Avaliação de Vozes Artificiais: Inteligibilidade, Compreensibilidade e Naturalidade

Harlei M. A. Leite [1], Sarah N. Carvalho [2,3], Tiago Cinto[1] e Dalton S. Arantes[1]

[1] Departamento de Comunicações
[2]Departamento de Engenharia de Computação e Automação Industrial
Faculdade de Engenharia Elétrica e de Computação
Universidade Estadual de Campinas (UNICAMP)
Campinas, SP, Brasil

[3]Departamento de Engenharia Elétrica
Instituto de Ciências Exatas e Aplicadas
Universidade Federal de Ouro Preto (UFOP)
João Monlevade, MG, Brasil

{hmleite,tcinto,dalton}@decom.fee.unicamp.br

{sarah}@decea.ufop.br

Abstract. *The evolution of the electronic media has led to new and challenging technology needs. Today there is a growing demand for high quality speech synthesizers, targeting applications in the most diverse fields of use, such as educational systems, telephony, accessibility and human-machine interaction. In this context, the present study evaluates the relationship between intelligibility, comprehensibility and naturalness of synthesized voices by concatenative technique and highlights the relevant features to be considered for choice of a synthetic voice suitable for an application and to improve speech synthesizers.*

Resumo. *A evolução dos meios de comunicação tem levado a novas exigências tecnológicas. Hoje em dia, há uma crescente demanda por sistemas sintetizadores de fala de alta qualidade para serem integrados nos mais diversos campos de uso, desde sistemas educativos à telefonia e acessibilidade. Neste contexto de aprimoramento das interações homem-máquina, o presente trabalho avalia a relação entre inteligibilidade, compreensibilidade e naturalidade de vozes sintetizadas pela técnica concatenativa e destaca as características relevantes a serem consideradas tanto para nortear a escolha da voz sintética adequada a uma aplicação, como para auxiliar no aprimoramento dos sintetizadores de fala.*

1. Introdução

Nos dias de hoje, as máquinas fazem parte da vida de grande parte da população mundial. Percebe-se que o ser humano está cada vez mais dependente de equipamentos como celulares, tablets, computadores, etc., pois estes dinamizam o mundo, facilitando a realização

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

de tarefas, encurtando as distâncias na comunicação entre pessoas e permitindo a transferência e aquisição de informação de forma ágil e eficiente. As máquinas estão inseridas no contexto de atividades de trabalho, cultura e lazer. Tudo isto só é possível graças ao desenvolvimento tecnológico e a facilidade crescente que a "pessoa comum" tem em utilizar estes aparelhos. O fato da interface eletrônica ser intuitiva e simples de ser utilizada é essencial para a popularidade destes equipamentos. O aprimoramento das interações homem-máquina busca cada vez mais permitir que as máquinas possuam capacidades humanas, de forma que elas possam falar, agir e pensar como as pessoas. Neste âmbito tecnológico, existem várias pesquisas e investimentos que visam aperfeiçoar o processamento de sinais de fala, a programação não linear integrada à inteligência artificial, o aprendizado de máquina, o desenvolvimento de sensores a serem integrados nos equipamentos, entre outras.

O contínuo progresso na área de processamento de sinais de fala tem permitido o aprimoramento dos sistemas TTS (*Text-to-Speech*), possibilitando a aplicação de voz artificial em diversos cenários, como no educacional, nos programas de acessibilidade para deficientes visuais, aprendizado de idioma, secretária eletrônica, entretenimento e até mesmo para fins militares [Lemmetty 1999] [Georgila et al. 2012] [Martins and Brasiliano 2012].

Os principais aspectos que influenciam na qualidade da voz sintética são inteligibilidade, naturalidade e compreensibilidade [Klatt 1987] [Mariniak 1993]. Eles são importantes não só por garantirem que a mensagem textual seja adequadamente transformada em mensagem oral mas, também, porque impactam na preferência e aceitabilidade do ouvinte da voz artificial. [Yu-Yun 2011].

A tarefa de avaliar um sintetizador de voz envolve a aplicação de diversos métodos de avaliação [Pisoni et al. 1985] [Sydeserff et al. 1992] [Stevens et al. 2005]. Atualmente, esses métodos estão divididos em duas classes: *Black Box* e *Glass Box*. Os métodos de avaliação da classe *Black Box* consideram somente a entrada (texto) e a saída (fala), sendo que toda a etapa de processamento é desconsiderada e a avaliação se dá pela percepção subjetiva da voz. Por outro lado, os métodos da classe *Glass Box* analisam não somente a entrada e a saída do sistema, mas todo o processamento de um sistema TTS. Normalmente, métodos *Glass Box* são utilizados em situações em que os métodos *Black Box* não conseguem identificar fragilidades de um sintetizador de fala [Heuven and Bezooijen 1995].

