. A síntese por formantes foi amplamente utilizada nas décadas passadas, mas atualmente vem sendo substituída pela síntese concatenativa, que utiliza base de dados de voz humana e permite a geração de fala com alto nível de naturalidade. A síntese HMM atualmente tem sido amplamente estudada e já vem sendo implementada em sistemas comerciais. Cada método tem suas vantagens e desvantagens que serão discutidas nas próximas seções.

3.3.1 Síntese Articulatória

A síntese articulatória procura modelar os órgãos vocais o mais próximo da realidade para produzir a fala sintética, tornando-o o método mais complicado de se implementar devido a complexidade dos órgãos vocais [1] [23] [24]. Apesar da síntese articulatória não produzir boa qualidade de fala, o seu estudo é fundamental para compreender o processo de produção de fala.

A implementação do método envolve a modelagem de articuladores humanos e cordas vocais. Os articuladores são usualmente modelados por meio de um conjunto de funções entre a glote e a boca. O primeiro modelo articulatório baseou-se em uma tabela de funções do trato vocal da laringe até os lábios para cada segmento fonético [4]. Os parâmetros dos articuladores podem ser abertura dos lábios, movimento dos lábios, altura da ponta da língua, posição da ponta da língua, altura da língua, posição da língua e a abertura entre a laringe e a passagem nasal. Os parâmetros de fonação ou de excitação podem ser a abertura da glote, tensão das cordas vocais e pressão pulmonar [23]. Os parâmetros articulatórios e de fonação são usualmente obtidos a partir de dados de observações do trato vocal durante a fala.

Devido a sua complexidade, não se tem conhecimento de sistemas comerciais que empregam esta técnica, no entanto, existem diversos estudos nesta área, podendo-se citar os trabalhos de Engwall [25], Mullen et al. [26], Aryal e Gutierrez-Osuna [27] e Palo [24].

3.3.2 Síntese por Formantes

A síntese por formantes utiliza um conjunto de regras que determinam os parâmetros necessários para sintetizar uma dada expressão. Em geral, a síntese por formantes é implementada seguindo uma estrutura sequencial ou paralela. No entanto, pode-se encontrar implementações mistas buscando melhorar a qualidade do sintetizador. A síntese por formantes não utiliza uma base de dados de voz humana, permitindo uma maior flexibilidade na geração dos mais variados tipos de sons [1].

Pelo menos três formantes são necessários para produzir fala inteligível e até cinco formantes são necessários para produzir fala de alta qualidade. Cada formante é usualmente modelado com um ressonador de dois pólos, que permite especificar a frequência do formante e a largura de banda [28].

Os parâmetros utilizados pela síntese de formantes são [29] [30]:

- Frequência fundamental da voz ($F0$)
- Quociente de abertura (OQ)
- Grau de excitação da voz ($V0$)
- Frequências dos formantes e amplitude ($F1...F3$ e $A1...A3$)
- Frequência de um ressonador de baixa frequência adicional (FN)
- Intensidade da região de baixa e alta frequência (ALF, AHF)

Um sintetizador por formantes sequencial consiste de um conjunto de formantes ressonadores passa-banda conectado em série, de forma que a saída de cada formante ressonador é aplicado na entrada do próximo formante ressonador. A estrutura cascata precisa somente das frequências dos formantes como informação de controle. A principal vantagem da estrutura sequencial é que as amplitudes dos formantes para cada vogal não precisam de controles individuais [29]. A Figura 3.2 mostra um sintetizador por formantes sequencial.

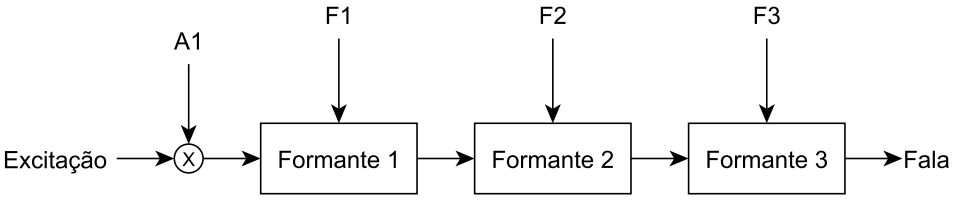


Figura 3.2: Estrutura de um Sintetizador por Formantes Sequencial. Fonte: Adaptado de [1].

Um sintetizador por formantes paralelo consiste de um conjunto de formantes ressonadores conectados em paralelo. O sinal de excitação é aplicado em todos os ressonadores simultaneamente e suas saídas são somadas. A estrutura paralela permite controlar a largura de banda e o ganho para cada formante ressonador individualmente, diferente da estrutura paralela. A

principal vantagem da estrutura em paralelo é a qualidade da modelagem de consoantes nasais, fricativas e oclusivas. No entanto, algumas vogais são melhores modeladas por meio da estrutura sequencial. A Figura 3.3 mostra um sintetizador por formantes paralelo [1].

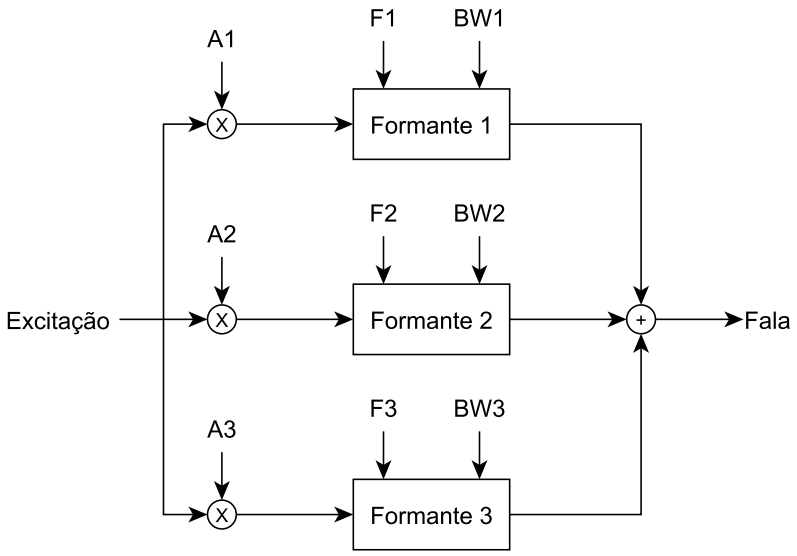


Figura 3.3: Estrutura de um Sintetizador por Formantes Paralelo. Fonte: Adaptado de [1].

Implementações passadas mostram que a estrutura sequencial e paralela não produzem bons resultados, o que motivou a criação de uma estrutura híbrida, que incorpora a estrutura sequencial e paralela, como a proposta por Klatt [31] em 1980. A qualidade do modelo híbrido proposto por Klatt foi promissor na época, e foi incorporado em diversos sintetizadores, tais como MITalk, DECtalk, Prose-2000 e Klattalk [28]. Além do modelo híbrido proposto por Klatt, diversos outros modelos híbridos foram implementados, podendo-se citar a implementação de Laine [32] incorporado no sintetizador SYNTE3 para a língua finlandesa [1].

3.3.3 Síntese Concatenativa

A síntese concatenativa é hoje o método mais utilizado por sintetizadores comerciais. O princípio da síntese concatenativa é a concatenação de segmentos de fala humana pré-gravada, sendo a qualidade do sintetizador diretamente relacionada à extensão e qualidade da base de dados de áudio. Por utilizar uma base de dados de voz real, a síntese concatenativa é a forma mais simples de gerar fala com alto nível de inteligibilidade e naturalidade.

Um dos principais aspectos da síntese concatenativa é encontrar o tamanho ideal de cada segmento de fala. Quanto maior o segmento de fala, maior é o espaço em memória exigido e melhor é a naturalidade da fala, pois menos pontos de concatenação são necessários. Por outro lado, quanto menor o segmento de fala, menor é o espaço em memória exigido e pior é a naturalidade da fala, pois mais pontos de concatenação são necessários.

A concatenação de palavras é a forma mais simples de formar sentenças. No entanto, gravar palavras inteiras só é viável em situações onde o vocabulário é pequeno e as sentenças a serem

formadas sejam previamente conhecidas. Além do mais, a pronúncia de uma palavra dita isoladamente pode não ser a mesma dentro de uma sentença. Por conta dessas dificuldades, além do espaço em memória requerido para armazenar vocabulários complexos, a concatenação de palavras não é adequada para sistemas onde o texto de entrada seja irrestrito [29].

Os primeiros sintetizadores faziam a concatenação de fonemas. No entanto, na segmentação dos fonemas ocorre a perda das informações de coarticulação. Sem as informações de coarticulação não é possível fazer a transição contínua entre os fonemas, comprometendo a inteligibilidade da fala [33].

Atualmente, a maior parte dos sintetizadores aplica a concatenação de polifones, difones e trifones para evitar o problema de transição entre os fonemas. Desta forma, as transições entre os fones são preservadas por meio da concatenação feita entre sinais com conteúdo espectral semelhante, obtendo-se assim um sinal de fala inteligível e natural, podendo-se melhorar ainda mais a transição entre os fones com o uso de algoritmos de concatenação capazes de modificar a envoltória espectral do sinal de fala, suavizando as descontinuidades.

Uma forma de melhorar a síntese concatenativa é usando a técnica TD-PSOLA (*Time-Domain Pitch-Synchronous Overlap-Add*) [34] [35]. O método é amplamente utilizado devido a sua eficiência computacional e por melhorar a qualidade da síntese concatenativa [36]. A técnica TD-PSOLA é uma variante do algoritmo PSOLA (*Pitch Synchronous Overlap Add*), que foi originalmente desenvolvido pela France Telecom (CNET) [1].

O método TD-PSOLA permite alterar a duração da fala e da frequência dos *pitches* dos fones, buscando satisfazer os requisitos das palavras que conterão aqueles fones. O método TD-PSOLA é dividido em 3 etapas [37] [38]: Na primeira etapa ocorre a marcação dos *pitches*; na segunda etapa ocorre a modificação do sinal, cancelando ou replicando janelas e na terceira etapa ocorre a concatenação dos sinais pela superposiçao de segmentos janelados, utilizando janelas de *Hanning*. A modificação da duração da fala é feita cancelando ou replicando algumas das janelas e a modificação da frequência dos *pitches* é feita aumentando ou diminuindo a superposição entre os segmentos janelados. A Figura 3.4 mostra o incremento e decremento do *pitch*.

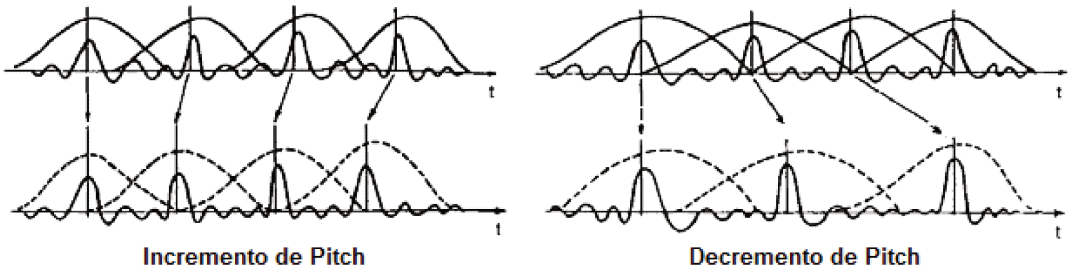


Figura 3.4: Modificação do *Pitch*. Fonte: Adaptado de [1].

Utilizando o método TD-PSOLA podemos mudar a prosódia da fala, de forma a expressar emoções, tais como raiva, tristeza, medo, felicidade, desgosto, dentre outras. A possibilidade de alteração da prosódia permite a criação de discursos mais efetivos, atraindo a atenção dos ouvintes e facilitando a comunicação.

3.3.4 Síntese HMM

A síntese HMM foi desenvolvida para superar as limitações da síntese concatenativa, tais como a necessidade de uma extensa base de dados de segmentos de fala humana pré-gravada e a dificuldade de inserção de emoção no discurso [22]. Ela foi proposta por um grupo de pesquisadores do Instituto de Nagoya e do Instituto de Tecnologia de Tóquio, coordenados por Keiichi Tokuda [39]. O sistema desenvolvido por eles é denominado HTS do acrônimo "*H triple S*" em referência ao nome *HMM-Based Speech Synthesis System*.

A estrutura de um sintetizador HMM é dividida em duas fases: treinamento e síntese. Na etapa de treinamento, os modelos HMM são construídos. A etapa de síntese é dividida em 4 fases: Na fase 1 o texto a ser sintetizado é transcrito foneticamente. Na fase 2 são definidas as durações dos estados para a sequência HMM. Na fase 3 obtêm-se os modelos espectrais e de excitação para cada estado de cada fonema do texto e na fase 4 a onda é sintetizada a partir dos parâmetros espectrais e de excitação usando o filtro MLSA (*Mel Log Spectrum Approximation*) [22]. A Figura 3.5 ilustra o processo.

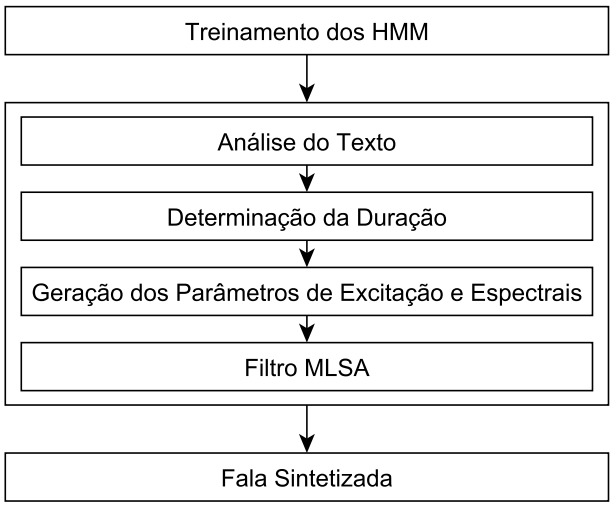


Figura 3.5: Estrutura de um Sintetizador HMM.

A síntese HMM necessita de uma base de dados de fala humana relativamente pequena para treinamento dos modelos HMM quando comparada com o exigido pela síntese concatenativa. Por ser um método de natureza estática e paramétrica, é possível modificar as características da voz, adaptar a prosódia e inserir emoções no discurso com facilidade. As técnicas mais utilizadas para efetuar essas modificações na voz são a de adaptação [40], *eigenvoices* [41], interpolação [42] [43] e regressão múltipla [44].

A principal desvantagem da síntese HMM é a falta de naturalidade da voz sintética causada pela perda de variabilidade dos parâmetros estáticos dos fones durante a fase de treinamento. Este excesso de suavização dá origem a uma fala sintética abafada e monótona [22]. Para obter uma fala mais natural e expressiva, diversos métodos foram propostos, tais como a pós-filtragem [45], Straight [46] e a utilização de parâmetros dinâmicos e da variância global [47] [48].

3.4 Considerações Finais

Dentre todos os métodos, a síntese concatenativa é a mais utilizada para desenvolver sistemas TTS comerciais. Por ser um método que concatena segmentos de fala humana, ela proporciona um alto nível de inteligibilidade, naturalidade e compreensibilidade. Por outro lado, a síntese HMM tem tido destaque no ambiente acadêmico, devido a sua estabilidade e flexibilidade em modificar as características da voz artificial e por necessitar de uma base de fala limitada (na ordem de uma a duas horas de gravação) comparado com a síntese concatenativa, que exige uma extensa base de fala. A síntese articulatória tem recebido uma ponta de destaque em alguns grupos específicos de pesquisa, que buscam compreender o processo de produção de fala e o *enhancement* com base na modelagem realista do aparelho fonador humano. Para a síntese por formantes, não foi encontrado nenhum trabalho relevante nos últimos anos.

Capítulo 4

Métodos de Avaliação de Voz

O processo de avaliação de voz sintetizada consiste em avaliar os níveis de inteligibilidade, compreensibilidade, adequabilidade e características subjetivas, como a naturalidade [4] [49] [19] [8]. O nível adequado para cada uma dessas características depende do tipo de aplicação. Em uma aplicação militar, ou um leitor para cegos, o nível de inteligibilidade e compreensibilidade tem uma maior relevância comparado com o nível de naturalidade. O nível de adequabilidade depende exclusivamente das tecnologias usadas. Um sistema TTS pode não fornecer suporte de integração para uma determinada linguagem de programação, ou pode até mesmo requerer uma plataforma de *hardware/software* específico para execução.

Atualmente, existem inúmeros desafios no processo de avaliação de voz sintetizada [6] [49] [1]. Os métodos de avaliação de voz são, na maioria das vezes, projetados para avaliar voz humana, e nem sempre abordam todos os requisitos de avaliação de voz sintetizada, pois o processo de avaliação de voz sintetizada não deve levar em consideração somente as características acústicas, mas sim toda a etapa de *Front-End*.

O procedimento de avaliação é feito com um conjunto de ouvintes, em uma sala previamente preparada, de preferência com isolamento acústico, de maneira individual. Para maior controle, todos os ouvintes devem usar fone de ouvido de alta qualidade, com um volume igual para todos, não podendo ser aumentado nem diminuído. Os testes devem ser feitos no menor tempo possível, para evitar distrações. Em alguns casos, ouvintes profissionais podem ser necessários, principalmente em situações onde o sintetizador de voz assume função crítica. Não deve ter no campo de visão do ouvinte qualquer tipo de artefato que possa distraí-lo.

O processo de avaliação deve ser feito somente uma única vez com cada ouvinte. Quando realizada mais de uma vez, pode ocorrer o efeito de aprendizagem, que se soma aos resultados da avaliação. Por outro lado, problemas de concentração podem decrementar os resultados. Por conta dessas influências, o uso de ouvintes profissionais é recomendado para casos onde o resultado deve ter um alto nível de confiabilidade.

Uma segunda forma de avaliar voz sintetizada (e voz humana) é por meio de sistemas STT (*Speech-to-Text*). Neste caso, o papel do ouvinte é feito pelo reconhecedor de voz. Apesar de ser uma boa alternativa do ponto de vista logístico, por não precisar selecionar uma população e evitar questões ligadas ao código de ética, os resultados podem sofrer um decremento por conta de falhas no sistema de reconhecimento. Além do mais, não é possível avaliar a naturalidade e

a compreensibilidade por meio de um sistema STT. Atualmente, avaliação automática usando sistemas STT é usada de forma conjunta com a avaliação por ouvintes.

Neste capítulo, os principais métodos de avaliação de voz são apresentados. Alguns métodos necessitam de materiais de base, tais como frases semanticamente confusas (*SUS - Semantically Unpredictable Sentences*) e tabelas de palavras. No Apêndice, um conjunto de materiais de base é disponibilizado para a língua portuguesa.

4.1 Métodos de Avaliação Segmental

Uma forma de avaliar a inteligibilidade de uma voz é por meio de métodos de avaliação segmental [6] [50]. Os métodos de avaliação segmental buscam mensurar a inteligibilidade de um único fonema ou segmento. Os métodos de avaliação segmental mais conhecidos são chamados de *Diagnostic Rhyme Test* (DRT) e *Modified Rhyme Test* (MRT). O método DRT avalia a inteligibilidade da consoante inicial e o método MRT avalia a consoante inicial e final. Ambos os métodos podem ser aplicados usando ouvintes não profissionais. Nas subsessões seguintes, ambos os métodos são apresentados com maior profundidade.

4.1.1 *Diagnostic Rhyme Test* (DRT)

O método *Diagnostic Rhyme Test* (DRT) foi apresentado por Fairbanks em 1958 e consiste em usar um conjunto de palavras isoladas para avaliar a inteligibilidade da consoante inicial [51] [50]. O método consiste em 96 pares de palavras de duas sílabas que diferem unicamente na consoante inicial. A Tabela 4.1 apresenta alguns exemplos de pares de palavras:

Tabela 4.1: Conjunto de Palavras para o Método DRT

	A	**B**
1	Pato	Gato
2	Ouro	Touro
3	Moda	Roda
4	Tato	Fato

O processo de aplicação do método consiste em reproduzir para o ouvinte uma palavra por vez, que por sua vez deve marcar na folha de respostas a palavra que julga ter ouvido dentre as duas opções do grupo de palavras. Para cada ouvinte obtemos uma taxa de acerto, que é a relação entre o número de palavras identificadas corretamente pelo número total de palavras. O nível de inteligibilidade da consoante inicial é dado pela média geral de todos os ouvintes [1].