Por meio da aplicação de testes de avaliação, pesquisadores encontraram uma forte relação entre inteligibilidade e compreensibilidade [Pisoni et al. 1985]. Embora existam estudos focados nessa relação, ainda não foi desenvolvido um método suficientemente confiável para avaliar a compreensibilidade, por envolver fatores cognitivos de difícil quantificação [Yu-Yun 2011].

Este artigo busca analisar as relações entre inteligibilidade, compreensibilidade e naturalidade, por meio da aplicação de métodos da classe (*Black Box*) em duas vozes comerciais que utilizam a técnica de síntese concatenativa e em uma voz humana, usada como referência. Os testes foram estruturados visando encontrar as fragilidades das vozes e evitar o fator de aprendizado do ouvinte e a sua familiarização com a voz durante o teste.

1.1. Síntese de Fala

Um sistema TTS é um sistema que recebe como entrada um texto escrito em linguagem natural e o sintetiza, gerando uma saída em áudio correspondente ao texto de entrada. Um sistema TTS padrão é dividido em dois estágios: *Front-End* e *Back-End*, conforme mostra a Figura 1.

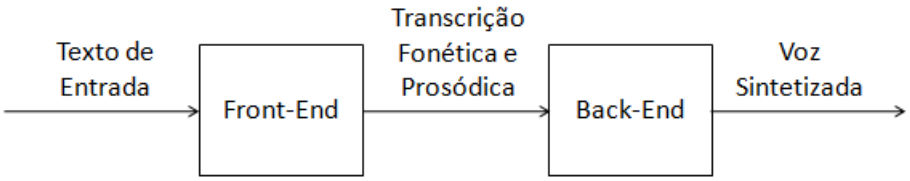

Figura 1. Arquitetura de um Sistema TTS.

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, sendo que os mais utilizados atualmente são (i) síntese concatenativa e (ii) síntese baseada em modelos ocultos de Markov (HMM).

Cada tipologia tem suas vantagens e desvantagens intrínsecas. Normalmente, é a aplicação e o escopo de utilização do sistema de síntese que determina qual delas deve ser adotada. Na Seção 1.2 a síntese concatenativa é explorada com maiores detalhes por ser o método adotado pelos sintetizadores avaliados.

1.2. Síntese Concatenativa

O processo de síntese concatenativa está baseado na concatenação de segmentos de fala natural pré-gravada. Esta técnica exige um banco de dados de áudio e texto de alta qualidade, uma vez que a qualidade da voz artificial gerada está diretamente relacionada com a extensão e qualidade das gravações armazenadas no banco de dados.

A principal vantagem da síntese concatenativa é a qualidade da fala gerada em termos de naturalidade do som. Entretanto, diferenças entre variações naturais da fala e a natureza das técnicas automáticas para segmentação das formas de ondas podem acarretar em falhas que comprometem a inteligibilidade do discurso.

Um dos aspectos mais importantes da síntese concatenativa é a segmentação da fala com o comprimento ideal. Durante a concatenação dos segmentos de fala deve-se evitar a ocorrência de transições bruscas, uma vez que os sons a serem concatenados não são naturalmente sequenciais. A maior parte dos sistemas comerciais realizam a concatenação de polifones que compreendem grupos de difones e trifones. Empregando o uso de polifones é possível realizar de maneira relativamente suave as transições entre os fones, sendo necessário trabalhar a junção de regiões do sinal com conteúdo espectral semelhante. Com esta técnica pode-se obter um sinal de fala mais natural e agradável.

Todavia, esta abordagem exige o uso de uma base de fala extensa que permita extrair os polifones nos diversos contextos.

2. Métodos de Avaliação da Voz Artificial

Sintetizadores de fala podem ser avaliados em termos de inteligibilidade, compreensibilidade e naturalidade [Klatt 1987] [Mariniak 1993]. Os métodos de avaliação devem considerar esses termos de acordo com a aplicação na qual o sintetizador vai ser utilizado. Em algumas aplicações, como leitores para cegos, a inteligibilidade é um fator de extrema importância. Por outro lado, em aplicações multimídia e de entretenimento, a naturalidade e a inteligibilidade são igualmente importantes. A tarefa de avaliação da voz pode ser feita a nível de fonema, palavra ou sentença, dependendo do método de avaliação utilizado [Lemmetty 1999].

A tarefa de avaliar uma voz contém inúmeros desafios [Jekosch 1993]. Em um sistema TTS não somente as caracteristícias acústicas são importantes, mas sim toda a etapa de pré-processamento e linguística (*Front-End*), de forma que o sistema como um todo influencia na qualidade final da voz sintetizada. Sendo assim, diferentes métodos de avaliação devem ser usados para se obter bons resultados [Lemmetty 1999].

O procedimento de avaliação da qualidade da voz artificial é usualmente feito por meio de testes subjetivos, em que um ouvinte escuta um conjunto de sílabas, palavras ou sentenças. O conjunto de dados para a avaliação normalmente é focado em consoantes, por serem elas as mais problemáticas no processo de síntese em relação as vogais [Carlson et al. 1990].

É importante que os testes sejam realizados uma única vez com cada indivíduo, a fim de evitar que ocorra a familiarização com a voz artificial, o que causaria um "efeito de aprendizagem", implicando significativamente na melhoria dos resultados. Por outro lado, problemas de concentração podem degradar os resultados significativamente. Por conta disso, a escolha entre ouvintes amadores e profissionais é importante, principalmente quando o sintetizador for utilizado em aplicações críticas.

Além da experiência dos ouvintes, é necessário que a voz sintética seja da língua materna dos ouvintes, pois isso impacta diretamente na compreensibilidade e na inteligibilidade.

Para o presente trabalho, os métodos empregados foram o MRT (*Modified Rhyme Test*), o WER (*Word Error Rates*) aplicado em sentenças SUS (*Semantically Unpredictable Sentence*), o Método de Avaliação de Compreensibilidade e o Método de Classificação de Naturalidade e Inteligibilidade por meio da escala MOS (*Mean Opinion Score*). Os testes foram dimensionados de forma a avaliar os três fatores que impactam diretamente na qualidade da fala sintética: inteligibilidade, compreensibilidade e naturalidade. Nas subseções seguintes, é apresentada uma breve explicação sobre a estruturação de cada método.

2.1. MRT (*Modified Rhyme Test*)

O método MRT tem por objetivo avaliar a inteligibilidade das consoantes iniciais e das consoantes finais de uma palavra [Logan et al. 1989] [Goldstein 1995]. O método consiste em conjuntos de palavras de duas sílabas, em que o ouvinte escuta a palavra e marca

a opção que ele escutou em um questionário de múltipla escolha de 6 items. Um conjunto de palavras deve testar somente a consoante inicial ou somente a consoante final. A Tabela 1 ilustra exemplos de palavras para o teste MRT. A primeira metade (conjuntos de 1 a 4) testa a consoante inicial e a segunda metade (conjuntos de 5 a 8) testa a consoante final. Somente uma palavra de cada conjunto deve ser sintetizada e ouvida.

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

Tabela 1. Conjunto de dados para o teste MRT.

Como resultado, o método permite visualizar a taxa de erro das consoantes iniciais, finais e a média geral. No entanto, o método não consegue avaliar a prosódia, pois se concentra em palavras isoladas fora de qualquer contexto.

2.2. WER (*Word Error Rates*)

O cálculo WER é derivado da *Levenshtein Distance* [Levenshtein 1966] e permite avaliar a inteligibilidade de um sintetizador de voz. Nesse teste, pede-se aos ouvintes que escutem as frases e as escrevam, para que posteriormente se compare a frase original e a frase redigida. O WER mede o grau de diferença entre elas, de acordo com a equação

$$WER = \frac{S + D + I}{N} \tag{1}$$

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 e N é o número total de palavras.

Para aplicar o WER é importante que as sentenças sintetizadas sejam gramaticalmente corretas, porém semanticamente confusas, de forma a evitar deduções que possam incrementar a taxa de acerto. Tais sentenças são chamadas de SUS (*Semantically Unpredictable Sentences*) [Benoit et al. 1996]. Para vários idiomas existem frases SUS tabeladas para aplicar nos testes, entretanto para a língua portuguesa tais sentenças não foram encontradas. A fim de contornar essa dificuldade, foram geradas sentenças usando os critérios do SUS para permitir a realização do teste e o cálculo do WER.