4.1.2 *Modified Rhyme Test* (MRT)

O método *Modified Rhyme Test* (MRT) é uma extensão do método DRT e consiste em usar um conjunto de palavras isoladas para avaliar a inteligibilidade da consoante inicial e da consoante final [51] [50]. O método consiste em 50 conjuntos de 6 palavras de duas sílabas que totalizam 300 palavras, sendo 25 conjuntos para avaliar a consoante inicial e 25 conjuntos para avaliar a consoante final. A Tabela 4.1 apresenta alguns exemplos de conjuntos de palavras.

Tabela 4.2: Conjunto de Palavras para o Método MRT

		A	B	C	D	E	F
Consoante Inicial	1	Pato	Gato	Rato	Tato	Fato	Mato
	2	Foca	Arca	Banca	Barca	Bica	Boca
	3	Cofre	Abre	Corre	Bagre	Chifre	Cobre
	4	Baixa	Buxa	Taxa	Coxa	Fixa	Flexa
Consoante Final	1	Dança	Dama	Dado	Data	Dano	Danar
	2	Manga	Mansão	Manso	Mapa	Marco	Março
	3	Quintal	Quinto	Quinze	Quite	Quina	Quilo
	4	Tela	Tecla	Teimar	Tédio	Tema	Temer

O processo de aplicação do método consiste em reproduzir para o ouvinte uma palavra por vez, que por sua vez deve marcar na folha de respostas a palavra que julga ter ouvido dentre as seis opções do grupo de palavras. Para cada ouvinte obtemos uma taxa de acerto da consoante inicial e final, além da média geral. O nível de inteligibilidade da consoante inicial e final é dado pela média geral de todos os ouvintes [1].

Para fins de exemplificação, a Figura 4.1 apresenta os resultados obtidos por Logan [50] utilizando o método MRT para avaliar a inteligibilidade de dez sintetizadores e uma voz natural. Os sintetizadores utilizados foram: *DECtalk Paul*, *DECtalk Betty*, *MITalk-79*, *Prose 3.0*, *Amiga*, *Infovox SA 101*, *TSI-Proto I*, *Smoothtalker*, *Votrax Type n Talk* e *Echo*.

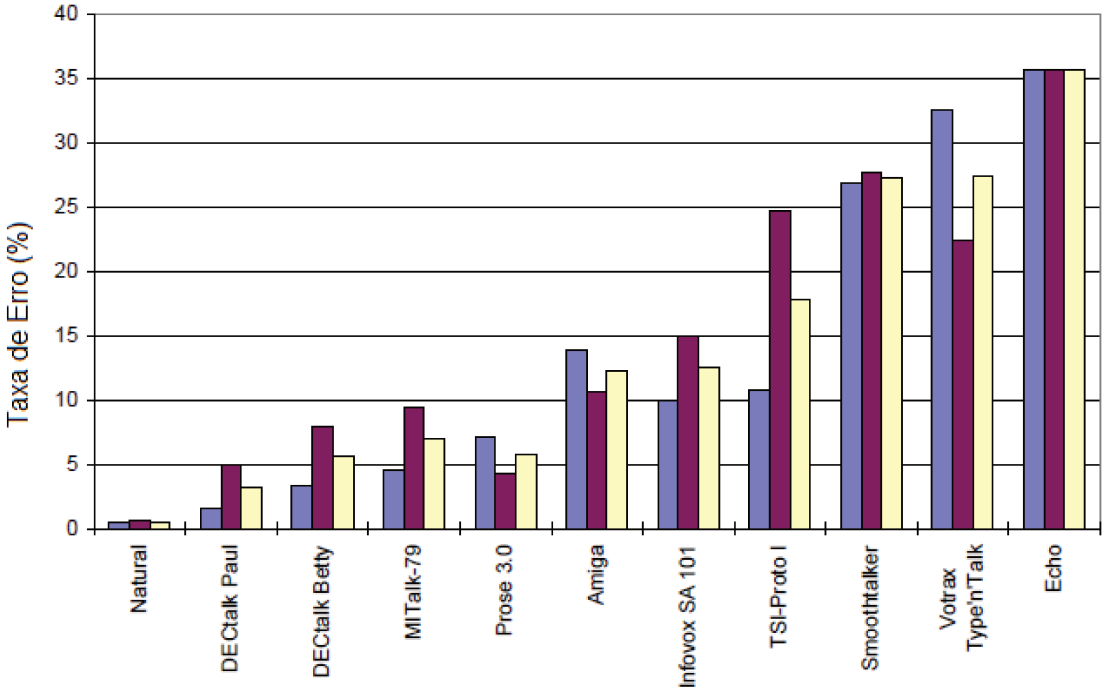


Figura 4.1: Resultado de Avaliação feita por Logan usando MRT. Fonte: Adaptado de [1].

Na Figura 4.1, a barra em azul corresponde a taxa de acerto da consoante inicial, a barra roxa corresponde a média entre a consoante inicial e final e a barra em amarelo corresponde a taxa

de acerto da consoante final. Ainda segundo Logan [50], quando o método é aplicado sem uma folha de alternativas, ou seja, respostas abertas, a taxa de acerto decrementa significativamente.

4.2 Métodos de Avaliação de Sentenças

Os métodos de avaliação de sentenças permitem avaliar o nível de inteligibilidade e compreensibilidade. Para avaliar o nível de inteligibilidade, sentenças com propriedades distintas são utilizadas. Atualmente existem diversas bases de dados de sentenças para avaliação de voz, dentre elas podemos citar: *Harvard Psychoacoustic Sentences*, *Haskins Sentences* e *Semmantically Unpredictable Sentences* (SUS). Na avaliação da compreensibilidade, um conjunto de notícias é sintetizado e questões referentes a notícia são feitas; a taxa de compreensibilidade é a taxa de acerto das perguntas. Nas subsessões seguintes, os métodos são apresentados com maior profundidade.

4.2.1 *Harvard Psychoacoustic Sentences*

A base de dados *Harvard Psychoacoustic Sentences* é um conjunto fechado de 720 sentenças separadas em 72 séries desenvolvido para testar a inteligibilidade de palavras no contexto de uma sentença [52]. A base foi desenvolvida para a língua inglesa e as sentenças são balanceadas foneticamente.

O processo de aplicação do método consiste em reproduzir para o ouvinte uma frase por vez, que por sua vez deve transcrever na folha de respostas a frase que julga ter ouvido. O processo de avaliação das frases transcritas podem ser por meio do cálculo *Word Error Rate* (WER) ou uma nota pode ser dada de acordo com a proximidade da frase transcrita com a frase falada. Alguns exemplos de sentenças da base de dados são apresentados abaixo [52]:

- *The boy was there when the sun rose*
- *A rod is used to catch pink salmon*
- *The source of the huge river is the clear spring*
- *Kick the ball straight and follow through*

Pelo fato da base de dados ser fechada, o efeito de aprendizagem pode ser perceptível em caso de repetição de testes com um mesmo ouvinte [53] [19]. Atualmente, não existe nenhuma versão da *Harvard Psychoacoustic Sentences* para outras línguas a não ser para a língua inglesa.

4.2.2 *Haskins Sentences*

Assim como a *Harvard Psychoacoustic Sentences*, a base de dados *Haskins Sentences* é um conjunto fechado de 200 sentenças separadas em 4 séries desenvolvido para testar a inteligibilidade de palavras no contexto de uma sentença [54]. No entanto, diferentemente da *Harvard Psychoacoustic Sentences*, as frases são sintaticamente corretas porém semanticamente confusas.

Na maioria das vezes, as frases nunca são utilizadas em um cenário real, e possivelmente nunca foram ouvidas pelo ouvinte [53].

O processo de aplicação do método é exatamente igual ao do *Harvard Psychoacoustic Sentences*, por meio do cálculo WER ou por nota, de acordo com a proximidade da frase transcrita com a frase falada. As primeiras quatro sentenças da base de dados são apresentadas abaixo [54]:

- *The live farm got the book*
- *The white peace spoke the share*
- *The black shout caught the group*
- *The end field sent the point*

Apesar das sentenças serem complicadas de memorizar por não existir um significado claro, elas devem ser usadas somente uma vez para que não ocorra o efeito de aprendizagem. Atualmente, não existe nenhuma versão da *Haskins Sentences* para outras línguas a não ser para a língua inglesa.

4.2.3 *Semantically Unpredictable Sentences* (SUS)

Diferentemente da *Harvard Psychoacoustic Sentences* e da *Haskins Sentences*, *Semantically Unpredictable Sentences* (SUS) não é uma base de dados. A SUS é uma especificação de como construir uma base de dados sintaticamente correta porém semanticamente confusa [55]. Apesar de existirem bases fechadas de sentenças SUS em diversas línguas, nenhuma delas é considerada oficial.

A definição original do SUS define que o teste deve conter sentenças dentro de 5 estruturas gramaticais, como descrito na Tabela 4.3. No entanto, vários estudos de análise de voz utilizam frases semanticamente confusas mas que não necessariamente estejam dentro da estrutura gramatical proposta na especificação original [8] [56] [12]. Isso se deve pela essência do uso de sentenças SUS. O principal objetivo de utilizar sentenças sintaticamente corretas porém semanticamente confusas é de não permitir a adivinhação de palavras com base no contexto da sentença, evitando assim o incremento da taxa de inteligibilidade por conta da adivinhação. Sendo assim, as sentenças não necessariamente devem estar dentro da estrutura gramatical proposta pela especificação original para que consigam atingir o objetivo.

Tabela 4.3: Estrutura Gramatical da Especificação Original do SUS [6]

	Estrutura	**Exemplo**
1	*Subject - verb - adverbial*	*The table walked through the blue truth*
2	*Subject - verb - Direct object*	*The strong way drank the day*
3	*Adverbial - verb - direct object*	*Never draw the house and the fact*
4	*Q-word - transitive verb - subject - direct object*	*How does the day love the bright word*
5	*Subject - verb - complex direct object*	*The plane closed the fish that lived*

As estruturas gramaticais foram impostas originalmente para que as sentenças pudessem ser geradas automaticamente com palavras aleatórias contidas em uma base de dados, de forma a desprezar o efeito de aprendizagem. No entanto, a geração automática de frases exige uma extensa base de dados a fim de não gerar sentenças repetidas. Sendo assim, a necessidade de um software gerador de sentenças e de uma extensa base de dados pode ser um fator complicante. Por conta disso, bases de dados de sentenças SUS são amplamente utilizadas.

4.2.4 *Word Error Rate* (WER)

O cálculo do *Word Error Rate* (WER) é derivado da *Levenshtein Distance* [57] e permite avaliar a inteligibilidade de um sintetizador de voz. A *Levenshtein Distance* mensura a diferença entre duas cadeias de caracteres (palavras) enquanto o WER mensura a diferença entre duas sentenças. Para se obter a taxa de inteligibilidade pelo cálculo WER, pede-se aos ouvintes que escutem uma sentença e as escrevam, para que posteriormente se compare a sentença original e a sentença redigida. O WER mede o grau de diferença entre elas, de acordo com a Equação 4.1.

$$WER = \frac{S + D + I}{N} \tag{4.1}$$

onde

- S é o número de substituições de palavras;
- D é o número de exclusões de palavras;
- I é o número de inserções de palavras;
- N é o número total de palavras da sentença original.

Na Tabela 4.4 apresentam-se alguns exemplos de aplicação do cálculo WER em diferentes sentenças SUS. As palavras sublinhadas correspondem às alterações nas sentenças. A coluna S corresponde ao número de substituições; a coluna D corresponde ao número de exclusões; a coluna I corresponde ao número de inserções e a coluna WER corresponde ao grau de diferença entre a sentença original e a sentença redigida.

Tabela 4.4: Exemplos de Aplicação do Cálculo WER

Sentença Original	**Sentença Redigida**	**S**	**D**	**I**	**WER**
O carro cantou no alto da maçã	O carro morou no alto da maçã	1	0	0	0.14
A goiaba bebeu uma parede	a goiaba uma parede	0	1	0	0.2
A mesa gritou para o coelho	A mesa gritou para o coelho branco	0	0	1	0.16

O cálculo WER pode ser aplicado em qualquer tipo de sentença. No entanto, é preferível que seja aplicado em sentenças sintaticamente corretas porém semanticamente confusas, tais como as da base de dados *Haskins Sentences* ou frases elaboradas seguindo as diretrizes das sentenças SUS. Essa estratégia é adotada para evitar que a taxa de acerto não seja influenciada pelo efeito de aprendizagem.

4.3 Avaliação de Compreensibilidade

Diferentemente da inteligibilidade, não se pode avaliar o nível de compreensibilidade que uma voz proporciona solicitando aos ouvintes que simplesmente escrevam as frases ou palavras ouvidas. Isso se deve pelo fato de que o não entendimento de uma palavra pode não comprometer o entendimento de uma sentença [58] [29].

Na avaliação de compreensibilidade se busca descobrir se a voz sintetizada foi capaz de transmitir a mensagem de forma coerente e clara e se permitiu ao ouvinte compreender o que foi dito. Para isso é necessário que os ouvintes sejam questionados quanto à interpretação e ao significado da mensagem transmitida com o sinal de fala sintetizado.

Para avaliar a compreensibilidade, o avaliador deve ouvir uma notícia para posteriormente responder questões sobre ela. As questões podem ser dissertativas ou objetivas e devem englobar somente o conteúdo dito na notícia sintetizada [29] [8]. Abaixo temos um exemplo de notícia e às questões referentes a ela.

> Para os notívagos de plantão, a madrugada desta terça-feira, 15, terá um espetáculo à parte. Um eclipse total da Lua poderá ser visto em toda a América a partir das 3 horas, horário de Brasília. O pesquisador do Observatório Nacional, Jair Barroso, explica que a visibilidade do fenômeno dependerá das condições do céu. "É importante que não tenham nuvens a ponto de impedir a visão".

> O que pode influenciar na visibilidade do fenômeno?
> R: _____

> O que pôde ser visto na madrugada do dia 15?
> (A) Um meteoro
> (B) Um cometa
> (C) Um eclipse
> (D) Uma estrela
> (E) Um planeta

O processo de avaliação depende do tipo da questão. Para questões dissertativas se adota o seguinte critério: Nota 2 se a resposta estiver plenamente correta, nota 1 se estiver parcialmente correta e nota 0 se estiver errada. Para questões objetivas, somente o certo e errado é considerado.

A taxa de compreensibilidade está diretamente relacionada com o número de acertos, sendo assim, a taxa de compreensibilidade é a média das notas dadas [8]. Apesar de ambas medirem o grau de compreensibilidade de uma voz, as questões objetivas dão margem a "chutes" que podem incrementar ou decrementar a taxa de compreensibilidade. Deve-se evitar o uso de notícias extensas, pois o ouvinte pode errar a resposta por conta do esquecimento e não pela falta de compreensão.

4.4 Estimação Categórica

Além dos métodos de avaliação de inteligibilidade e compreensibilidade, outras características podem ser avaliadas, tais como a naturalidade, suavidade, nitidez, dentre outras [59]. Para a avaliação dessas características, usa-se a escala *Mean Opinion Score* (MOS).

Apesar de ter sido originalmente desenvolvida para avaliar a qualidade dos padrões de codificação, a escala MOS é amplamente utilizada para avaliar a qualidade de uma voz sintetizada. A escala MOS possuí 5 níveis, conforme mostra a Figura 4.2. O processo de aplicação consiste em pedir para que um ouvinte ouça uma sentença e atribua uma nota para a característica que está sendo julgada.

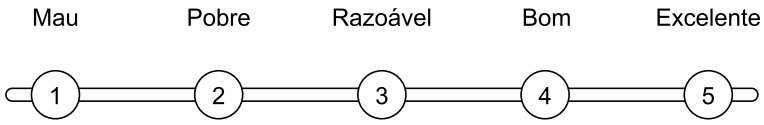


Figura 4.2: Escala MOS.

Na estimação categórica, diversos atributos podem ser avaliados independentemente pela escala MOS. Recomenda-se avaliar por meio da escala MOS as características que não podem ser avaliadas com o uso de métodos objetivos, como os usados para avaliar a inteligibilidade e a compreensibilidade. A Tabela 4.5 apresenta algumas características que podem ser avaliadas pela escala MOS. Apesar das classificações para cada característica variar na nomenclatura, todas elas são escalas de 5 níveis, assim como a escala MOS.

Tabela 4.5: Exemplos de Características de Voz [7]

Atributo	**Classificação**
Pronunciação	Não irritante ... Muito irritante
Velocidade	Lento demais ... Rápido demais
Nitidez	Muito nítido ... Pouco nítido
Naturalidade	Muito natural ... Pouco natural
Estresse	Pouco extressante ... Muito extressante
Inteligibilidade	Excelente ... Pobre
Compreensibilidade	Excelente ... Pobre
Suavidade	Muito suave ... Pouco suave

A inteligibilidade e a compreensibilidade não devem ser avaliadas somente com o uso de escala MOS, pois características como a naturalidade podem levar a falsa sensação de inteligibilidade e compreensibilidade. Por outro lado, pode ser interessante avaliar a inteligibilidade e a compreensibilidade usando escala MOS e métodos objetivos, como os descritos nas seções anteriores, para analisar a influência da naturalidade na inteligibilidade e compreensibilidade [12]. Em situações onde somente a inteligibilidade e a compreensibilidade importam, as avaliações usando a escala MOS podem ser omitidas.

4.5 Avaliação em Campo

Quando o processo de avaliação de voz for para uma aplicação específica, é importante que a voz seja avaliada em conjunto com a aplicação e no ambiente no qual a aplicação será usada. Em um cenário de uso real, diversos fatores podem influenciar na capacidade da voz transmitir sua mensagem de forma correta. Por exemplo, um leitor para cegos é composto de dois módulos principais: reconhecedor de caracteres e sistema TTS. A qualidade do leitor depende da qualidade do reconhecedor e do sistema TTS. No entanto, se o leitor identificar uma palavra errada, o sistema TTS também irá pronunciá-la erroneamente, de forma que o problema se encontra no módulo de reconhecimento, não do sistema TTS. É importante conhecer toda a plataforma hospedeira do sistema TTS para identificar as potenciais falhas.

Um segundo exemplo é o uso de sistemas TTS na telefonia. Atualmente, grande parte das empresas utilizam sistemas STT e TTS para direcionar chamadas para departamentos específicos ou até mesmo para solucionar problemas casuais. No entanto, em um sistema de telefonia, degradações da voz podem ocorrer na transmissão, e se a voz já tiver algum tipo de degradação própria, a sua compreensão pode ser inviabilizada.

Um terceiro exemplo é quando a voz é utilizada em conjunto com uma face, que pode ser real ou virtual. Estudos comprovam que, quando a voz é avaliada acompanhada de uma face, seja ela real ou virtual, a taxa de inteligibilidade aumenta consideravelmente, principalmente quando a face transmite informações, como sincronismo labial e emoção [60] [61] [62].

Na etapa de avaliação em campo também se deve questionar a real necessidade de um sistema TTS. Por mais que os sistemas TTS atuais cheguem próximo da voz humana em relação a inteligibilidade e compreensibilidade, a voz natural transmite maior confiança e credibilidade. Em situações onde o fator humano e emoções devem ser transmitidos, talvez sistemas TTS não seja a melhor escolha.

4.6 Considerações Finais

Como apresentado neste capítulo, atualmente existem inúmeros métodos de avaliação de voz. Cada método tem por finalidade quantificar determinadas características da voz, como a sua inteligibilidade, compreensibilidade, naturalidade, dentre outras. O maior desafio no processo de avaliação de voz é identificar quais métodos devem ser utilizados para avaliar suficientemente uma voz para uma determinada aplicação. Na literatura se encontram muitos trabalhos apresentando métodos de avaliação isolados, mas poucos trabalhos buscam esclarecer de que forma os métodos devem ser aplicados em conjunto dada uma determina situação. Pelo fato da avaliação de voz ser subjetiva em muitos sentidos, o processo de avaliação automática de características como prosódia e naturalidade ainda é um desafio.

Capítulo 5

Relações entre Inteligibilidade, Compreensibilidade e Naturalidade

Chang [8] apresenta em seu artigo intitulado "*Evaluation of TTS Systems in Intelligibility and Comprehension Tasks*" a relação entre inteligibilidade e compreensibilidade. Para seu estudo, ele utilizou uma voz natural para servir como controle, e os sintetizadores HTS-2008 e Multisyn.

A hipótese inicial do autor era que o nível de compreensibilidade tivesse uma relação direta com o nível de inteligibilidade. No entanto, os resultados foram contra a hipótese inicial. O nível de inteligibilidade variou entre os sintetizadores, como mostra a Tabela 5.1, e não se observou uma diferença significativa no nível de compreensibilidade.

Tabela 5.1: Resultados para Inteligibilidade por Chang [8]

	WER	**Desvio Padrão**
Voz Humana	4.2%	10%
HTS-2008	6.7%	11.4%
Multisyn	14.3%	21.6%

Para complementar o estudo de Chang, propusemos uma avaliação para encontrar as relações entre inteligibilidade, compreensibilidade e naturalidade. Os resultados mostraram que o nível de naturalidade, assim como a inteligibilidade, influenciam o nível da compreensibilidade. Os resultados são apresentados e comentados nas subseções seguintes.

5.1 Metodologia do Experimento

Para avaliar a relação entre inteligibilidade, compreensibilidade e naturalidade, foi realizada uma pesquisa de campo. Ao todo, 30 pessoas, com idades variando entre 20 e 40 anos, responderam a um questionário *online* na plataforma *Google* Formulários.

Foram selecionados somente indivíduos que tinham a língua portuguesa como língua materna e com pelo menos o ensino médio completo, de forma a garantir a fluência no idioma na modalidade escrita e falada. Cada ouvinte foi orientado quanto à metodologia dos testes e o questionário eletrônico não permitia que o ouvinte avançasse sem ter respondido todas as questões. Os ouvintes podiam escutar cada frase uma única vez.

Na avaliação foram utilizadas 3 vozes. Uma voz de controle que consistia na gravação da voz humana e duas vozes sintéticas comerciais de empresas líderes no setor. A fim de preservar as empresas, as vozes aparecerão identificadas como voz do sintetizador A (Voz A) e voz do sintetizador B (Voz B).

O questionário aplicado avaliava as três vozes em sequência para cada teste. Para cada voz foi aplicado um número idêntico de questões, com nível de dificuldade equivalente. Para o MRT havia 8 conjuntos de palavras, sendo 4 conjuntos para avaliar a consoante inicial e 4 conjuntos para avaliar a consoante final. Para avaliar a inteligibilidade pelo cálculo WER, foram aplicadas 6 frases para cada voz. No teste de compreensibilidade cada ouvinte escutou 3 notícias diferentes de cada voz. No teste subjetivo de naturalidade e inteligibilidade, cada ouvinte escutou uma frase para cada quesito. A duração do teste foi de cerca de 20 minutos para cada ouvinte, gerando um total de 1.710 respostas.

5.2 Resultados e Discussão

A Figura 5.1 resume os resultados dos experimentos usando o método MRT. Observa-se que a voz humana obteve o pior desempenho na avaliação, provavelmente porque o locutor não deu a ênfase necessária ao pronunciar cada sílaba, levando, deste modo, os ouvintes a não entenderem corretamente as palavras dissílabas pronunciadas isoladamente. No entanto, podemos observar que todas as vozes apresentaram um alto nível de inteligibilidade, chegando até mesmo a 100% de acerto, no caso da voz sintética B nos testes da primeira sílaba e na voz sintética A nos testes da segunda sílaba. A voz natural, apesar de ter alcançado a menor nota de inteligibilidade, ainda assim conseguiu uma taxa de acerto de 92,40% tanto nos testes da primeira sílaba quanto da segunda sílaba.

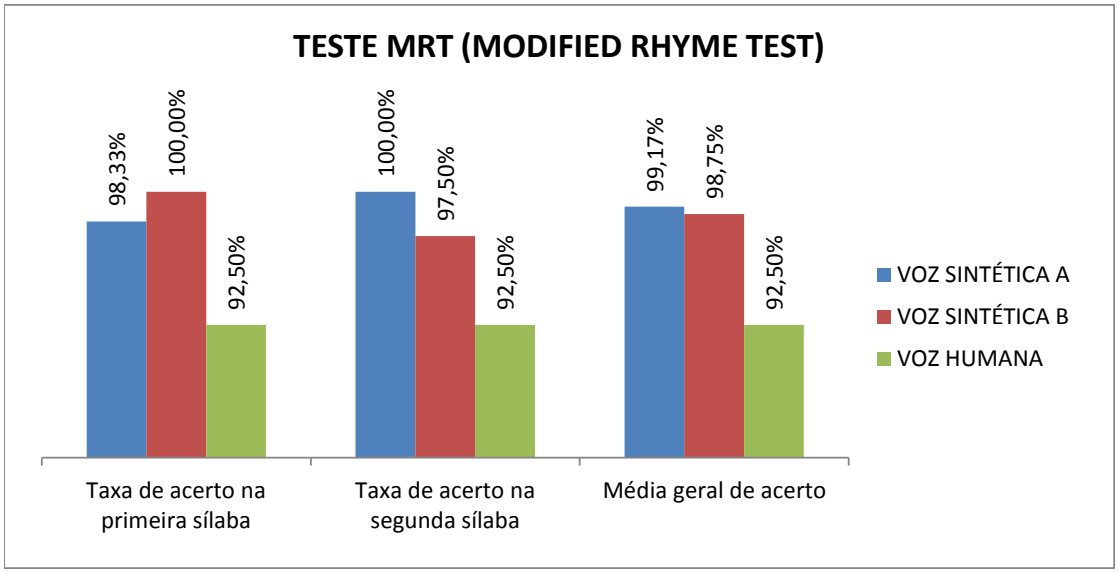


Figura 5.1: Resultados para o Teste MRT (*Modified Rhyme Test*).

A Figura 5.2 resume os resultados dos experimentos usando o método WER com sentenças SUS. Observa-se que, diferentemente dos resultados obtidos com o método MRT, os resultados

utilizando o método WER mostraram que a voz humana se saiu melhor que as duas vozes sintetizadas. Isso pode ter ocorrido por conta da junção natural das palavras proporcionada pela voz humana. No entanto, podemos observar que todas as vozes obtiveram um baixo nível de erro, chegando a no máximo 10,59% no caso da voz sintética A. Com base nos testes MRT e WER, podemos concluir que as três vozes apresentam uma boa taxa de inteligibilidade.

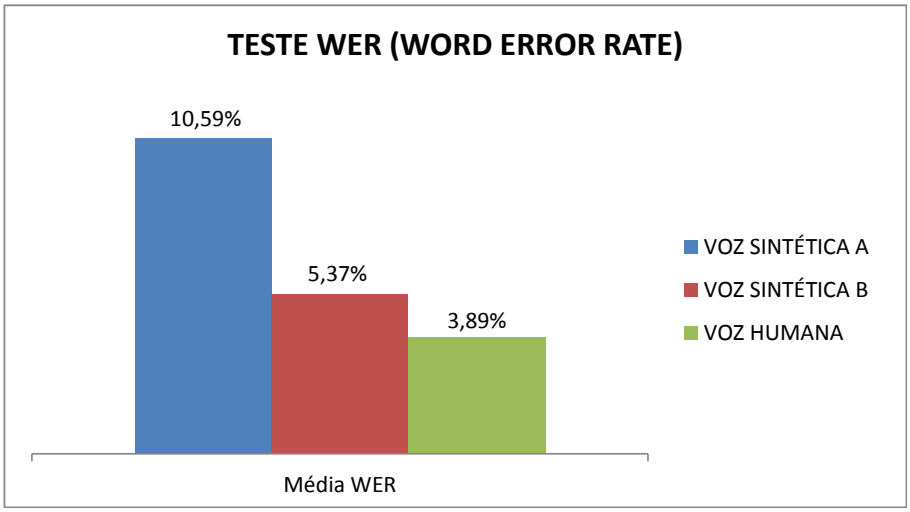


Figura 5.2: Resultados para o Teste WER (*Word Error Rate*).

A Figura 5.3 resume os resultados dos experimentos usando o método de avaliação de compreensibilidade. Observa-se que a taxa de compreensibilidade teve uma grande variação entre as vozes. A voz sintética A foi a que mais destoou dentre as outras, apesar de ter tido uma taxa semelhante nos teste de inteligibilidade. A voz humana foi a que obteve a maior taxa de compreensibilidade entre todas, possivelmente pela naturalidade e a emoção transmitida durante a leitura das notícias.

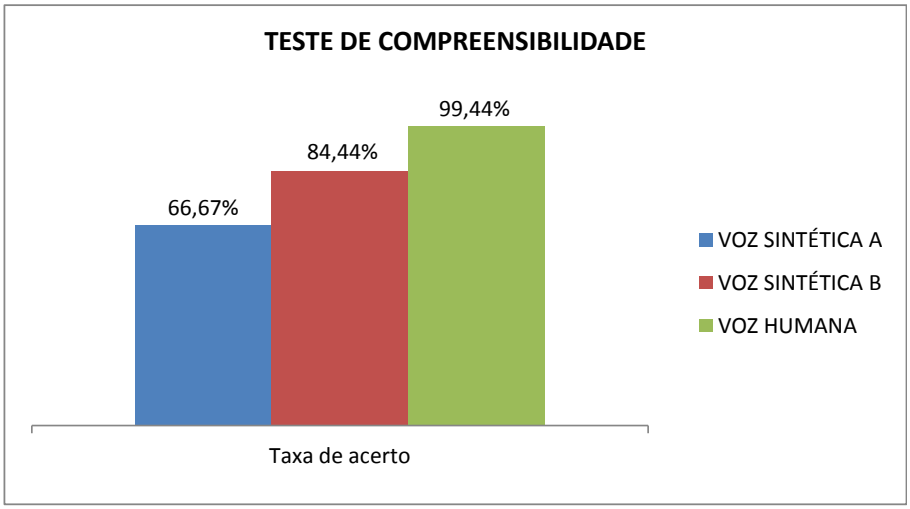


Figura 5.3: Resultados para o Teste de Compreensibilidade.

Neste estudo observa-se o inverso dos resultados obtidos por Chang [8]. Enquanto Chang

obteve resultados diferentes para a inteligibilidade e resultados semelhantes para a compreensibilidade, obtivemos resultados semelhantes para a inteligibilidade e diferentes para a compreensibilidade. Isso mostra que não somente a inteligibilidade influencia na compreensibilidade. Para buscar compreender o que influencia na compreensibilidade, foi feita uma avaliação da naturalidade.

A Figura 5.4 resume os resultados dos experimentos usando o método de avaliação de naturalidade. As notas, de 0 a 5, sendo 0 a pior nota e 5 a melhor nota, mostram que a voz natural foi a que obteve a melhor naturalidade, como o esperado. Dentre os sintetizadores, a voz sintética A foi a que obteve a melhor nota.

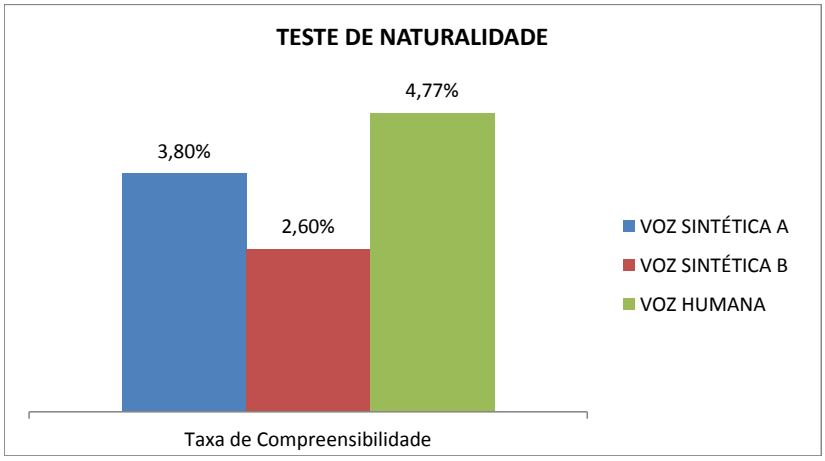


Figura 5.4: Resultados para o Teste de Naturalidade.

Também foi feita uma avaliação para medir a inteligibilidade e a compreensibilidade por meio de testes subjetivos usando a escala MOS. Na Figura 5.5 podemos visualizar a taxa de inteligibilidade usando escala MOS.

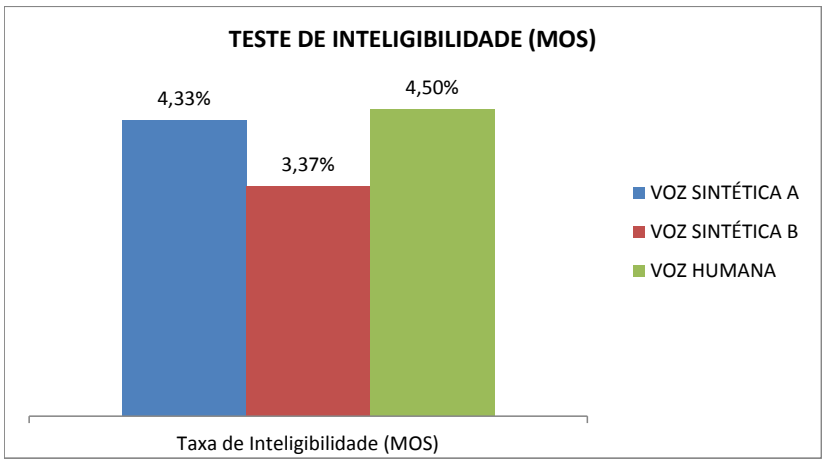


Figura 5.5: Resultados para o Teste de Inteligibilidade.

Observa-se que pelo teste subjetivo, as vozes mais bem avaliadas foram as que obtiveram o melhor resultado no teste subjetivo de naturalidade. Desta forma, pode-se concluir que os

avaliadores não se sentem seguros da inteligibilidade da voz quando a naturalidade dela está comprometida.

Pode-se intuir também que a nota subjetiva atribuída à voz sintética B no quesito inteligibilidade foi penalizada pelo fato dela ser menos natural. Sendo assim, pelo menos nas avaliações subjetivas, a naturalidade é um parâmetro a se considerar nas análises de inteligibilidade e compreensibilidade. A Figura 3 permite ver que existe uma relação inversa entre a avaliação subjetiva da inteligibilidade e a taxa de acerto nos testes.

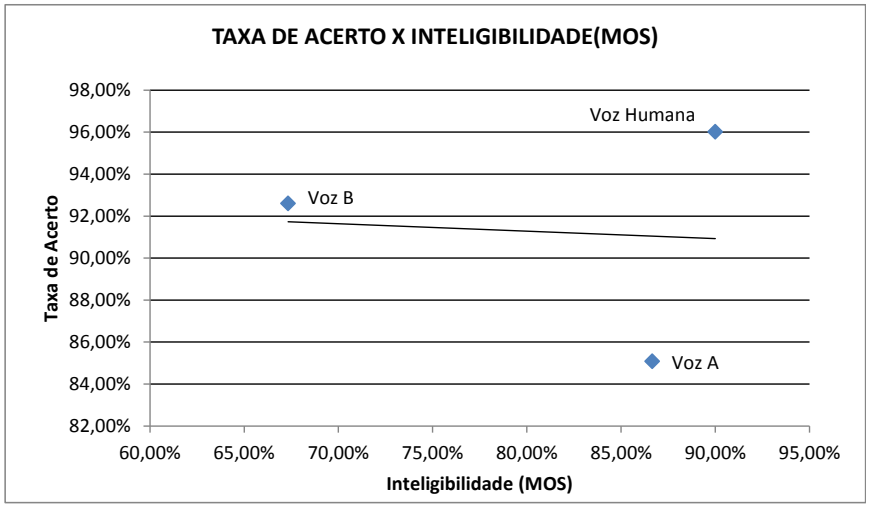


Figura 5.6: Inteligibilidade - Avaliação Subjetiva X Objetiva.

A Figura 5.7 permite visualizar que realmente existe uma correlação direta entre as notas atribuídas para os quesitos naturalidade e inteligibilidade da voz na avaliação subjetiva, o que comprova a hipótese anterior, de que os avaliadores não se sentem seguros quanto a inteligibilidade quando a naturalidade está comprometida.

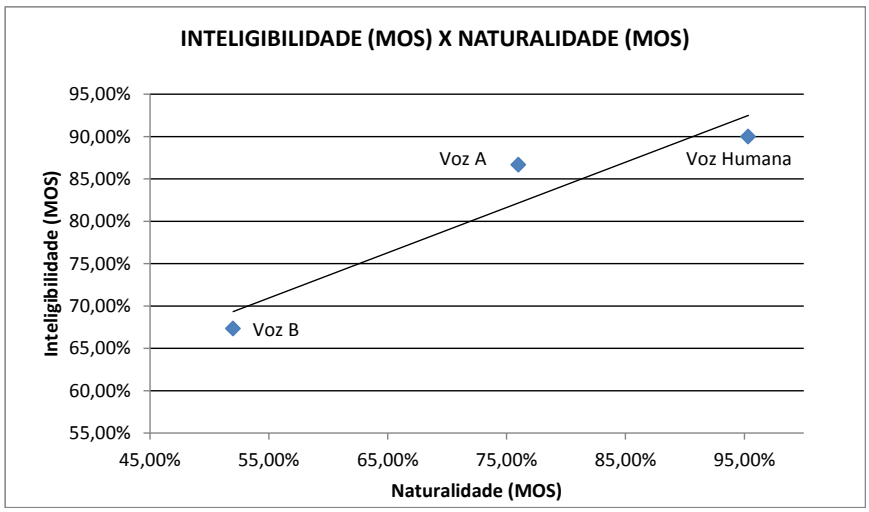


Figura 5.7: Correlação entre Naturalidade e Inteligibilidade.

Outro resultado interessante que pode ser inferido observando a Figura 5.3 são as notas obtidas no teste de compreensibilidade. Neste teste a voz sintética A demonstrou uma taxa

de erro em torno de 44% e a voz sintética B uma taxa de erro em torno de 20%, enquanto que a voz humana obteve acerto de praticamente 100%. Este resultado parece contraditório, ou ao menos insatisfatório comparado com as notas obtidas pelas vozes artificiais nos testes de inteligibilidade.

Para compreender esta aparente contradição, pode-se comparar os resultados obtidos no teste WER e no teste de compreensibilidade. Em ambos os testes o ouvinte escuta frases, sendo que no primeiro as palavras são descontextualizadas e as frases não têm sentido lógico, enquanto que no teste de compreensibilidade as palavras formam frases com significado. Pode-se ver na Figura 5.8 que existe uma correlação direta entre a inteligibilidade e a compreensibilidade.

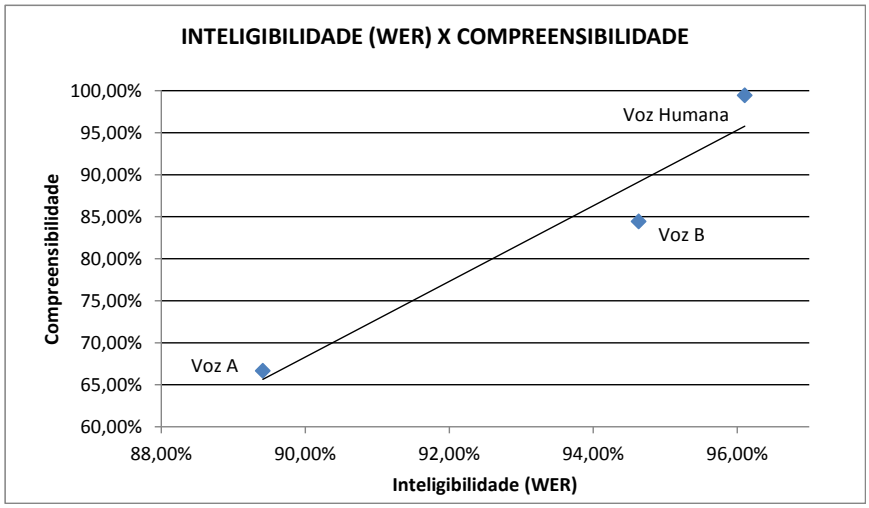


Figura 5.8: Correlação entre Compreensibilidade e Inteligibilidade.