Exemplo da aplicação do WER em uma sentença semanticamente confusa:

Sentença esperada: A camiseta ganhou um fone de gramado.

Sentença ouvida: A camiseta ganhou um fone ** cinemado.

No exemplo acima ocorre 1 substituição (gramado para cinemado), nenhuma inserção e uma exclusão (faltou o pronome “de”). Calculando por meio da Equação (1), obtém-se $WER = 0.28$.

2.3. Método de Avaliação de Compreensibilidade

Diferentemente da inteligibilidade, não se pode avaliar o nível de compreensibilidade que um sintetizador de voz proporciona solicitando aos ouvintes que simplesmente escrevam as frases ou palavras ouvidas. No teste de compreensibilidade se deseja avaliar se a voz sintetizada foi capaz de transmitir a mensagem de forma coerente e clara e 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, diversas notícias populares de aproximadamente 180 caracteres foram sintetizadas. Para cada notícia, uma pergunta dissertativa de resposta direta foi elaborada sobre o assunto. Nesse teste, os ouvintes escutam a notícia uma única vez e respondem as perguntas logo em seguida.

O processo de avaliação do método consiste em dar notas para as respostas com o seguinte critério: Nota 2 se a resposta estiver plenamente correta, nota 1 se estiver parcialmente correta e nota 0 se estiver errada. A taxa de compreensibilidade está diretamente correlacionada com o número de acertos, sendo assim a taxa de compreensibilidade é a média das notas dadas [Yu-Yun 2011].

2.4. Avaliação Subjetiva

A avaliação da naturalidade do sintetizador de voz foi feita por meio de perguntas subjetivas diretas utilizando a escala MOS. Nesse teste, o ouvinte escutou uma frase sintetizada de 11 segundos e baseando-se nessa frase e nas frases anteriormente ouvidas atribuiu-se uma nota para a naturalidade da voz na escala de 1 a 5, sendo: 1- Mau; 2- Pobre; 3- Razoável; 4- Bom e 5- Excelente.

O nível de naturalidade foi calculado como sendo o valor médio das notas atribuídas no teste. De modo equivalente, foi questionado aos ouvintes que dessem uma nota para a inteligibilidade da voz. Baseando-se na médias das notas dadas, se encontrou o nível de inteligibilidade subjetivo para cada voz.

A escala MOS pode ser utilizada para avaliar qualquer característica de um sintetizador. Por outro lado, ela pode oferecer dados menos precisos comparado com os testes MRT e WER para avaliar a inteligibilidade. No entanto, para avaliar a naturalidade e agradabilidade, ainda não foi encontrado um método para concorrer com a escala MOS por ser uma questão subjetiva.

3. Metodologia

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, com plena capacidade de realização do teste responderam a um questionário *online*. 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 1 e voz do sintetizador 2.

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.

4. Resultados

A Figura 2 resume os resultados dos experimentos realizados para as duas vozes sintéticas comerciais e para a voz humana (controle).

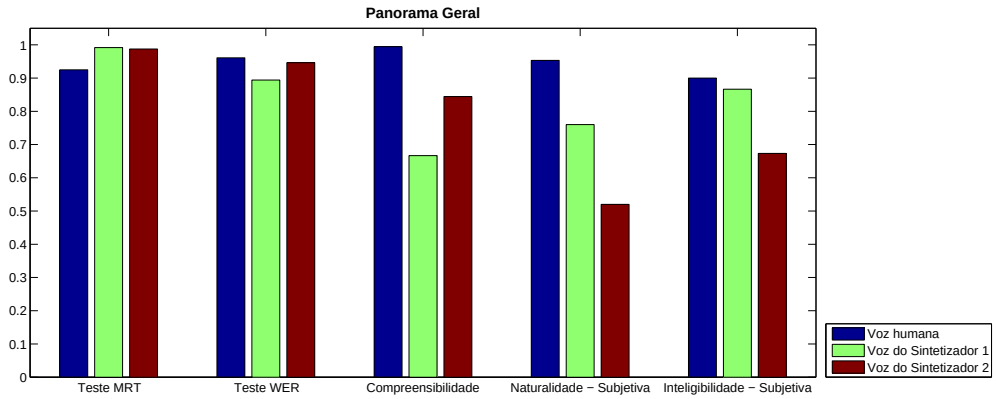

Figura 2. Panorama complexivo dos resultados.