Todavia, observa-se que o eixo Y (Compreensibilidade) exibe valores percentuais inferiores aos correspondentes no eixo X (Inteligibilidade). A proporção entre eles são equivalentes somente para o caso da voz humana, para quais ambas as variáveis possuem acerto acima de 95%. Sendo assim, podemos dizer que a compreensibilidade está associada a peculiaridades cognitivas que vão além da inteligibilidade. Apesar de não podermos afirmar com propriedade que a naturalidade afeta diretamente a compreensibilidade, podemos dizer que existem indícios, visto que nos testes subjetivos de inteligibilidade e compreensibilidade as vozes com maior nível de naturalidade se saíram melhores.

Além do mais, é interessante notar que, entre as vozes artificiais, a voz sintética B apresentou um desempenho melhor nos testes de inteligibilidade e compreensibilidade, entretanto, recebeu notas menores nos testes de preferência subjetiva, tanto para naturalidade como para inteligibilidade. Isto indica que a naturalidade da voz do sintetizador B é inferior à do sintetizador A e que os ouvintes deram um peso importante na naturalidade da voz ao exprimirem suas preferências.

Por fim, a Figura 5.9 permite comparar o espectrograma e a forma de onda dos sinais de fala gerados pela voz humana e pelos sintetizadores correspondentes a palavra “montanha”. A voz humana e a voz gerada pelo sintetizador A são masculinas e a voz B é feminina. Observa-se no espectrograma que as vozes sintéticas aproximam-se bem da voz humana, apesar das formantes e a transição entre o “tanha” de “montanha” não estarem tão bem definidas.

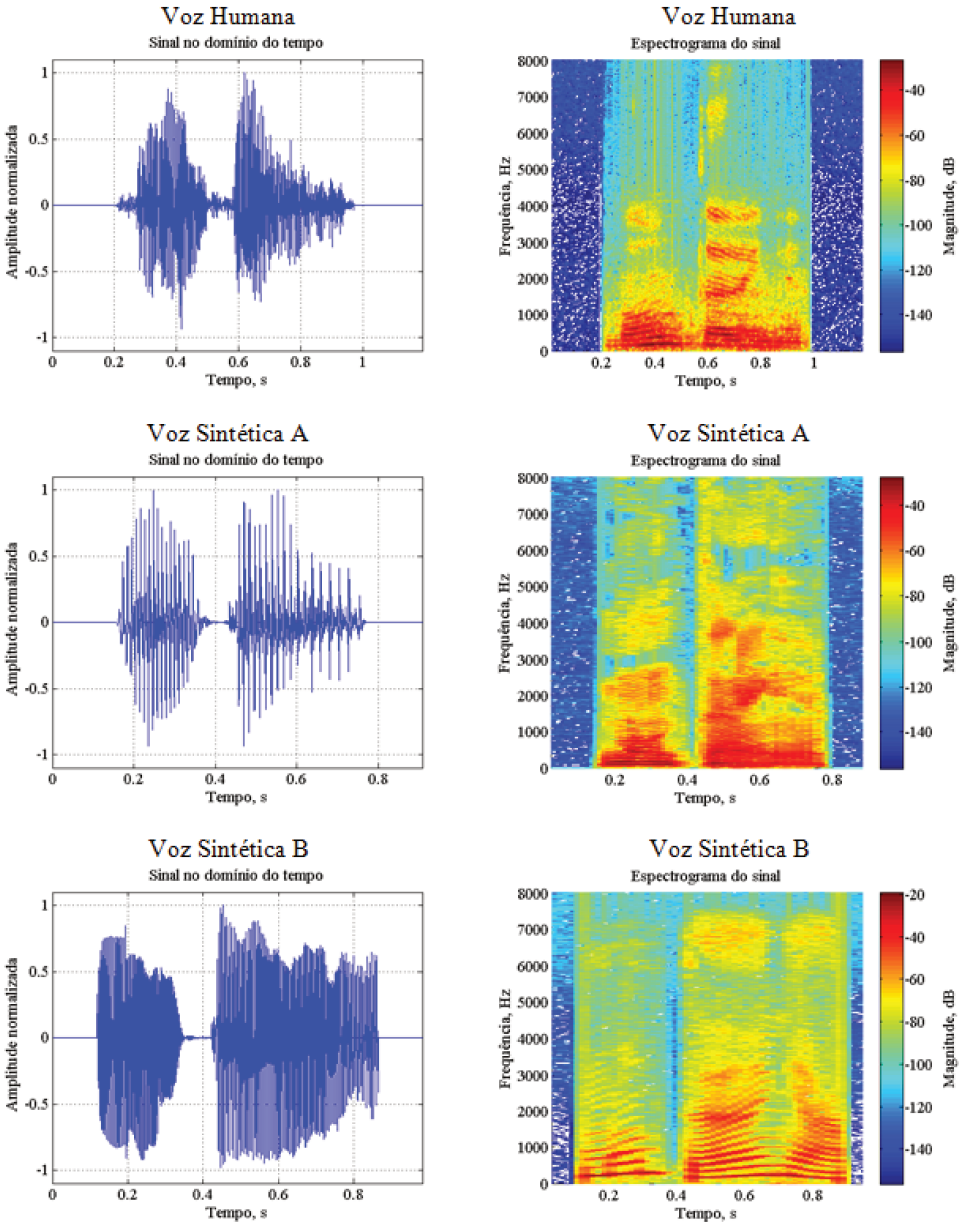


Figura 5.9: Comparação da Forma de Onda e do Espectrograma da Palavra “montanha” sintetizada.

5.3 Considerações Finais

A partir dos resultados pode-se concluir que os testes objetivos de inteligibilidade levaram a uma taxa de acerto equivalente entre as vozes artificiais e a voz humana, indicando a boa qualidade e eficiência dos sintetizadores. Entretanto, no teste de compreensibilidade nota-se que as vozes sintéticas não são capazes de atingir os mesmos resultados que a voz humana, indicando que, além da inteligibilidade, a naturalidade, a prosódia do discurso e outros fatores são importantes no mecanismo cognitivo humano. Apesar das vozes sintéticas apresentarem um desempenho inferior à da voz humana nos testes de compreensibilidade, as taxas de acerto são aceitáveis, mostrando que o uso de vozes artificiais em situações reais é plenamente possível.

Capítulo 6

Metodologia de Avaliação de Voz

O processo de avaliar uma voz sintética envolve o uso de métodos de avaliação com ouvintes. Atualmente, um grande número de métodos de avaliação pode ser encontrado na literatura, conforme apresentado no Capítulo 4. O uso correto destes métodos, assim como a escolha do método ideal para cada situação são fatores fundamentais para uma boa avaliação. Este capítulo apresenta uma metodologia de avaliação de voz sintetizada, composta de vários métodos de avaliação, e busca estruturar o processo de avaliação em 6 Etapas. A metodologia foi testada e os resultados são apresentados.

6.1 Metodologia Proposta

A metodologia proposta consiste em 6 Etapas: Análise do Cenário, Seleção de Vozes, Identificação de Métodos de Avaliação, Avaliação, Análise dos Resultados e Escolha. A Figura 6.1 apresenta as 6 Etapas da Metodologia.

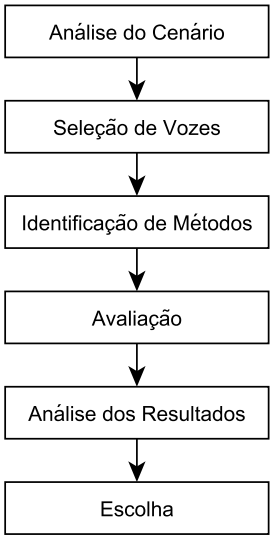


Figura 6.1: Metodologia Proposta.

Para facilitar a aplicação da metodologia proposta, um *template* foi desenvolvido. Neste *template*, todas as instruções referentes a todas as etapas estão presentes. O *template* se encontra no Apêndice desta dissertação. As subseções seguintes apresentam cada etapa da metodologia.

6.1.1 Etapa 1 - Análise do Cenário

A Etapa 1, denominada como "Análise do Cenário", tem por objetivo fornecer uma visão de alto nível do sistema hospedeiro, e o que levou a escolha de usar sistemas TTS. A Etapa 1 é de fundamental importância, pois os métodos de avaliação a serem aplicados dependem das características que desejam ser avaliadas. As informações requeridas nesta etapa são:

1. Nome do Projeto
2. Motivação para o uso de Sintetizador de Voz
3. Idioma da Voz
4. Sexo da Voz
5. Características Requeridas
6. Plataforma do Projeto
7. Informações Complementares

No primeiro item, o avaliador define o nome do projeto hospedeiro para fins de documentação. No segundo item, o avaliador descreve a motivação que levou a escolha de usar um sistema TTS. É de fundamental importância que o sistema TTS seja utilizado somente em situações onde a voz humana não possa ser utilizada de maneira adequada. Em geral, sistemas TTS são usados em situações onde não é previsível o texto a ser narrado (como em atendimento automatizado) ou em situações onde o roteiro do texto muda constantemente e a regravação da voz humana se torna inviável.

No terceiro item, o avaliador define o idioma da voz. No quarto item, o avaliador define se a voz deverá ser masculina, feminina, ou se é indiferente. No quinto item, o avaliador define quais serão as caracteristicas requeridas. Nela o avaliador define se somente a inteligibilidade e a compreensibilidade são importantes ou se caracteristicas subjetivas como a pronúncia, velocidade da fala e naturalidade devem ser levadas em conta.

No sexto item, o avaliador indica qual será a plataforma do projeto, uma vez que um determinado sistema TTS pode não oferecer suporte para todas as plataformas. E no sétimo item, o avaliador fornece informações complementares, se necessário, tais como limitações de projeto, tempo máximo a ser gasto com o processo de avaliação e até mesmo o valor máximo a ser investido no sistema TTS.

6.1.2 Etapa 2 - Seleção de Vozes

A Etapa 2, denominada como "Seleção de Vozes", tem por objetivo listar o conjunto de vozes a serem avaliadas. As vozes devem ser selecionadas de forma a satisfazer os requisitos definidos na Etapa 1. As informações requeridas para cada voz selecionada são:

- Nome do Sintetizador
- Nome da Voz
- Idioma
- Sexo
- Plataformas Suportadas
- Informações Técnicas
- Tipo de Licença

No primeiro item o avaliador define o nome do sistema TTS. No segundo item o nome da voz deve ser definido, pelo fato de um sistema TTS poder ter mais de uma voz. No terceiro item o avaliador indica o idioma da voz, que deve estar de acordo com o terceiro item da Etapa 1. No quarto item o avaliador define se a voz é masculina ou feminina.

No quinto item o avaliador lista as plataformas suportadas pelo sistema TTS, que devem estar de acordo com o sexto item da Etapa 1. No sexto item o avaliador descreve as informações técnicas, como a interface a ser usada entre o sistema hospedeiro e o sistema TTS, por exemplo. No sétimo item o avaliador indica o tipo de licença do produto, que pode ser comercial ou *open source*.

6.1.3 Etapa 3 - Identificação de Métodos

A Etapa 3, denominada como "Identificação de Métodos", tem por objetivo apresentar os métodos de avaliação a serem usados, de acordo com o quinto item da Etapa 1. As opções, assim como os métodos de avaliação de acordo com cada opção seguem abaixo:

- Métodos para avaliar inteligibilidade, compreensibilidade e teste em campo
  - WER (*Word Error Rate*)
  - Questões sobre notícias sintetizadas
  - Teste em campo
- Métodos para avaliar intelibilidade, compreensibilidade, características subjetivas e teste em campo
  - WER (*Word Error Rate*)
  - Questões sobre notícias sintetizadas

- Escala MOS (*Mean Opinion Score*) para avaliações subjetivas
- Teste em campo

O teste em campo está presente nas duas opções por ser de fundamental importância. Uma vez que as vozes são avaliadas individualmente, fatores do ambiente como ruídos e conversas paralelas são minimizados e podem levar a um falso resultado da real eficácia do uso do sistema TTS. Além do mais, dependendo do sistema hospedeiro, a qualidade do sistema TTS pode ser incrementada ou decrementada dependendo dos equipamentos de áudio disponíveis.

6.1.4 Etapa 4 - Avaliação

A Etapa 4, denominada como "Avaliação", consiste no processo de avaliação propriamente dita. Nela, o avaliador deverá preparar um formulário, manualmente ou com o uso de algum software gerador de formulários e aplicar a um grupo de ouvintes, que podem ser ouvintes profissionais ou amadores.

É recomendado que as avaliações sejam feitas individualmente, em uma sala isolada, evitando qualquer tipo de distração ao ouvinte. Também é importante que os ouvintes usem fone de ouvido, todos eles com o mesmo volume. Os áudios devem ser executados um de cada vez em uma ordem pré-definida.

A metodologia não define a quantidade de testes que devem ser feitos e nem mesmo a população adequada, pois são informações dependentes do tipo do projeto.

6.1.5 Etapa 5 - Análise dos Resultados

A Etapa 5, denominada como "Análise dos Resultados", consiste no processo de avaliar os dados obtidos na Etapa 4. Para cada método, um tipo de análise deve ser feita.

Para calcular a taxa de acerto de uma voz usando o método WER, primeiramente se calcula o WER para cada frase, e em seguida se calcula a média WER para cada indivíduo. Em seguida, se calcula a média geral com base nas médias de cada ouvinte. A média geral é a taxa de erro da voz.

Para calcular a taxa de compreensibilidade, primeiramente se calcula a média de compreensibilidade para cada ouvinte. Em seguida, se calcula a média geral com base nas médias de cada ouvinte. A média geral é a taxa de compreensibilidade da voz.

Na avaliação das características subjetivas, se usa escala MOS. A média da pronúncia, velocidade da fala e naturalidade é dada pela média geral de todos os indivíduos.

Por fim, o avaliador terá um conjunto de resultados que permitirá encontrar a melhor voz para cada uma das características avaliadas.

Os métodos de avaliação usados na metodologia proposta foram explicados em detalhes no Capítulo 4 desta dissertação.

6.1.6 Etapa 6 - Escolha

A Etapa 6, denominada como "Escolha", tem por objetivo identificar a melhor voz para o projeto. A melhor voz não necessariamente deve ser a voz que obteve o melhor resultado

em todas as características avaliadas, mas sim a voz que obteve os melhores resultados nas características desejadas para o projeto.

6.2 Aplicação da Metodologia

Para testar a metodologia proposta, esta foi aplicada em uma plataforma de ensino a distância usando avatares. O *template* da metodologia preenchido com as informações do projeto são apresentados no Apêndice desta dissertação. Nas subseções seguintes, o sistema hospedeiro é apresentado, assim como os resultados da avaliação.

6.2.1 Descrição do Sistema Hospedeiro

A plataforma de ensino a distância usando avatares foi desenvolvida por Cinto [5] como trabalho de mestrado e consiste em dois módulos: (1) Editor de Aulas e (2) Sala de Aula Virtual.

O Editor de Aulas foi idealizado de forma a apoiar e facilitar o trabalho dos professores na criação de conteúdos instrucionais, apresentados dentro da Sala de Aula Virtual. Por meio do Editor de Aulas, o professor pode modelar a linha temporal de cenas executadas pelos avatares, além de criar atividades interativas aos alunos e inserir materiais de consulta [5]. O Editor de Aulas é um aplicativo *desktop* desenvolvido em C# para o ambiente *Windows*. A Figura 6.2 apresenta o Editor.

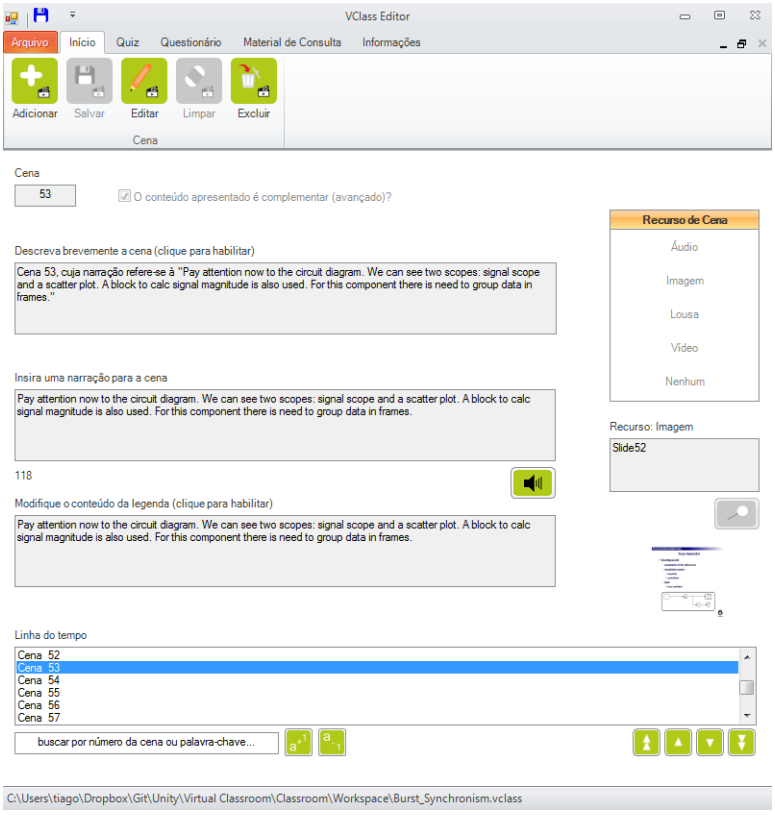


Figura 6.2: Editor de Aulas [5].

A Sala de Aula Virtual foi idealizada de forma a apoiar os alunos no processo de visualização dos conteúdos criados pelo Editor de Aulas. Trata-se de uma representação virtual de uma sala de aula, uma vez que possui grande parte de seus principais itens como lousas, projetores, computadores, telas de projeção, mesas, cadeiras, dentre outras. A Sala de Aula Virtual foi desenvolvida usando o *Unity3D* como *engine* de renderização e processamento gráfico, além da linguagem C# em forma de *script*. A Sala de Aula Virtual é um aplicativo *desktop*, mas que pode ser portado facilmente para as mais diversas plataformas, como *Web* e *Mobile*, por conta da flexibilidade proporcionada pela *Unity3d*. A Figura 6.3 apresenta a Sala de Aula Virtual.

Figura 6.3: Sala de Aula Virtual [5].

O avatar faz uso de sistemas TTS para narrar as aulas. O processo de geração da narração é feita no Editor de Aulas, que por sua vez envia os arquivos de áudio para a Sala de Aula Virtual. A comunicação entre o Editor de Aulas e o sintetizador de voz é feita por meio da interface SAPI5 (*Speech Application Programming Interface*) desenvolvida pela *Microsoft*. Sendo assim, somente o Editor de Aulas possui relação direta com o sintetizador de voz.

Por se tratar de uma aplicação educacional, onde o foco principal é transmitir conhecimento de maneira eficiente, a voz sintética deve ser inteligível, compreensível, ter boa pronúncia, não ser rápida nem lenta e ser natural. A voz deve soar agradável durante a aula e apresentar boa compreensão quando integrado a plataforma de ensino a distância. Também deve permitir que o conteúdo da aula seja transmitido de forma semelhante a uma aula presencial.

6.2.2 Metodologia do Experimento

O ambiente de ensino a distância usando avatares requer uma voz masculina e uma voz feminina. Desta forma, foram avaliados dois sintetizadores (cada qual com duas vozes feminina e duas vozes masculina, totalizando quatro vozes).

Por motivos legais, foi optado por preservar o real nome de cada sintetizador e voz. Desta forma, os sintetizadores serão referenciados como Sintetizador A e Sintetizador B, e as vozes como A1F (Voz feminina do sintetizador A), A1M (Voz masculina do sintetizador A), B1F

(Voz feminina do sintetizador B) e B1M (Voz masculina do sintetizador B). Além das 4 vozes sintéticas, utilizamos uma voz humana feminina para servir de controle, referenciada como Voz Humana.