Observa-se que a voz humana obteve o melhor desempenho em todos os testes, exceto para o teste MRT. Nesse, a taxa de acertos foi ligeiramente inferior, provavelmente porque o locutor não deu a enfase necessária ao pronunciar cada sílaba, levando, deste modo, os ouvintes a não entenderem corretamente as palavras dissílabas pronunciadas isoladamente.

Ambas as vozes artificiais obtiveram taxas de acerto nos testes de inteligibilidade (MRT e acerto de palavras) em torno de 90%, indicando a boa qualidade da síntese dessas vozes comerciais. Na avaliação subjetiva da inteligibilidade, a voz do sintetizador 1 recebeu uma nota também em torno de 90%, enquanto que a voz do sintetizador 2 obteve uma nota inferior a 70%, bem abaixo do valor obtido no teste objetivo. A Figura 3 permite visualizar que existe uma relação inversa entre a avaliação subjetiva da inteligibilidade e a taxa de acertos nos testes.

A provável causa deste resultado parece ser indicada pela análise das notas subjetivas para o quesito Naturalidade da Voz. Percebe-se que para esta característica a voz do sintetizador 1 recebeu nota de 87% e a voz artificial do sintetizador 2 recebeu nota de

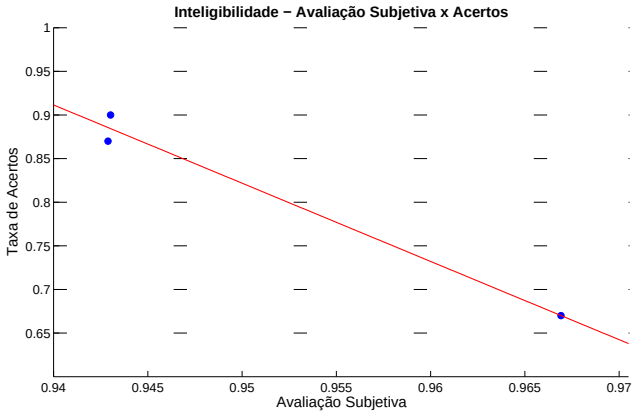

Figura 3. Inteligibilidade - Avaliação Subjetiva x Objetiva.

52%. 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 ao sintetizador 2 no quesito inteligibilidade foi penalizada pelo fato dela ser menos natural.

A Figura 4 permite visualizar que realmente existe uma correlação direta entre as notas atribuídas para os quesitos naturalidade e inteligibilidade da voz na avaliação subjetiva.

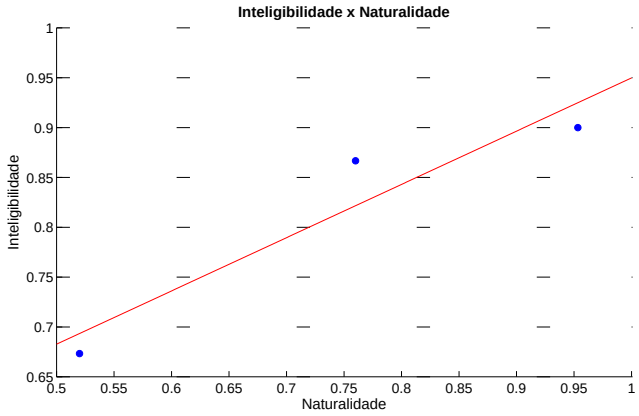

Figura 4. Correlação entre Naturalidade e Inteligibilidade.

Um outro resultado interessante que pode ser inferido observando-se a Figura 2, são as notas obtidas no teste de compreensibilidade. Nesse teste o sintetizador 1 demonstrou uma taxa de erro de 44% e o sintetizador 2 uma taxa de erro de cerca 20%, enquanto que a voz humana obteve acerto de praticamente 100%. Este resultado parece contraditório, ou ao menos insatisfatório à luz das notas obtidas pelas vozes artificiais nos testes de inteligibilidade.

Para melhor 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 segundo as palavras formam frases com significado.

Pode-se ver na Figura 5 que existe uma correlação direta entre a inteligibilidade e a compreensibilidade.