Ao todo, 32 alunos, com idades variando entre 20 e 40 anos, todos de cursos de graduação e pós-graduação relacionados a computação e engenharia, responderam a um questionário *online* na plataforma *Google* Formulários. Foram selecionados somente indivíduos que tinham a língua portuguesa como língua materna, de forma a garantir a fluência no idioma na modalidade escrita e falada.

Cada ouvinte foi orientado quanto à metodologia dos testes e o questionário eletrônico não permitia que o ouvinte avançasse sem ter respondido todas as questões. Os ouvintes podiam escutar cada fala uma única vez. O formulário aplicado seguiu todas as orientações da metodologia proposta.

6.2.3 Resultados da Avaliação

A Figura 6.4 resume os resultados dos experimentos usando o método WER com sentenças SUS. Observa-se que a voz humana obteve a menor taxa de erro (1,15%), como esperado. Dentre as vozes sintéticas, as vozes A1M e B1M obtiveram uma taxa de erro semelhante, 1,51% e 1,60% respectivamente, enquanto a voz B1F atingiu 2,58% e a A1F obteve a maior taxa de erro dentre todas (6,35%). Apesar da voz A1F ter apresentado uma taxa de erro alta comparada com as outras vozes, pode-se dizer que todas as vozes apresentaram bons níveis de inteligibilidade.

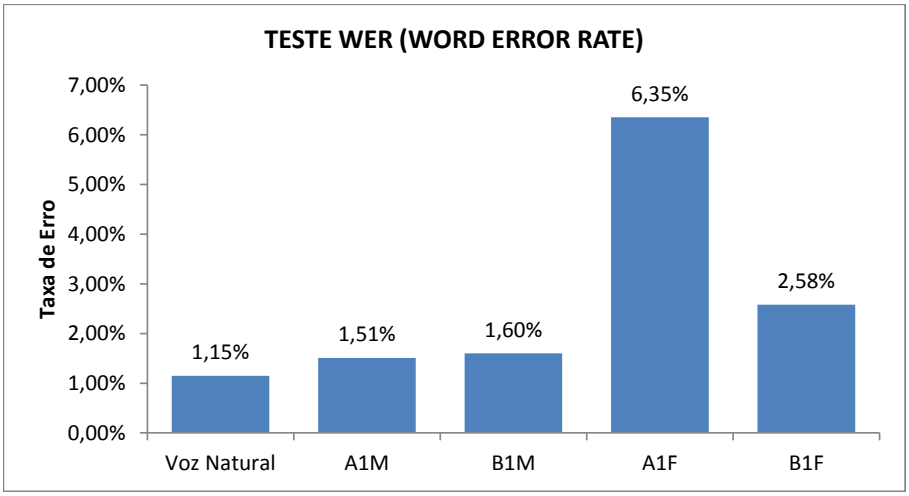


Figura 6.4: Resultados para o Teste WER.

A Figura 6.5 resume os resultados dos experimentos usando o método de avaliação de compreensibilidade. Observa-se que a voz humana obteve a maior taxa de compreensibilidade (95,31%), como esperado. Dentre as vozes sintéticas, as vozes A1M e A1F obtiveram a maior taxa de compreensibilidade, 95,31% e 77,34% respectivamente, enquanto a voz B1F atingiu 75% de compreensibilidade e a voz B1M obteve a pior taxa de compreensibilidade de todas (61,33%). Pode-se dizer que a voz A1M apresenta uma ótima taxa de compreensibilidade, A1F e B1F uma boa taxa de compreensibilidade e a B1M uma taxa de compreensibilidade razoável.

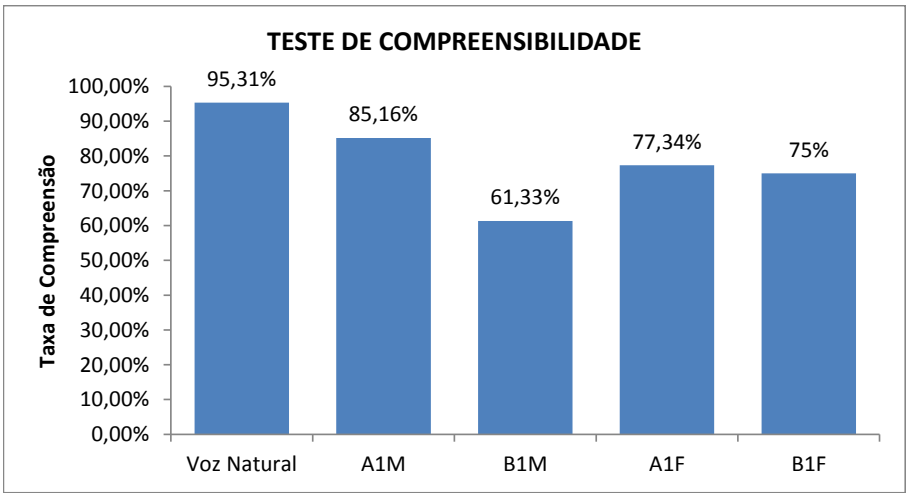


Figura 6.5: Resultados para o Teste de Compreensibilidade.

A Figura 6.6 resume os resultados para as características subjetivas: pronúncia, velocidade da fala e naturalidade. Em relação a pronúncia, a voz humana obteve a melhor avaliação (4,16), considerada boa, como esperado. As vozes A1M, B1M e B1F ficaram em torno da nota 3, considerada razoável, enquanto que a voz A1F ficou em torno de 2, considerada pobre.

Em relação a velocidade, a escala considera 1 como muito lento e 5 como muito rápido, de forma que 3 seja a velocidade ideal. Sendo assim, a voz humana foi a que obteve a melhor avaliação (3,19), considerada excelente, como esperado. No entanto, as vozes artificiais também ficaram em torno do 3, o que mostra que a velocidade de todas as vozes artificiais foram consideradas excelentes quanto a velocidade da pronúncia.

Em relação a naturalidade, a voz humana obteve a melhor avaliação (3,50), considerada razoável. Esperava-se uma nota entre 4 e 5 para a voz humana, mas características como sotaque, altura da voz, entonação, dentre outros, podem influenciar na naturalidade [8]. As vozes B1M e B1F ficaram em torno da nota 3, enquanto que as vozes A1M e A1F obtiveram, respectivamente, 2,88 e 1,88, consideradas pobre e mau, respectivamente.

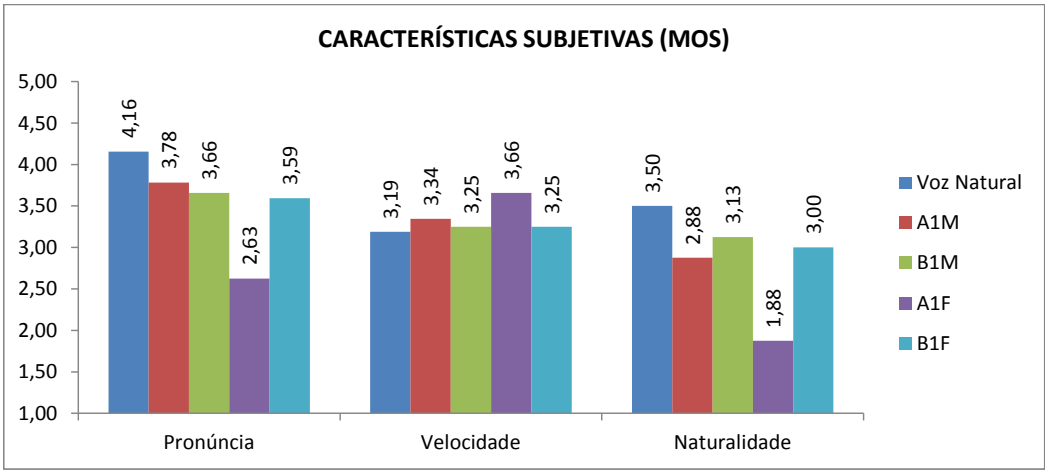


Figura 6.6: Resultados para as Avaliações Subjetivas.

Na avaliação em campo, perguntou-se aos alunos sobre a agradabilidade e a compreensão da voz do avatar. Para o teste, foram desenvolvidas 5 aulas na plataforma de ensino a distância, de aproximadamente 2,30 minutos cada, sobre assuntos ligados a computação e engenharia. Para cada aula, os ouvintes deram uma nota de 1 a 5 (escala MOS) para a agradabilidade e compreensão da voz do avatar. A Figura 6.7 resume os resultados da avaliação em campo.

Em relação a agradabilidade da voz do avatar, a voz humana obteve a melhor avaliação (4,13), considerada boa, como esperado. As vozes B1M e B1F obtiveram notas consideradas razoáveis, 3,34 e 3,53, respectivamente. Por outro lado, as vozes A1M e A1F obtiveram as piores avaliações, 2,69 e 2,66, respectivamente, consideradas pobres.

Em relação a compreensão da voz do avatar, a voz humana obteve a melhor avaliação (4,38), considerada boa, como esperado. As vozes B1M e B1F também obtiveram notas consideradas boas, 4,06 e 4,34, respectivamente. Por outro lado, as vozes A1M e A1F obtiveram notas consideradas razoáveis, 3,63 e 3,44, respectivamente.

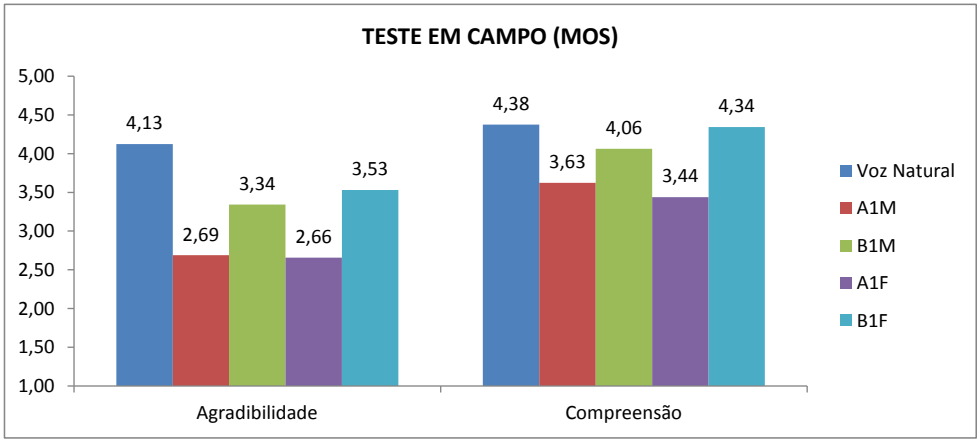


Figura 6.7: Resultados para as Avaliações em Campo.

Além das questões referentes a agradabilidade e compreensão da voz do avatar, também foi desenvolvida uma questão para cada aula. A questão era referente ao conteúdo da aula e tinha por objetivo medir a retenção do conhecimento. A Figura 6.8 resume os resultados da avaliação.

Podemos observar que a questão relativa a aula narrada com voz natural teve uma taxa de acerto considerada boa (81,82%), mas ainda assim, não foi o melhor resultado. Isso possivelmente se deve a dificuldade da questão. É difícil garantir que todas as questões sejam da mesma dificuldade, além do mais, o conhecimento entre os ouvintes pode variar, apesar de todos terem perfil acadêmico semelhante.

As vozes A1M, B1M e A1F obtiveram uma excelente taxa de acerto, 90,91%, 93,94% e 90,91%, respectivamente, enquanto que a voz B1F obteve a menor taxa de acerto (78,79%).

Também foi possível observar na prática, apesar de não ter sido documentado formalmente, que todas as aulas virtuais despertaram a curiosidade dos ouvintes e transmitiram conhecimento de maneira eficiente. Após a avaliação, diversos ouvintes pediram informações extras em relação ao conteúdo das aulas e dos sistemas TTS utilizados. Isso mostra que o ambiente de ensino usando avatares com vozes sintetizadas obteve sucesso na interação com os alunos e na transmissão de conhecimento.

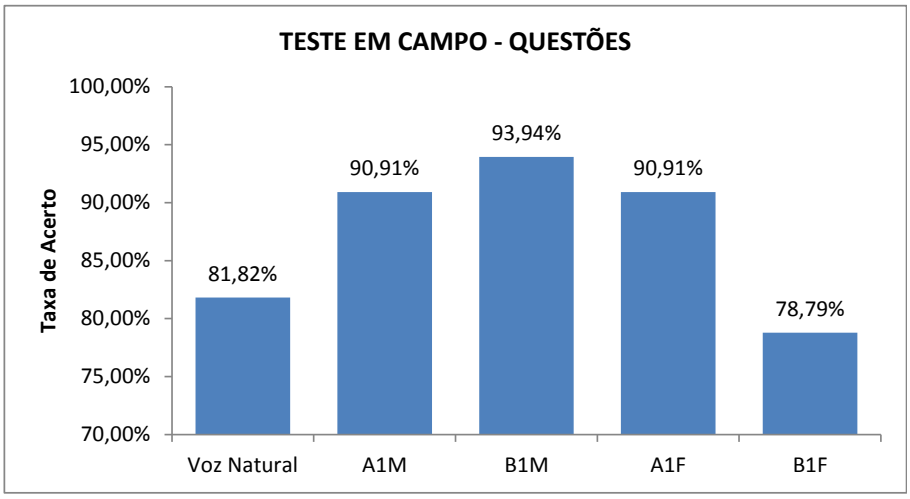


Figura 6.8: Resultados para as Avaliações em Campo - Questões.

Com base nas informações obtidas nas avaliações, foram identificadas as melhores vozes masculinas e femininas para cada avaliação. A Tabela 6.1 apresenta a compilação dos resultados.

Tabela 6.1: Compilação dos resultados (Melhores Vozes)

Método de Teste	**Voz Masculina**	**Voz Feminina**
WER	A1M (1,50)	B1F (2,57)
Compreensão	A1M (85,15%)	A1F (77,34%)
Pronúncia (MOS)	A1M (3,78)	B1F (3,59)
Velocidade da fala (MOS)	B1M (3,25)	B1F (3,25)
Naturalidade (MOS)	B1M (3,12)	B1F (3,00)
Agradabilidade da fala (Teste em Campo - MOS)	B1M (3,34)	B1F (3,53)
Compreensão da fala (Teste em Campo - MOS)	B1M (4,06)	B1F (4,34)
Questões Objetivas (Teste em Campo)	B1M (93,94%)	A1F (90,91%)

Com base nesses dados, podemos fazer a escolha da voz ideal levando em consideração as características vocais consideradas mais importantes. Como dito na Subseção 6.2.1, para a plataforma de ensino usando avatares, a voz, tanto masculina quanto feminina, deve ser inteligível, compreensível, ter boa pronúncia, não ser rápida nem lenta e ser natural. A voz deve soar agradável durante a aula e apresentar boa compreensão quando integrado a plataforma de ensino a distância. Também deve permitir que o conteúdo da aula seja transmitido de forma semelhante a uma aula presencial.

Sendo assim, a voz B1M (Masculina) do sintetizador A foi a escolhida por ter apresentado o melhor desempenho nos testes de velocidade da fala, naturalidade, agradabilidade (MOS - Teste em Campo), Compreensão (MOS - Teste em Campo), o que levou a uma taxa de 93,94% de acertos nas questões objetivas na avaliação em campo, enquanto que a voz B1F (Feminina) do sintetizador B foi a escolhida por ter apresentado o melhor desempenho nos testes WER, Pronúncia, Velocidade, Naturalidade, Agradabilidade (MOS - Teste em Campo), Compreensibilidade (MOS - Teste em Campo), o que levou a uma taxa de 79% de acertos nas questões objetivas na avaliação em campo.

6.3 Proposta de Automatização da Metodologia

O processo de avaliação de voz é extenso e repetitivo. A metodologia proposta utiliza métodos que podem ser passíveis de automatização. Uma plataforma que auxilie no preparo dos formulários de avaliação e na análise dos dados coletados é de fundamental importância por diminuir o tempo gasto no processo de busca da voz ideal e por diminuir a margem de erro no processo de avaliação dos dados.

Atualmente, com a popularização da *internet*, o processo de avaliação de voz pode ser feito de maneira descentralizada. No entanto, em situações onde se requer um maior controle, a avaliação deve ser feita em uma sala isolada, de maneira que todos os ouvintes sejam supervisionados durante o processo.

Pensando nas duas possibilidades, propomos uma arquitetura inicial de uma plataforma de avaliação de voz. A arquitetura, apesar de sua natureza distribuída, pode ser utilizada tanto em avaliações descentralizadas como em avaliações centralizadas. A Figura 6.9 apresenta a plataforma em alto nível.

Como podemos ver, o avaliador acessa a sua aplicação (*Web*) que está hospedada em um servidor, pelo seu computador pessoal. Por meio dele, o avaliador pode: (1) criar uma nova avaliação e (2) analisar os dados coletados.

Quando o avaliador cria um novo formulário de avaliação, ele define quais serão os métodos de avaliação a serem usados (os métodos são os que a metodologia proposta utiliza), gera as vozes no sistema TTS desejado e envia para o servidor. Em relação ao teste de inteligibilidade usando WER e sentenças SUS, a geração de sentenças deve ser feita automaticamente com base em um conjunto de palavras armazenadas em uma base de dados no servidor. A aplicação *Web* deve auxiliar na criação de todo o formulário.

Uma vez que o formulário esteja finalizado, o avaliador envia um *link* do formulário para os ouvintes poderem acessar de seu computador pessoal. Após os ouvintes responderem ao formulário, os dados são enviados novamente para o servidor, que efetua as análises pertinentes para cada método de avaliação. Em caso de uma avaliação centralizada, os ouvintes acessam o *link* do formulário na sala controlada pelo avaliador.

Após feita a avaliação, o avaliador, por meio de seu painel administrativo, poderá consultar todos os dados já analisados, a fim de escolher a voz ideal. O processo de escolha da voz ideal não deve ser automatizado, pois depende das características exigidas pelo projeto. Por fim, um relatório de toda a avaliação seguindo o *template* proposto deve ser emitido.

É importante que a plataforma de avaliação utilize métodos de *streaming* de áudio, para que os ouvintes não precisem baixar os arquivos contendo as fala sintetizadas. Também é importante que a interface do formulário de avaliação seja fácil e intuitiva, de forma a não necessitar de treinamento prévio.

O áudio reproduzido para o ouvinte deve ser exatamente igual ao áudio gerado pelo sistema TTS. Nenhum método de compressão com perdas deve ser utilizado pela plataforma. Por fim, a plataforma deve ser compatível com qualquer navegador *Web* e apresentar uma rápida navegação.

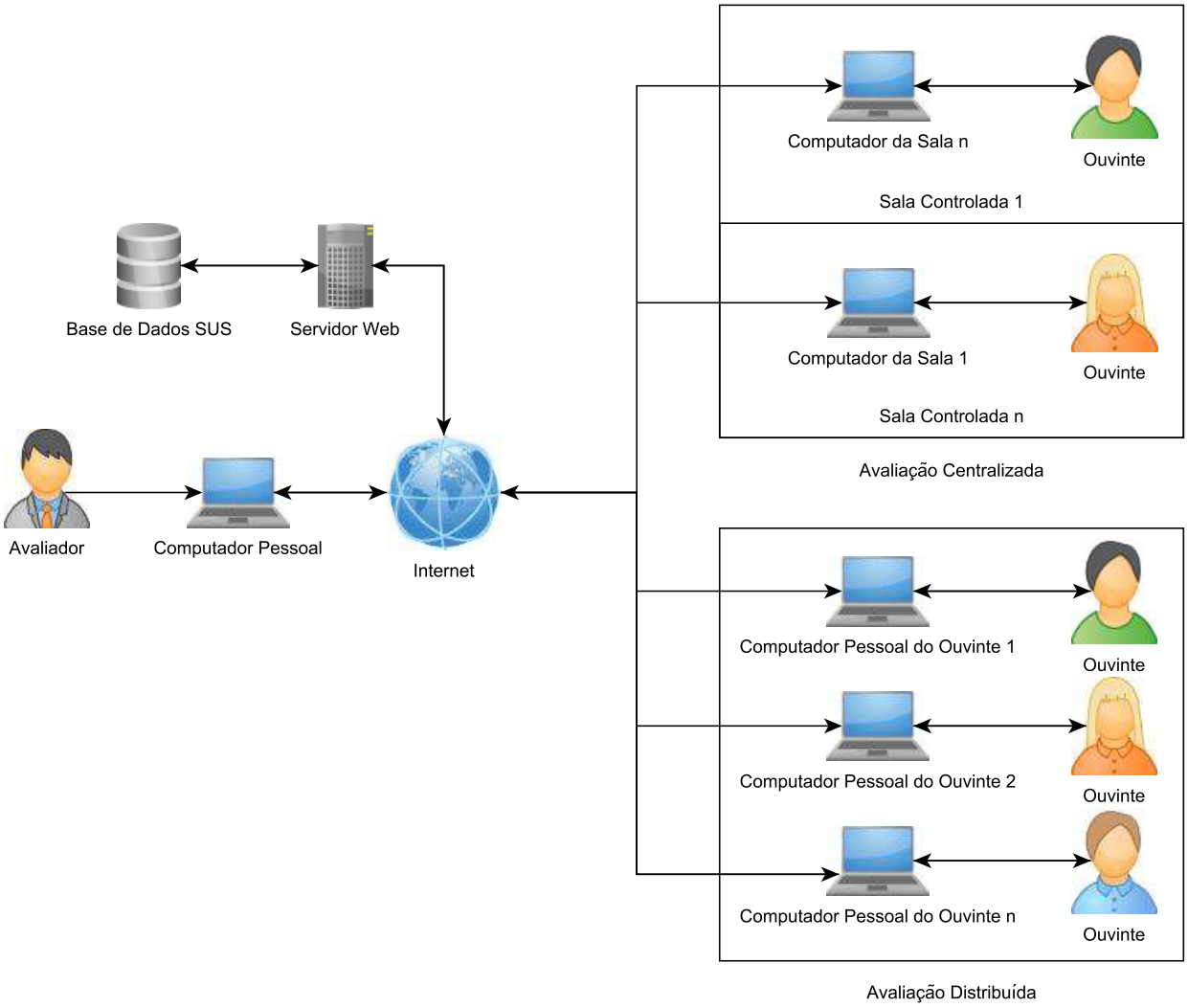


Figura 6.9: Proposta de Arquitetura para Automatização da Metodologia Proposta.

6.4 Considerações Finais

O processo de escolha de uma voz sintetizada é um processo longo e demorado. Atualmente, a literatura aborda diversos métodos de avaliação, mas são poucos os esforços que mostram como avaliar uma voz para um determinado projeto. Desta forma, este capítulo apresentou uma metodologia de avaliação, assim como uma proposta de automatização da metodologia, que busca auxiliar no processo de escolha de voz sintética para um determinado projeto. A metodologia foi testada em um projeto real e, segundo Cinto [5], a voz escolhida por meio da metodologia apresentou bons resultados na plataforma de ensino a distância. O *template* preenchido da avaliação pode ser encontrado no Apêndice desta dissertação.

Capítulo 7

Conclusão

Com o avanço do poder computacional, a síntese de fala artificial deixou os laboratórios de pesquisa e passou a fazer parte da vida das pessoas. Hoje, ela é encontrada nos mais diversos tipos de aplicações como sistemas ATM, leitor para cegos, brinquedos, dentre outras.

Um grande esforço da comunidade acadêmica foi necessário para a sua popularização. Na literatura, os primeiros indícios da tentativa de geração de fala artificial foram em 1779, com Christian Kratzenstein e seus aparatos mecânicos. Desde então, inúmeros pesquisadores trabalharam no desenvolvimento de métodos de geração de fala artificial.

Com o surgimento de novos sintetizadores, métodos de avaliação de fala artificial se tornaram necessários. Com isso, uma nova área de pesquisa surgiu, a de avaliação de voz sintetizada, na qual este trabalho se insere.

Este trabalho teve por objetivo apresentar uma metodologia de avaliação de voz sintética, assim como compreender as relações entre inteligibilidade, compreensibilidade e naturalidade.

A metodologia de avaliação de voz sintética surgiu da necessidade de encontrar uma voz ideal para uma plataforma de ensino a distância usando avatares. A metodologia proposta neste trabalho engloba as etapas de planejamento da avaliação, aplicação e análise dos dados. A metodologia foi posta em prática e as vozes selecionadas por ela satisfizeram os requisitos de voz do ambiente de ensino.

Quanto as relações entre inteligibilidade, compreensibilidade e naturalidade, chegamos a conclusão de que o ouvinte se sente seguro quanto a inteligibilidade e a compreensibilidade quando a naturalidade não está comprometida. Esta informação pode ser útil no desenvolvimento de novos sintetizadores e também no processo de avaliação de voz sintetizada, uma vez que os dados referentes à avaliação da inteligibilidade, quando medido por meio de testes subjetivos, como a escala MOS, podem ser decrementados por conta da naturalidade.

Por fim, uma arquitetura de avaliação de voz sintetizada é proposta, que tem por finalidade auxiliar no processo de preparação, aplicação e análise dos dados. A arquitetura surgiu devido a dificuldade encontrada em aplicar o questionário e em analisar os dados.

7.1 Trabalhos Futuros

Com base na experiência adquirida ao longo do projeto e também nas dificuldades encontradas, listamos as seguintes vertentes que podem ser trabalhadas futuramente:

- Analisar outras características além da naturalidade que podem influenciar na inteligibilidade e na compreensibilidade de forma a melhor compreender as características fundamentais que devem ser trabalhadas no desenvolvimento e na análise de sintetizadores de voz.
- Analisar a eficiência da metodologia propostas em outros tipos de projetos e, se necessário, fazer adaptações na metodologia.
- Desenvolver uma plataforma de avaliação de voz sintetizada, de forma a facilitar a criação da avaliação, a aplicação e a análise dos dados.
- Analisar a eficiência do uso de sistemas STT para avaliar a inteligibilidade de uma voz sintetizada e o impacto dos erros de reconhecimento dos sistemas STT na avaliação das vozes.
- Estudar formas de avaliação automática de fala de forma a diminuir o tempo e o custo das avaliações.

7.2 Trabalhos Publicados

Artigos Publicados em Revista

Cinto, T. ; **Leite, H. M. A.** ; Arantes, D. S. ; Oliveira Junior, H. P. ; Peixoto, C. S. A. Proposta de Editoração de Livros Acadêmicos Interativos WYSIWYM Baseada em XML. Revista Ciência e Tecnologia (RCT). ISSN 2236-6733, v. 16, p. 1-8, 2013.

Trabalhos Completos Publicados em Anais de Congressos

Leite, H. M. A.; Carvalho, S. N.; Cinto, T.; Arantes, D. S. Avaliação de Vozes Artificiais: Inteligibilidade, Compreensibilidade e Naturalidade. In: Computer on the Beach, 2014, Florianópolis. Anais do Computer on the Beach 2014, 2014. p. 144-153.

Cinto, T.; **Leite, H. M. A.**; Peixoto, C. S. A.; Arantes, D. S. Virtual 3D Learning Environments Using Avatars. In: The 3rd International Conference on E-Learning and E-Technologies in Education (ICEEE), 2014, Kuala Lumpur. Proceedings of The 3rd International Conference on E-Learning and E-Technologies in Education (ICEEE), 2014. p. 206-215.

Cinto, T.; **Leite, H. M. A.**; Peixoto, C. S. A.; Arantes, D. S. O Uso de Avatares Computacionais como Ferramenta de Instrução no Ensino à Distância. In: V Seminário Internacional de Educação a Distância: Meios, Atores e Processos, 2013, Belo Horizonte. Anais do V Seminário Internacional de Educação à Distância: Meios, Atores e Processos, 2013.

Moreira, V. R.; Cinto, T.; **Leite, H. M. A.**; Arantes, D. S. Aprimorando o Ensino de Engenharia com Novas Abordagens Usando Recursos Computacionais. In: VI Congresso Tecnológico InfoBrasil TI & Telecom, 2013, Fortaleza. Anais do VI Congresso Tecnológico InfoBrasil TI &Telecom, 2013.

Resumos Expandidos Publicados em Anais de Congressos

Leite, H. M. A.; Cinto, T.; Peixoto, C. S. A.; Arantes, D. S. Ambientes Virtuais de Aprendizagem com Uso de Avatares e Vozes Sintetizadas. In: V Seminário Internacional de Educação a Distância: Meios, Atores e Processos, 2013, Belo Horizonte. Anais do V Seminário Internacional de Educação à Distância: Meios, Atores e Processos, 2013.

Artigos Convidados (Aguardando Aprovação)

Leite, H. M. A.; Cinto, T.; Carvalho, S. N.; Peixoto, C. S. A.; Arantes, D. S. Uso de Vozes Sintetizadas em Ambientes Virtuais de Ensino a Distância. Revista Ciência e Tecnologia (RCT). ISSN 2236-6733, 2014.

Cinto, T.; **Leite, H. M. A.**; Peixoto, C. S. A.; Arantes, D. S. Virtual Learning Environments: Proposals for Authoring and Visualization of Educational Content. International Journal of Digital Information and Wireless Communications (IJDIWC). ISSN 2225-658X, 2014.

Bibliografia

[1] S. Lemmetty. Review of speech synthesis technology. Master's thesis, Helsinki University of Technology, 1999.

[2] J. L. Flanagan, J. B. Allen, and M. A. H. Johnson. *Speech Analysis, Synthesis, and Perception*. Springer-Verlag, New York, 1972.

[3] H. Dudley, R. R. Riesz, and S. S. A. Watkins. A synthetic speaker. *Journal of The Franklin Institute*, 227(6):739–764, 1939.

[4] D. Klatt. Review of text-to-speech conversion for english. *Journal of the Acoustical Society of America (JASA)*, 82(3):737–793, 1987.

[5] T. Cinto. Ambientes virtuais de aprendizagem: Propostas de editoração e visualização de conteúdo educacional para aulas presenciais e online. Master's thesis, Universidade Estadual de Campinas, 2014.

[6] U. Jekosch. Speech quality assessment and evaluation. In *Proceedings of Eurospeech*, pages 1387–1394, 1993.

[7] V. Kraft and T. Portele. Quality evaluation of five german speech synthesis systems. *Acta Acustica*, 3:351–365, 1995.

[8] Y. Chang. Evaluation of tts systems in intelligibility and comprehension tasks. In *Proceedings of the 23rd Conference on Computational Linguistics and Speech Processing*, pages 64–78. Association for Computational Linguistics, 2011.

[9] G. E. Moore. Cramming more components onto integrated circuits. *Proceedings of the IEEE*, 86(1):82–85, 1998.

[10] J. Grudin. A moving target - the evolution of human-computer interaction. In *Human-Computer Interaction Handbook (3rd Edition)*, 2012.

[11] A. Chapanis. Interactive human communication. In *Computer-supported Cooperative Work: A Book of Readings*, pages 127–140. Morgan Kaufmann Publishers Inc., 1988.

[12] H. M. A. Leite, S. N. Carvalho, T. Cinto, and D. S. Arantes. Avaliação de vozes artificiais: Inteligibilidade, compreensibilidade e naturalidade. *Anais do Computer on the Beach 2014*, pages 144–153, 2014.

[13] H. M. A. Leite, T. Cinto, C. S. A. Peixoto, and D. S. Arantes. Ambientes virtuais de aprendizagem com uso de avatares e vozes sintetizadas. *Anais do V Seminário Internacional de Educação a Distância: Meios, Atores e Processos*, pages 1321–1326, 2013.

[14] V. R. Moreira, T. Cinto, H. M. A. Leite, and D. S. Arantes. Aprimorando o ensino de engenharia com novas abordagens usando recursos computacionais. *Anais do VI Congresso Tecnológico InfoBrasil TI & Telecom*, 2013.

[15] T. Cinto, H. M. A. Leite, C. S. A. Peixoto, and D. S. Arantes. O uso de avatares computacionais como ferramenta de instrução no ensino a distância. *Anais do V Seminário Internacional de Educação a Distância: Meios, Atores e Processos*, pages 858–869, 2013.

[16] T. Cinto, H. M. A. Leite, C. S. A. Peixoto, and D. S. Arantes. Virtual 3d learning environments using avatars. *Proceedings of The 3rd International Conference on E-Learning and E-Technologies in Education (ICEEE)*, pages 206–215, 2014.

[17] M. R. Schroeder. A brief history of synthetic speech. *Speech Commun.*, 13(1-2):231–237, 1993.

[18] W. Willis. On vowel sounds, and on reed-organ pipes. *Trans. Camb. Phil. Soc. III*, pages 1–38, 1838.

[19] W. B. Kleijn and K. K. Paliwal. *Speech Coding and Synthesis*. Elsevier Science, 1995.

[20] J. Santen, R. Sproat, J. Olive, and J. Hirschberg. *Progress in Speech Synthesis*. Springer-Verlag New York Inc., 1997.

[21] F. Chen and K. Jokinen. *Speech Technology*. Springer Science+Business Media, 2010.

[22] S. N. Carvalho. Estudo de um sistema de conversão texto-fala baseado em hmm. Master's thesis, Universidade Estadual de Campinas, 2013.

[23] B. Kroger. Minimal rules for articulatory speech synthesis. In *Proceedings of EUSIPCO92*, volume 1, pages 331–334. Elsevier Science Publisher, 1992.

[24] Pertti Palo. A review of articulatory speech synthesis. Master's thesis, Helsinki University of Technology, 2006.

[25] O. Engwall. Assessing mri measurements: Effects of sustentation, gravitation and coarticulation. In *Speech Production: Models, Phonetic Processes And Techniques*, pages 301–314. Psychology Press, 2006.

[26] J. Mullen, D. M. Howard, and D. T. Murphy. Waveguide physical modeling of vocal tract acoustics: flexible formant bandwidth control from increased model dimensionality. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(3):964–971, 2006.

[27] S. Aryal and R. Gutierrez-Osuna. Articulatory inversion and synthesis: Towards articulatory-based modification of speech. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 7952–7956, 2013.

[28] R. Donovan. *Trainable Speech Synthesis*. PhD thesis, Cambridge University, 1996.

[29] J. Allen, S. Hunnicutt, and D. Klatt. *From Text to Speech: The MITalk System*. Cambridge University Press, New York, 1987.

[30] W. J. Holmes, J. N. Holmes, and M. W. Judd. Extension of the bandwidth of the jsru parallel-formant synthesizer for high quality synthesis of male and female speech. In *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*, pages 313–316, 1990.

[31] D. Klatt. Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America (JASA)*, 67(3):971–995, 1980.

[32] U. K. Laine. Parcas, a new terminal analog model for speech synthesis. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '82*, pages 940–943, 1982.

[33] S. S. Silva. Um estudo de modelos básicos de prosódia para o português brasileiro. Master's thesis, Universidade Federal do Rio de Janeiro, 2004.

[34] S. A. Toma, G. I. Tarsa, E. Oancea, D. Munteanu, F. Totir, and L. Anton. A td-psola based method for speech synthesis and compression. In *Communications (COMM), 2010 8th International Conference on*, pages 123–126, 2010.

[35] F. Violaro and O. Boeffard. A hybrid model for text-to-speech synthesis. *Speech and Audio Processing, IEEE Transactions on*, 6(5):426–434, 1998.

[36] R. Kortekaas and A. Kohlrausch. Psychoacoustical evaluation of the pitch-synchronous overlap-and-add speech-waveform manipulation technique using single-formant stimuli. *Journal of the Acoustical Society of America (JASA)*, 101(4):2202–2213, 1997.

[37] E. Moulines and F. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. In *Proceedings os Eurospeech 89*, pages 453–467, 1990.

[38] H. Valbret, E. Moulines, and J. P. Tubach. Voice transformation using psola technique. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, pages 145–148, 1992.

[39] K. Tokuda, T. Kobayashi, and S. Imai. Speech parameter generation from hmm using dynamic features. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, pages 660–663, 1995.

[40] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi. Adaptation of pitch and spectrum for hmm-based speech synthesis using mllr. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on*, pages 805–808, 2001.

[41] K. Shichiri, A. Sawabe, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. Eigenvoices for hmm-based speech synthesis. In *Proceedings of EUROSPEECH 2002*, pages 1269–1272, 2002.

[42] M. Tachibana, J. Yamagishi, T. Masuko, and T. Kobayashi. Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing. *IEICE - Trans. Inf. Syst.*, E88-D(11):2484–2491, 2005.

[43] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. Speaker interpolation in hmm-based speech synthesis system. In *Proceedings of EUROSPEECH 1997*, pages 2523–2526, 1997.

[44] T. Nose, J. Yamagishi, and T. Kobayashi. A style control technique for speech synthesis using multiple regression hsmm. In *Proceedings of Interspeech 2006*, pages 1324–1327, 2006.

[45] T. Yoshimura. *Simultaneous Modeling of Phonetic and Prosodic Parameters, and Characteristic Conversion for HMM-Based Text-to-Speech Systems*. PhD thesis, Nagoya Institute of Technology, 2002.

[46] H. Banno, H. Hata, M. Morise, T. Takahashi, T. Irino, and H. Kawahara. Implementation of realtime straight speech manipulation system: Report on its first implementation. *The Acoustical Society of Japan*, 28(3):140–146, 2007.

[47] T. Toda and K. Tokuda. A speech parameter generation algorithm considering global variance for hmm-based speech synthesis. *IEICE - Trans. Inf. Syst*, pages 2801–2804, 2007.

[48] T. Toda. Modeling of speech parameter sequence considering global variance for hmm-based speech synthesis. In *Hidden Markov Models, Theory and Applications*, pages 131–150. InTech, 2011.

[49] A. Mariniak. A global framework for the assessment of synthetic speech without subjects. In *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 1683–1686, 1993.

[50] J. Logan, B. Greene, and D. Pisoni. Segmental intelligibility of synthetic speech produced by rule. *Journal of the Acoustical Society of America, JASA*, 86(2):566–581, 1989.

[51] M. Goldstein. Classification of methods used for assessment of text-to-speech systems according to the demands placed on the listener. *Speech Commun.*, 16(3):225–244, 1995.

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

[53] D. B. Pisoni and S. Hunnicutt. Perceptual evaluation of mitalk: The mit unrestricted text-to-speech system. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '80.*, pages 572–575, 1980.

[54] P. W. Nye and J. H. Gaitenby. The intelligibility of synthetic monosyllabic words in short, syntactically normal sentences. *Haskins Laboratories Status Report on Speech Research*, pages 169–190, 1974.

[55] C. Benoit, M. Grice, and V. Hazan. The sus test: A method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences. *Speech Communication*, 18(4):381–392, 1996.

[56] P. B. Mareüil, B. Philippe, C. Alessandro, A. Raake, G. Bailly, M. Garcia, and M. Morel. Lrec. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '80.*, 2006.

[57] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics - Doklady*, 10(8):707–710, 1966.

[58] J. Bernstein and D. B. Pisoni. Unlimited text-to-speech system: Description and evaluation of a microprocessor based device. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '80.*, volume 5, pages 576–579, 1980.

[59] H. Klaus, H. Klix, J. Sotscheck, and K. Fellbaum. An evaluation system for ascertaining the quality of synthetic speech based on subjective category rating tests. In *EUROSPEECH*, volume 3, pages 1679–1682, 1993.

[60] J. Beskow, M. Dahlquist, B. Granstrom, M. Lundeberg, K. Spens, and T. Ohman. The teleface project - disability, feasibility and intelligibility. In *In proceedings of Fonetik 97*, 1997.

[61] J. Beskow, K. O. E. Elenius, and S. McGlashan. Olga - a dialogue system with an animated talking agent. *TMH-QPSR*, 38(2-3):1–6, 1997.

[62] B. L. Goff and C. Benoit. A text-to-audiovisual-speech synthesizer for french. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 4, pages 2163–2166, 1996.