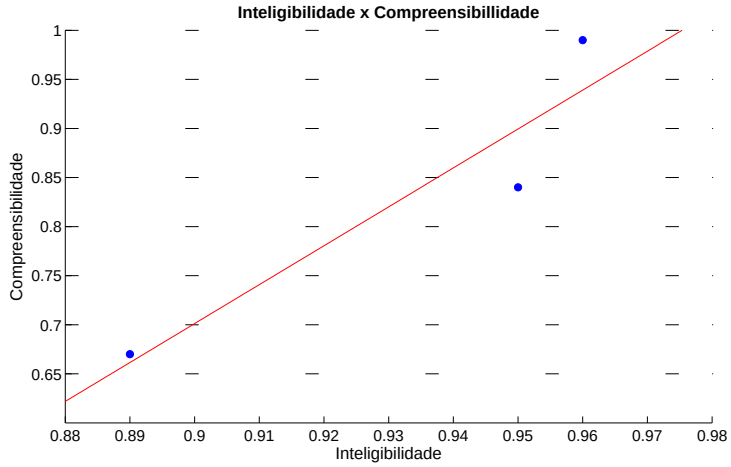

Figura 5. Correlação entre Compreensibilidade e Inteligibilidade.

Todavia, observa-se que o eixo Y exibe valores percentuais inferiores aos correspondentes no eixo X. A proporção entre eles são equivalentes somente para o caso da voz humana, para a qual ambas as variáveis possuem acerto acima de 95%. Isso vem corroborar o fato de que a compreensibilidade está associada a peculiaridades cognitivas que vão além da inteligibilidade.

5. Conclusões

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, indicando que o uso de vozes artificiais em aplicativos é plenamente possível.

É interessante notar que, entre as vozes artificiais, a voz do sintetizador 2 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 2 é inferior à do sintetizador 1 e que os ouvintes deram um peso importante na naturalidade da voz ao exprimirem suas preferências. Testes futuros devem ser aplicados com o intuito de identificar as características que influenciam a compreensibilidade do discurso, a fim de aprimorar as vozes sintéticas e torná-las tão compreensíveis quanto as vozes humanas.

Referências

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

Carlson, R., Granstrom, B., and Nord, L. (1990). Evaluation and development of the kth text-to-speech system on the segmental level. In *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*, pages 317–320 vol.1.

Georgila, K., Black, A., Sagae, K., and Traum, D. R. (2012). Practical evaluation of human and synthesized speech for virtual human dialogue systems. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 3519–3526, Istanbul, Turkey. European Language Resources Association (ELRA).

Goldstein, M. (1995). 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.

Heuven, V. J. V. and Bezooijen, R. V. (1995). Quality evaluation of synthesized speech. In Kleijn, W. B. and Paliwal, K. K., editors, *Speech Coding and Synthesis*. Elsevier Science.

Jekosch, U. (1993). Speech quality assessment and evaluation. In *Proceedings of Eurospeech 93*, volume 93, pages 1387–1394. ISCA.

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

Lemmetty, S. (1999). Review of speech synthesis technology. Master's thesis, Department of Electrical and Communications Engineering, Helsinki University of Technology.

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

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

Mariniak, A. (1993). A global framework for the assessment of synthetic speech without subjects. In *Proceedings of Eurospeech 93*, volume 93, pages 1683–1686. ISCA.

Martins, V. F. and Brasiliano, A. (2012). Interface do usuário baseada em voz como ferramenta para promover o ensino/aprendizagem de língua estrangeira. *Revista Eletrônica do Alto Vale do Itajaí (REAVI)*, 1(1):34–42.

Pisoni, D. B., Nusbaum, H. C., and Greene, B. G. (1985). Perception of synthetic speech generated by rule. *Proceedings of the IEEE*, 73(11):1665–1676.

Stevens, C., Lees, N., Vonwiller, J., and Burnham, D. (2005). On-line experimental methods to evaluate text-to-speech (TTS) synthesis: effects of voice gender and signal quality on intelligibility, naturalness and preference. *Computer Speech & Language*, 19(2):129–146.

Sydeserff, H. A., Caley, R. J., Isard, S. D., Jack, M. A., Monaghan, A. I. C., and Verhoeven, J. (1992). Evaluation of speech synthesis techniques in a comprehension task. *Speech Commun.*, 11(2-3):189–194.

Yu-Yun, C. (2011). Evaluation of tts systems in intelligibility and comprehension tasks. In *Proceedings of the 23rd Conference on Computational Linguistics and Speech Processing*, pages 64–78, Stroudsburg, PA, USA. Association for Computational Linguistics.