Capítulo 8

Apêndice A

8.1 Grupo de Palavras para o Método MRT

Consoante Inicial

Tabela 8.1: Conjunto de Palavras para o Método MRT - Consoante Inicial

A	**B**	**C**	**D**	**E**	**F**
Pato	Gato	Rato	Tato	Fato	Mato
Achar	Bichar	Duchar	Flechar	Fechar	Fichar
Sogra	Flora	Fora	Hora	Mora	Nora
Dado	Gado	Brado	Bordo	Lodo	Adô
Duro	Furo	Juro	Ouro	Puro	Touro
Abar	Abrar	Afar	Alar	Alçar	Mixar
Abril	Barril	Canil	Viril	Quadril	Funil
Anel	Mel	Cartel	Rímel	Cordel	Cruel
Manta	Gata	Mata	Porta	Pata	Ata
Jornal	Aral	Sinal	Astral	Banal	Nasal
Foca	Arca	Banca	Barca	Bica	Boca
Manga	Alga	Larga	Briga	Carga	Chega
Tecla	Bula	Bala	Clara	Mala	Bola
Fada	Cada	Moda	Roda	Frota	Rota
Credor	Ator	Odor	Amor	Autor	Bolor
Cofre	Abre	Corre	Bagre	Chifre	Cobre
Barco	Coco	Oco	Louco	Rouco	Ronco
Alça	Rosa	Tosa	Asa	Balsa	Mesa
Filtros	Afros	Puros	Arcos	Astros	Bairros
Leite	Lente	Mente	Goste	Pente	Rente
Raio	Fio	Rio	Freio	Áudio	Frio
Tese	Classe	Crise	Case	Use	Frise
Baixa	Buxa	Taxa	Coxa	Fixa	Flexa
Piso	Riso	Osso	Grosso	Liso	Bolso
Varal	Canal	Bucal	Normal	Casal	Coral

Consoante Final

Tabela 8.2: Conjunto de Palavras para o Método MRT - Consoante Final

A	**B**	**C**	**D**	**E**	**F**
Balão	Balsa	Banal	Bambu	Banca	Banco
Café	Cacto	Caju	Calça	Caixa	Calar
Dança	Dama	Dado	Data	Dano	Danar
Fácil	Fada	Face	Fala	Falha	Falta
Galpão	Gambá	Galão	Gaita	Gado	Gabar
Hora	Horror	Horta	Honra	Homem	Hoje
Jogar	Jogo	Joia	Jóquei	Jovem	Jornal
Ladrão	Lado	Lago	Lama	Laço	Laje
Manga	Mansão	Manso	Mapa	Marco	Março
Nariz	Nascer	Nata	Nado	Nação	Narrar
Pagar	País	Padrão	Painel	Paixão	Palco
Quadra	Quadril	Quadro	Qualquer	Quando	Quanto
Raça	Ração	Rachar	Raio	Raiva	Raiz
Saldo	Salão	Sala	Safra	Safar	Salmão
Taça	Tacha	Talco	Talher	Talvez	Tampar
Vale	Valer	Vaia	Vago	Vagem	Vagão
Quintal	Quinto	Quinze	Quite	Quina	Quilo
Cofre	Colar	Colcha	Colchão	Coisa	Cola
Demais	Deitar	Degrau	Dele	Dentro	Depois
Feixe	Feitor	Fera	Férias	Feliz	Ferro
Moda	Modo	Moer	Molho	Moral	Morar
Tela	Tecla	Teimar	Tédio	Tema	Temer
Chover	Chuchu	Chutar	Chuva	Chute	Chumbo
Pires	Piso	Pique	Pipa	Pino	Pilão
Rubi	Ruga	Rugir	Ruir	Rumor	Rural

8.2 Sentenças SUS

1. A porta superou os poemas verdes
2. O advogado nadou na montanha das águas
3. A televisão suave ultrapassou o carro
4. O computador quebrou o rio
5. A escada queria comer alfândegas estranhas
6. O caderno pedalou até a lua
7. O patrão batucou na janela de barro
8. A camiseta ganhou um fone de gramado
9. O celular andou nas águas altas
10. A caderneta simulou bolas quadradas
11. O supermercado subiu a montanha azul
12. A estatística mergulhou na porta de água
13. O carro mergulhou no alto da montanha
14. A mesa gritou para o leão
15. O monitor digitou na água
16. A cadeira navegou na nuvem
17. O brinquedo pilotou um avião no mar
18. A escola quebrou a nuvem
19. O carro cantou no alto da maçã
20. A mesa gritou para o coelho
21. O rinoceronte caiu da nuvem verde
22. A tampa conversou com o teclado
23. O computador mergulhou a caneta no armário
24. A cratera digitou um desenho
25. A goiaba bebeu uma parede
26. O prego cantou um carro

27. A pedra escreveu uma impressora
28. O javali falou com a carteira
29. A salada andou no mar
30. O dinheiro saiu da lua
31. A aliança cantou a água
32. O tomate subiu a escada
33. O computador ligou para a árvore
34. A figura colou as nuvens
35. O cachorro comeu o avião
36. A pipa nadou no lago
37. O feijão comeu o arroz
38. A menina prendeu a vassoura
39. O alpinista mergulhou na terra
40. O advogado superou os poemas amarelos
41. A mesa quebrou o rio
42. O leite nadou nas alturas
43. A cama mastigou o jardim
44. O brigadeiro pediu uma bola quadrada
45. A cadeira andou no lago
46. O livro simulou uma janela de barro
47. O monitor sorriu para o elefante
48. A matemática escalou a água
49. O restaurante comeu montanhas ao molho
50. A faxineira limpou as nuvens
51. O jogador caiu no teto
52. A calça queimou os televisores
53. O abajur rolou pelo mar

8.3 Dados Utilizados para Encontrar as Relações entre Inteligibilidade, Compreensibilidade e Naturalidade

Palavras MRT para Avaliar a Voz Humana

Tabela 8.3: Palavras usadas no Método MRT - Consoante Inicial

A	B	C	D	E	F
Abar	Abrar	Afar	Alar	Alçar	Mixar
Fada	Cada	Moda	**Roda**	Frota	Rota
Leite	**Lente**	Mente	Goste	Pente	Rente
Piso	Riso	Osso	Grosso	Liso	Bolso

Tabela 8.4: Palavras usadas no Método MRT - Consoante Final

A	B	C	D	E	F
Fácil	Fada	**Face**	Fala	Falha	Falta
Quadra	Quadril	Quadro	Qualquer	Quando	Quanto
Cofre	**Colar**	Colcha	Colchão	Coisa	Cola
Rubi	Ruga	**Rugir**	Ruir	Rumor	Rural

Palavras MRT para Avaliar o Sintetizador 1

Tabela 8.5: Palavras usadas no Método MRT - Consoante Inicial

A	B	C	D	E	F
Dado	Gado	**Brado**	Bordo	Lodo	Adô
Foca	Arca	Banca	Barca	Bica	**Boca**
Cofre	Abre	**Corre**	Bagre	Chifre	Cobre
Baixa	Buxa	**Taxa**	Coxa	Fixa	Flexa

Tabela 8.6: Palavras usadas no Método MRT - Consoante Final

A	B	C	D	E	F
Dança	Dama	Dado	**Data**	Dano	Danar
Manga	Mansão	Manso	Mapa	Marco	Março
Quintal	**Quinto**	Quinze	Quite	Quina	Quilo
Tela	Tecla	Teimar	Tédio	**Tema**	Temer

Palavras MRT para Avaliar o Sintetizador 2

Tabela 8.7: Palavras usadas no Método MRT - Consoante Inicial

A	B	C	D	E	F
Pato	Gato	**Rato**	Tato	Fato	Mato
Abril	Barril	Canil	Viril	Quadril	**Funil**
Jornal	Aral	Sinal	Astral	Banal	Nasal
Piso	Riso	Osso	Grosso	Liso	**Bolso**

Tabela 8.8: Palavras usadas no Método MRT - Consoante Final

A	B	C	D	E	F
Fácil	Fada	Face	**Fala**	Falha	Falta
Hora	Horror	**Horta**	Honra	Homem	Hoje
Taça	Tacha	Talco	Talher	Talvez	Tampar
Feixe	Feitor	Fera	**Férias**	Feliz	Ferro

Senteças SUS usadas para Avaliar a Voz Humana

- O carro mergulhou no alto da montanha
- A mesa gritou para o leão
- O monitor digitou na água
- A cadeira navegou na nuvem
- O brinquedo pilotou um avião no mar
- A escola quebrou a nuvem

Senteças SUS usadas para Avaliar o Sintetizador 1

- A porta superou os poemas verdes
- O advogado nadou na montanha das águas
- A televisão suave ultrapassou o carro
- O computador quebrou o rio
- A escada queria comer alfândegas estranhas
- O caderno pedalou até a lua

Senteças SUS usadas para Avaliar o Sintetizador 2

- O patrão batucou na janela de barro
- A camiseta ganhou um fone de gramado
- O celular andou nas águas altas
- A caderneta simulou bolas quadradas
- O supermercado subiu a montanha azul
- A estatística mergulhou na porta de água

Questões usadas para Avaliar a Voz Humana

> "A tecnologia NFC está se mostrando cada vez mais útil no dia a dia e agora mais uma novidade está chamando a atenção: o pagamento de passagens de ônibus, trens, barcas e vans legalizadas apenas com o smartphone"
> Questão: Como é feito o pagamento através da tecnologia NFC?

> "A loja online Apple Store saiu do ar na manhã desta terça-feira (22), poucas horas antes do evento de apresentação dos novos produtos da empresa. Marcada para as 15h, horário de Brasília, a conferência deve marcar a chegada dos novos iPad 5, iPad mini 2, MacBooks Pro e Air, além do anúncio da disponibilidade do novo Mac OS X"
> Questão: Que horas a Apple vai apresentar os novos produtos?

> "O Instagram é uma rede social gratuita para compartilhamento de fotos e agora também vídeos otimizado para o novo iOS 7. Com ele, é possível aplicar filtros em suas imagens e filmagens e depois publicá-las em seu perfil, onde seus amigos podem visualizá-las, curtí-las e comentá-las"
> Questão: O que é possível compartilhar no Instagram?

Questões usadas para Avaliar o Sintetizador 1

> "Físicos americanos revelaram, nesta quinta-feira (22), a criação do relógio atômico experimental mais preciso do mundo, com variação inferior a um segundo em 13,8 bilhões de anos, a idade estimada do Universo"
> Questão: Qual a variação do relógio atômico mais preciso do mundo?

"A Agência Espacial Europeia (ESA, na sigla em inglês) vai utilizar as impressoras 3D para criar peças de metal para aeronaves. Durante evento no Museu de Ciência de Londres, a entidade apresentou nesta terça-feira (15) o projeto Amaze, que irá aperfeiçoar a imprissão de peças dentro dos próximos cinco anos"
Questão: O que a ESA vai imprimir na impressora em 3D?

"Às vésperas de o Twitter estrear na Bolsa de Valores, o microblog tem deixado de ser a plataforma de rede social preferida para resolver problemas de clientes pelas empresas, que preferem o Facebook para isso, apontou uma pesquisa da Direct Talk, que desenvolve sistemas de atendimento digital ao consumidor"
Questão: Qual a plataforma preferida pelas empresas para interagir com seus clientes?

Questões usadas para Avaliar o Sintetizador 2

"A média da velocidade da internet no Brasil cresceu 11%, para 2,4 megabits por segundo (Mbps) no trimestre entre abril e junho de 2013, segundo pesquisa divulgada nesta quarta-feira (16) pela empresa de internet Akamai. O resultado coloca o Brasil na 80 posição na lista de países analisados. Na média anual, a velocidade da internet do país cresceu 15%"
Questão: Qual a velocidade média da internet brasileira?

"A pesquisa anual da fabricante de equipamentos de rede Ericsson sobre o consumo de serviços de telecom do brasileiro, revelou que 73% das pessoas no país assistem TV interagindo com outros usuários em redes sociais. Esse número era de 48% na pesquisa 2011, o que representa uma alta de 25%"
Questão: O que a pesquisa anual da fabricante de equipamentos de rede Ericsson revelou?

"A Nokia Siemens Networks vai abrir uma linha de montagem no Brasil, junto com a fabricante Flextronics International, para construir a próxima geração de redes de telefonia móvel, disse um executivo sênior em entrevista na segunda-feira (27)"
Questão: Que tipo de produto a Nokia Siemens Networks vai construir no Brasil?

Sentenças Usadas para Avaliar a Naturalidade e a Inteligibilidade

- Sentença usada para avaliar a voz humana: Oi, meu nome é Bill. Você me escutou falando diversas palavras e frases enquanto respondia o questionário de avaliação. Foi um prazer falar com você. Muito obrigado.
- Sentença usada para avaliar o sintetizador 1: Oi, meu nome é João. Você me escutou falando diversas palavras e frases enquanto respondia o questionário de avaliação. Foi um prazer falar com você. Muito obrigado.
- Sentença usada para avaliar o sintetizador 2: Oi, meu nome é Fernanda. Você me escutou falando diversas palavras e frases enquanto respondia o questionário de avaliação. Foi um prazer falar com você. Muito obrigado.

8.4 Dados Utilizados para Validar a Metodologia Proposta

Senteças SUS usadas para Avaliar a Voz Humana

- O carro cantou no alto da maçã
- A mesa gritou para o coelho
- O rinoceronte caiu da nuvem verde
- A tampa conversou com o teclado
- O computador mergulhou a caneta no armário
- A cratera digitou um desenho
- A goiaba bebeu uma parede

Senteças SUS usadas para Avaliar o Sintetizador A (Voz A1M)

- O computador ligou para a árvore
- A figura colou as nuvens
- O cachorro comeu o avião
- A pipa nadou no lago
- O feijão comeu o arroz
- A menina prendeu a vassoura
- O alpinista mergulhou na terra

Senteças SUS usadas para Avaliar o Sintetizador A (Voz A1F)

- O prego cantou um carro
- A pedra escreveu uma impressora
- O javali falou com a carteira
- A salada andou no mar
- O dinheiro saiu da lua
- A aliança cantou a água
- O tomate subiu a escada

Senteças SUS usadas para Avaliar o Sintetizador B (Voz B1M)

- O advogado superou os poemas amarelos
- A mesa quebrou o rio
- O leite nadou nas alturas
- A cama mastigou o jardim
- O brigadeiro pediu uma bola quadrada
- A cadeira andou no lago
- O livro simulou uma janela de barro

Senteças SUS usadas para Avaliar o Sintetizador B (Voz B1F)

- O monitor sorriu para o elefante
- A matemática escalou a água
- O restaurante comeu montanhas ao molho
- A faxineira limpou as nuvens
- O jogador caiu no teto
- A calça queimou os televisores
- O abajur rolou pelo mar

Questões usadas para Avaliar a Voz Humana

"O Google liberou nesta quarta-feira, 10, o primeiro kit para desenvolvimento do seu "celular de montar". O Projeto Ara, criado pela Motorola no ano passado quando a empresa americana ainda pertencia ao Google, planeja smartphones com módulos que podem ser substituídos pelos usuários se ele quiser uma câmera, um processador melhor ou uma bateria extra, em uma experiência totalmente customizável"
Questão: Quais empresas estão envolvidas no projeto Ara?

"Quem pretende realizar o Exame Nacional do Ensino Médio (Enem) em 2014 já pode começar a se organizar para os simulados da prova. A prova é realizada em dois dias e, segundo professores e especialistas, é preciso treino para se sair bem"
Questão: Em quantos dias será realizada a prova do Enem?

"Para os notívagos de plantão, a madrugada desta terça-feira, 15, terá um espetáculo à parte. Um eclipse total da Lua poderá ser visto em toda a América a partir das 3 horas, horário de Brasília. O pesquisador do Observatório Nacional, Jair Barroso, explica que a visibilidade do fenômeno dependerá das condições do céu. "É importante que não tenham nuvens a ponto de impedir a visão"
Questão: Que evento ocorrerá na madrugada desta terça-feira?

"Usuários do Gmail acusaram o Google de violação de privacidade e de ignorar leis federais e estaduais ao analisar suas mensagens para finalidades publicitárias. O Google argumentou que os usuários implicitamente consentiram com essa prática, reconhecendo que esta era parte do processo de entrega de emails"
Questão: Porque os usuários do Gmail acusaram a Google?

Questões usadas para Avaliar o Sintetizador A (A1M)

"A Universidade de Coimbra, em Portugal, será a primeira instituição estrangeira a usar a nota do Exame Nacional do Ensino Médio (Enem) como critério de acesso ao ensino superior a partir de 2014. A medida só valerá para candidatos brasileiros. O Enem já é usado como parte do processo seletivo de todas as universidades federais do Brasil desde o ano passado"
Questão: A partir de quando Coimbra usará as notas do Enem?

"Para celebrar o Dia da Terra, a Agência Espacial Americana (Nasa) divulgou nesta terça-feira, 22, uma imagem do planeta azul. Na foto, capturada por um satélite, é possível ver as Américas e desvendar as condições climáticas do planeta"
Questão: Como a Nasa celebrou o dia da terra?

"A existência de vida fora da Terra, um dos maiores mistérios para os seres humanos, nunca esteve tão próxima de se confirmar. Cientistas anunciaram nesta quinta-feira, 17, a descoberta do primeiro planeta fora do sistema solar de tamanho similar ao da Terra e onde pode haver água em estado líquido na superfície. Isso significa que o planeta pode ser habitável"
Questão: O que os cientistas descobriram?

"Trata-se de uma forma de carbono, um condutor de eletricidade e calor melhor do que qualquer outro. E não é apenas o material mais duro do mundo, como também um dos mais flexíveis. O grafeno poderá revolucionar a indústria eletrônica com a produção de aparelhos flexíveis, computadores quânticos superpoderosos, roupas eletrônicas e computadores capazes de fazer interface com as células do nosso corpo"
Questão: Qual materia revolucionará a indústria eletrônica?

Questões usadas para Avaliar o Sintetizador A (A1F)

"Quando cursava publicidade, um professor falou para eu abrir um blog, que estava deixando de ser só diário virtual. No mesmo dia, fiz em casa um no Blogspot. Cresci primeiro regionalmente. Cheguei a São Paulo já grande"
Questão: Quem deu a idéia para a aluna abrir o blog?

"O limite anual para abatimento de gastos com educação no Imposto de Renda já se aproxima do valor de uma única mensalidade nas principais escolas do Brasil. Na declaração deste ano, pode-se deduzir até R$ 3.230 da base de cálculo do tributo, considerando apenas esse benefício"
Questão: Qual o limite anual para abatimento de gastos com educação no imposto de renda?

"O governo do Estado de São Paulo chamará no próximo mês mais 30 mil professores aprovados em concurso público no ano passado para atuar em sua rede de ensino. Os novos docentes deverão lecionar no ensino médio e nos últimos anos do ensino fundamental"
Questão: Quantos professores serão convocados?

"Dois em cada três usuários brasileiros do LinkedIn têm diploma universitário. Os pós-graduados somam 25% dos brasileiros na plataforma. Os dados foram divulgados pelo LinkedIn nesta sexta-feira (18) junto ao anúncio de que o serviço atingiu 300 milhões de usuários no mundo todo."
Questão: Quantos usuários tem a rede Linkedln?

Questões usadas para Avaliar o Sintetizador B (B1M)

"O nível do sistema Cantareira registrou um novo recorde negativo nesta segunda-feira (28) operando com 11% de sua capacidade total. De acordo com a Sabesp, esta é a primeira vez na história que o manancial atinge esta marca"
Questão: Qual foi o nível do sistema cantareira no dia 28?

"A cidade de São Paulo registrou a madrugada mais fria do ano nesta segunda-feira (28). De acordo com a medição automática do Inmet (Instituto Nacional de Meteorologia), os termômetros registraram 13,6º C, no Mirante de Santana, na zona norte. A temperatura foi verificada por volta das 5h"
Questão: Quantos graus foi registrado na madrugada de segunda feira (28)?

"Um pouco mais de lenha acaba de ser colocada na fogueira do parque Augusta. Diferente dos dois grupos que têm protestado até agora ali, um novo movimento de moradores começa a se organizar para pedir que o local tenha sim prédios em parte da área. Essa também é a intenção das construtoras Cyrela e Setin, donas do terreno desde o final do ano passado"
Questão: O que o novo movimento de moradores pedem?

"Um grupo de 18 universidades públicas e privadas brasileiras decidiu colocar parte de seus professores em um curso para que eles adotem novos modelos de aulas. A ideia é que os docentes que fizerem o curso deixem de aplicar as aulas tradicionais, expositivas, e passem a adotar metodologias como o aprendizado baseado em resolução de problemas e a instrução por pares"
Questão: Quantas universidades públicas e privadas vão participar do curso?

Questões usadas para Avaliar o Sintetizador A (B1F)

"A Apple está preparando o lançamento de uma nova versão do MacBook Air para a próxima terça-feira. A informação publicada no site aponta o recebimento de grandes quantidades do produto nas lojas da empresa nos Estados Unidos"
Questão: O que a Apple está preparando?

"A empresa israelense desenvolveu uma impressora de bolso que imprime documentos em até 45 segundos. A Zuta Labs criou a Pocket Printer, uma impressora do tamanho de uma xícara de café que é compatível com qualquer tipo de aparelho eletrônico e pode ser usada em qualquer lugar a qualquer hora"
Questão: Qual o tamanho da impressora?

"O ex-presidente Luiz Inácio Lula da Silva disse a amigos próximos nesse fim de semana que será o candidato do PT à Presidência da República, conforme a jornalista Joyce Pascowitch. A colunista afirmou no site Glamurama que o presidente expressou sua intenção de voltar ao posto, apesar de Lula der dito que vai "ser cabo eleitoral da Dilma""
Questão: O que Lula disse aos amigos de acordo com a jornalista?

"Apesar da falta de consenso no documento final da NETmundial, o evento teve momentos marcantes. Na cerimônia de abertura, antes de iniciar seu discurso, a presidente Dilma Rousseff sancionou o Marco Civil da Internet, aprovado na terça-feira, 22, no Senado Federal"
Questão: O que a presidenta Dilma sancionou?

Sentenças Usadas para Avaliar a Pronúncia, Velocidade e Naturalidade

- Sentença usada para avaliar a voz humana: Oi, meu nome é Maria. Você me escutou falando diversas frases enquanto respondia ao questionário de avaliação. Foi um prazer falar com você! Muito obrigado.
- Sentença usada para avaliar o sintetizador A (A1M): Oi, meu nome é João. Você me escutou falando diversas frases enquanto respondia ao questionário de avaliação. Foi um prazer falar com você! Muito obrigado.
- Sentença usada para avaliar o sintetizador A (A1F): Oi, meu nome é Márcia. Você me escutou falando diversas frases enquanto respondia ao questionário de avaliação. Foi um prazer falar com você! Muito obrigado.
- Sentença usada para avaliar o sintetizador B (B1M): Oi, meu nome é Pedro. Você me

escutou falando diversas frases enquanto respondia ao questionário de avaliação. Foi um prazer falar com você! Muito obrigado.

- Sentença usada para avaliar o sintetizador A (B1F): Oi, meu nome é Bruna. Você me escutou falando diversas frases enquanto respondia ao questionário de avaliação. Foi um prazer falar com você! Muito obrigado.

8.5 Processo de Geração de Fala

Devido a deficiências do módulo de *Front-End*, muitas vezes a transcrição da linguagem escrita para a linguagem falada necessita de auxílio humano. Na geração das falas para validar a metodologia proposta, nenhum auxílio humano foi empregado, pois a finalidade era justamente conhecer as deficiências dos sistemas TTS testados.

No entanto, para uma aplicação prática, é conveniente adaptar o texto a fim de conseguir uma melhor locução. Devido ao fato de cada sistema TTS possuir o seu módulo *Front-End*, não existe um conjunto fechado de regras de como escrever um texto passível de uma ótima locução, mas sim um conjunto recomendações que, de maneira geral, facilita a transcrição. Abaixo listamos algumas delas:

- Evitar entrar com um texto longo, visto que é muito mais difícil para o sintetizador encontrar o contexto de cada palavra em um texto longo comparado com um texto pequeno.
- Sempre que possível dividir um paragrafo grande em vários pequenos.
- Quando precisar de uma pausa no meio da locução, forçar com o uso de vírgulas ou até mesmo pontos.
- Nunca esquecer sinais de interrogação e exclamação.
- Quase sempre é preciso transcrever siglas manualmente.
- A locução de nomes quase sempre é problemática, é recomendado que se teste várias formas de escrita até conseguir a pronúncia correta.

A descoberta das deficiências de cada sintetizador é feita por meio do uso. No entanto, com base na experiência adquirida neste trabalho, percebe-se que com o uso das recomendações acima, é possível melhorar consideravelmente a qualidade da locução.

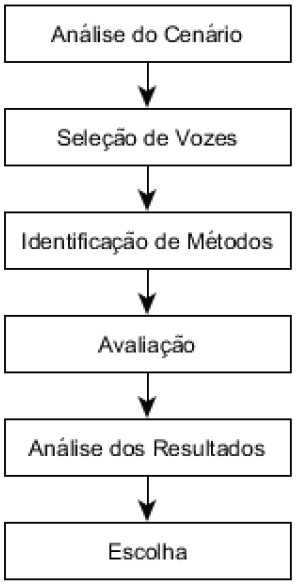
8.6 *Template* da Metodologia Proposta

Nas páginas seguintes o *template* da metodologia proposta é apresentado. Os dados nele contidos são referentes a avaliação que buscou encontrar a voz sintetizada ideal para a plataforma de ensino usando avatares.

TEMPLATE PARA AVALIAÇÃO DE VOZ SINTETIZADA

Edição 4 – Maio de 2014

Harlei Miguel de Arruda Leite
Prof. Dr. Dalton Soares Arantes



Trabalho financiado pela
FAPESP e Padtec S.A. (Proc. 2007/56018-4), CAPES e CNPq.

Campinas
Maio de 2014

TEMPLATE PARA AVALIAÇÃO DE VOZ SINTETIZADA

Etapa 1 de 6
Definir o escopo do projeto e as principais características que o sintetizador de voz deve ter, assim como questões da plataforma a ser utilizada

1.1	Nome do Projeto
Projeto Ambientes de Ensino a Distância Usando Avatares	

1.2	Motivação para o uso de Sintetizador de Voz
Facilidade no desenvolvimento de aulas; Facilidade de alterar o roteiro de aula	

1.3	Idioma da Voz
Português (Brasil)	

1.4	Sexo da Voz
(X) Masculino (X) Feminino ( ) Indiferente	

1.5	Características Requeridas
( ) Inteligibilidade e Compreensibilidade (X) Inteligibilidade, Compreensibilidade e Características Subjetivas	

1.6	Plataforma do Projeto
( ) Servidor (Para aplicações cliente servidor) (X) Desktop ( ) Aplicação Web ( ) Dispositivos móveis ( ) Sistema embarcado ( ) Outros _____	

1.7	Informações Complementares
O sistema TTS deve dar suporte para o sistema operacional Windows	

TEMPLATE PARA AVALIAÇÃO DE VOZ SINTETIZADA

Etapa 2 de 6
Definir o conjunto de vozes a serem avaliadas. As vozes devem ser selecionadas de forma a satisfazerem os requisitos definidos na Etapa 1

Voz Natural para Controle	
Nome da voz	Cláudia
Idioma	Português (Brasil)
Sexo	( ) Masculino (X) Feminino

Nome do sintetizador	Sintetizador A
Nome da voz	Voz A1F
Idioma	Português (Brasil)
Sexo	( ) Masculino (X) Feminino
Plataformas suportadas	Windows, Linux e Mac
Informações técnicas	Síntese concatenativa
Tipo de licença	Comercial

Nome do sintetizador	Sintetizador A
Nome da voz	Voz A1M
Idioma	Português (Brasil)
Sexo	(X) Masculino ( ) Feminino
Plataformas suportadas	Windows, Linux e Mac
Informações técnicas	Síntese concatenativa
Tipo de licença	Comercial

Nome do sintetizador	Sintetizador B
Nome da voz	Voz B1F
Idioma	Português (Brasil)
Sexo	( ) Masculino (X) Feminino
Plataformas suportadas	Windows e Mac
Informações técnicas	Síntese concatenativa
Tipo de licença	Comercial

Nome do sintetizador	Sintetizador B
Nome da voz	Voz B1M
Idioma	Português (Brasil)
Sexo	(X) Masculino ( ) Feminino
Plataformas suportadas	Windows e Mac
Informações técnicas	Síntese concatenativa
Tipo de licença	Comercial

TEMPLATE PARA AVALIAÇÃO DE VOZ SINTETIZADA

Etapa 3 de 6
Selecionar um grupo de critérios de avaliação de voz de acordo com os requisitos definidos na Etapa 1

( )	**Critérios para Avaliar Inteligibilidade, Compreensibilidade e Teste em Campo**
• WER (*Word Error Rates*) • Questões sobre notícias sintetizadas • Teste em campo	

(x)	**Critérios para Avaliar Inteligibilidade, Compreensibilidade e Características Subjetivas**
• WER (*Word Error Rates*) • Questões sobre notícias sintetizadas • Escala MOS (*Mean Opinion Score*) para avaliações subjetivas • Teste de campo	

TEMPLATE PARA AVALIAÇÃO DE VOZ SINTETIZADA

Etapa 4 de 6
Aplicar os critérios de avaliação de acordo com as diretrizes do Template. Cada voz deve ser avaliada individualmente por todos os critérios selecionados na Etapa 3

Critério de Avaliação utilizando WER (Word Error Rates)
O cálculo WER permite calcular a diferença entre a frase sintetizada e a frase que o ouvinte julga ter ouvido. O cálculo é dado por $WER = \frac{S+D+I}{N}$, onde S é o número de substituições de palavras; D é o número de exclusões de palavras; I é o número de inserções de palavras e N é o número total de palavras. É recomendado que o cálculo seja feito de maneira automatizado. As frases usadas devem seguir o padrão SUS (*Semantically Unpredictable Sentences*).

Questões sobre Notícias Sintetizadas
O critério para avaliar a compreensibilidade consiste em sintetizar uma notícia de 150 palavras e pedir ao ouvinte para responder questões a respeito das notícias. As questões podem ser dissertativas ou objetivas. Para as questões dissertativas, as notas devem ser dadas como segue: Nota 2 se a resposta estiver plenamente correta, nota 1 se estiver parcialmente correta e nota 0 se estiver errada. Para questões objetivas, somente o certo e errado é considerado.

Escala MOS (*Mean Opinion Score*) para Avaliar Características Subjetivas
Para avaliar a pronúncia, velocidade e naturalidade, deve-se sintetizar uma frase curta e pedir para o ouvinte atribuir uma nota para as características de 1 a 5 (escala MOS), sendo: Pronúncia: 1 – Péssima pronúncia e 5 – Ótima pronúncia. Velocidade: 1 – Lento demais e 5 – Rápido demais. Naturalidade: 1 – Pouco natural e 5 – Muito natural.

Teste em Campo
Aplicar um questionário para avaliar a voz sintetizada assim como a sua integração com o produto e o ambiente. As perguntas devem cobrir tanto o aspecto da voz como o produto como um todo. O teste pode ser aplicado em um grupo simultaneamente.

TEMPLATE PARA AVALIAÇÃO DE VOZ SINTETIZADA

Etapa 5 de 6
Analisar os dados brutos obtidos pelos critérios de avaliação. É recomendado que toda a análise seja feita por um software estatístico ou de planilha eletrônica

Critério de Avaliação utilizando WER (Word Error Rates)

Preencher a tabela abaixo para cada voz

Voz Natural para Controle	
Nome da voz natural	Cláudia
Taxa de acerto	1,14

Nome do sintetizador	Sintetizador A
Nome da voz	Voz A1F
Taxa de acerto	6,35

Nome do sintetizador	Sintetizador A
Nome da voz	Voz A1M
Taxa de acerto	1,50

Nome do sintetizador	Sintetizador B
Nome da voz	Voz B1F
Taxa de acerto	2,57

Nome do sintetizador	Sintetizador B
Nome da voz	Voz B1M
Taxa de acerto	1,60

TEMPLATE PARA AVALIAÇÃO DE VOZ SINTETIZADA

Questões sobre Notícias Sintetizadas

Preencher a tabela abaixo para cada sintetizador

Voz Natural para Controle	
Nome da voz natural	Cláudia
Taxa de acerto	95,31%
Taxa de erro	4,68%

Nome do sintetizador	Sintetizador A
Nome da voz	Voz A1F
Taxa de acerto	77,34%
Taxa de erro	22,65%

Nome do sintetizador	Sintetizador A
Nome da voz	Voz A1M
Taxa de acerto	85,15%
Taxa de erro	14,84%

Nome do sintetizador	Sintetizador B
Nome da voz	Voz B1F
Taxa de acerto	75,00%
Taxa de erro	25,00%

Nome do sintetizador	Sintetizador B
Nome da voz	Voz B1M
Taxa de acerto	61,32%
Taxa de erro	38,67%

TEMPLATE PARA AVALIAÇÃO DE VOZ SINTETIZADA

Escala MOS (*Mean Opinion Score*) para Avaliar Características Subjetivas

Preencher a tabela abaixo para cada sintetizador

Voz Natural para Controle	
Nome da voz natural	Cláudia
Média da pronúncia (MOS)	4,15
Média da velocidade (MOS)	3,18
Média da naturalidade (MOS)	3,5

Nome do sintetizador	Sintetizador A
Nome da voz	Voz A1F
Média da pronúncia (MOS)	2,62
Média da velocidade (MOS)	3,65
Média da naturalidade (MOS)	1,87

Nome do sintetizador	Sintetizador A
Nome da voz	Voz A1M
Média da pronúncia (MOS)	3,78
Média da velocidade (MOS)	3,34
Média da naturalidade (MOS)	2,87

Nome do sintetizador	Sintetizador B
Nome da voz	Voz B1F
Média da pronúncia (MOS)	3,59
Média da velocidade (MOS)	3,25
Média da naturalidade (MOS)	3,00

Nome do sintetizador	Sintetizador B
Nome da voz	Voz B1M
Média da pronúncia (MOS)	3,65
Média da velocidade (MOS)	3,25
Média da naturalidade (MOS)	3,12

TEMPLATE PARA AVALIAÇÃO DE VOZ SINTETIZADA

Avaliação em Campo

Preencher a tabela abaixo para cada sintetizador

Voz Natural para Controle	
Nome da voz natural	Cláudia
Média da agradabilidade (MOS)	4,12
Média da compreensão (MOS)	4,37
Taxa de acerto nas questões	82%

Nome do sintetizador	Sintetizador A
Nome da voz	Voz A1F
Média da agradabilidade (MOS)	2,65
Média da compreensão (MOS)	3,43
Taxa de acerto nas questões	91%

Nome do sintetizador	Sintetizador A
Nome da voz	Voz A1M
Média da agradabilidade (MOS)	2,68
Média da compreensão (MOS)	3,62
Taxa de acerto nas questões	91%

Nome do sintetizador	Sintetizador B
Nome da voz	Voz B1F
Média da agradabilidade (MOS)	3,53
Média da compreensão (MOS)	4,34
Taxa de acerto nas questões	79%

Nome do sintetizador	Sintetizador B
Nome da voz	Voz B1M
Média da agradabilidade (MOS)	3,34
Média da compreensão (MOS)	4,06
Taxa de acerto nas questões	94%

TEMPLATE PARA AVALIAÇÃO DE VOZ SINTETIZADA

Resumo dos Dados

Preencher a tabela abaixo para cada sintetizador

Voz Natural para Controle	
Nome da voz	Cláudia
Taxa de acerto WER	1,14
Taxa de compreensão	95,31%
Média da pronúncia (MOS)	4,15
Média da velocidade (MOS)	3,18
Média da naturalidade (MOS)	3,5

Nome do sintetizador	Sintetizador A	Nome da voz	Voz A1F
Taxa de acerto WER			6,35
Taxa de compreensão			77,34%
Média da pronúncia (MOS)			2,62
Média da velocidade (MOS)			3,65
Média da naturalidade (MOS)			1,87

Nome do sintetizador	Sintetizador A	Nome da voz	Voz A1M
Taxa de acerto WER			1,50
Taxa de compreensão			85,15%
Média da pronúncia (MOS)			3,78
Média da velocidade (MOS)			3,34
Média da naturalidade (MOS)			2,87

Nome do sintetizador	Sintetizador B	Nome da voz	Voz B1F
Taxa de acerto WER			2,57
Taxa de compreensão			75,00%
Média da pronúncia (MOS)			3,59
Média da velocidade (MOS)			3,25
Média da naturalidade (MOS)			3,00

Nome do sintetizador	Sintetizador B	Nome da voz	Voz B1M
Taxa de acerto WER			1,60
Taxa de compreensão			61,32%
Média da pronúncia (MOS)			3,65
Média da velocidade (MOS)			3,25
Média da naturalidade (MOS)			3,12

TEMPLATE PARA AVALIAÇÃO DE VOZ SINTETIZADA

AVALIAÇÃO EM ESTÚDIO

Sintetizador Voz	WER	Compreensão	Pronúncia (MOS)	Velocidade (MOS)	Naturalidade (MOS)
Voz Natural	1,14	95,31	4,15	3,18	3,5
Sint. A / A1F	6,35	77,34	2,62	3,65	1,87
Sint. A / A1M	1,50	85,15	3,78	3,34	2,87
Sint. B / B1F	2,57	75,00	3,59	3,25	3,00
Sint. B / B1M	1,60	61,32	3,65	3,25	3,12
Melhor Resultado	A1M	A1M	A1M	B1F / B1M	B1M

CLASSIFICAÇÃO GERAL		
1	Sintetizador A / Voz A1M	Melhor voz Masculina
2	Sintetizador B / Voz B1M	
3	Sintetizador B / Voz B1F	Melhor voz Feminina
4	Sintetizador A / Voz A1F	

TEMPLATE PARA AVALIAÇÃO DE VOZ SINTETIZADA

AVALIAÇÃO EM CAMPO

Sintetizador Voz	Agradabilidade da voz (MOS)	Compreensão da voz (MOS)	Taxa de acerto sobre conteúdo da aula
Voz Natural	4,12	4,37	81,82%
Sint. A / A1F	2,65	3,43	90,91%
Sint. A / A1M	2,68	3,62	90,91%
Sint. B / B1F	3,53	4,34	78,79%
Sint. B / B1M	3,34	4,06	93,94%
Melhor Resultado	Sint. B / B1F	Sint. B / B1F	Sint. B / B1M

CLASSIFICAÇÃO GERAL		
1	Sintetizador B / B1F	Melhor voz Feminina
2	Sintetizador B / B1M	Melhor voz Masculina
3	Sintetizador A / A1M	
4	Sintetizador A / A1F	

TEMPLATE PARA AVALIAÇÃO DE VOZ SINTETIZADA

Etapa 6 de 6
Com base nos dados da Etapa 5, selecionar a voz que melhor cumpre os requisitos do projeto. A escolha deve levar em conta todos os requisitos da Etapa 1

Voz Selecionada

Voz Masculina

Sintetizador B / B1M

Resumo da avaliação da voz

Taxa de acerto WER	1,60
Taxa de compreensão	61,32%
Média da pronúncia (MOS)	3,65
Média da velocidade (MOS)	3,25
Média da naturalidade (MOS)	3,12
Média da agradabilidade da voz (Teste em Campo - MOS)	3,34
Média da compreensão da voz (Teste em Campo - MOS)	4,06
Taxa de acerto sobre o conteúdo da aula (Teste em Campo)	93,94%

Informações Complementares Sobre a Escolha

A voz B1M (Masculina) do sintetizador A foi a escolhida por ter apresentado o melhor desempenho nos testes de velocidade da fala, naturalidade, agradabilidade (MOS – Teste em Campo), Compreensão (MOS – Teste em Campo), o que levou a uma taxa de 93,94% de acertos nas questões objetivas na avaliação em campo.

TEMPLATE PARA AVALIAÇÃO DE VOZ SINTETIZADA

Voz Selecionada

Voz Feminina

Sintetizador B / B1F

Resumo da avaliação da voz

Taxa de acerto WER	2,57
Taxa de compreensão	75,00%
Média da pronúncia	3,59
Média da velocidade	3,25
Média da naturalidade	3,00
Média da agradabilidade da voz (Teste em Campo - MOS)	3,53
Média da compreensão da voz (Teste em Campo - MOS)	4,34
Taxa de acerto sobre o conteúdo da aula (Teste em Campo)	78,79%

Informações Complementares Sobre a Escolha

A voz B1F (Feminina) do sintetizador B foi a escolhida por ter apresentado o melhor desempenho nos testes WER, Pronúncia, Velocidade, Naturalidade, Agradabilidade (MOS – Teste em Campo), Compreensibilidade (MOS – Teste em Campo), o que levou a uma taxa de 79% de acertos nas questões objetivas na avaliação em campo.