Um Sintetizador de Voz Baseado em HMMs Livre: Dando Novas Vozes para Aplicações Livres no Português do Brasil

Ericson Sarmento Costa[1], **Anderson de Oliveira Monte**[1], **Nelson Neto**[1], **Aldebaro Klautau**[1]

[1]Universidade Federal do Pará - UFPA
Rua Augusto Correa, 1 - 660750-110 - Belém, PA, Brasil
http://www.laps.ufpa.br/falabrasil

{ericson,aomonte,nelsonneto,aldebaro}@ufpa.br

Abstract. *Speech Synthesis is a mature field, with great results, and many differents techniques. One of most recent techniques is the HMM-based (Hidden Markov Models-based) approach, which is very interesting by its good results when using limited voice databases. This work aims to demonstrate the use of HMM-based methods applied to Brazilian Portuguese creating a TTS (Text To Speech) system and integrating it with the ORCA Screen Reader.*

Resumo. *A area de síntese de voz é uma área madura, com ótimos resultados, e diversas técnicas de implementação. Uma das técnicas mais recentes que tem chamado atenção da comunidade pela facilidade de aplicação e bons resultados é a técnica de síntese baseada em HMMs. Para aplicação dessa técnica podem ser utilizadas bases de voz de baixa qualidade, e com poucas amostras, e ainda assim, obter resultados de boa qualidade, o que facilita e muito a criação de variadas vozes para serem utilizadas em aplicações que fazem uso do recurso Texto para Fala. Este trabalho objetiva demonstrar o uso desta técnica no Português Brasileiro, através da ferramenta HTS, gerando um sistema Texto para Fala de boa qualidade. E como demonstração, o sistema Texto para Fala criado é integrado com o software ORCA, que é um excelente leitor de tela de domínio livre.*

1. Introdução

Sistemas TTS (”Text To Speech”) são sistemas que transformam um texto simples em voz falada. Estes sistemas são muito úteis do ponto de vista da interação entre homem e computador, pois dão uma dimensão mais natural e humana a interação. Podem ser acoplados como módulos em sistemas de diálogo e constituir o computador uma ferramenta de uso extremamente simples. Podem, também, ser utilizados como leitores de tela a fim de auxiliar deficientes físicos no uso do computador [ORCA 2012], [DOSVOX 2012].

A pesquisa acadêmica em sistemas TTS não é nova, mesmo para o Português Brasileiro, onde as técnicas mais empregadas são a síntese concatenativa e a síntese baseada em formantes. Estes trabalhos já alcançaram um alto grau de maturidade, gerando sistemas TTS de alta qualidade [Gomes et al. 1998], [Solewicz et al. 1994], [Egashira and Violaro 1995], [Albano and Aquino 1997], [Barbosa et al. 1999], [Seara et al. 2007].

Nos últimos anos, um método emergente, baseado em aprendizado de máquina, a síntese baseada em HMMs (”Hidden Markov Models”) [Yoshimura et al. 1999], tem se

mostrado promissor pela qualidade do resultado gerado e pela facilidade de aplicação, porque suporta o uso de bases de voz pequenas em comparação as demais técnicas, e de pior qualidade, como por exemplo, com uma base de gravação caseira. Além disso, a voz gerada no TTS fica muito similar à voz do locutor, o que dá ao sistema um ganho a mais em termos de interação, onde a aplicação que usa interface de voz pode ser melhor aceita por ter características da voz de alguma pessoa estimada.

Muitos trabalhos relacionados a síntese de voz baseada em HMMs têm sido realizados objetivando desenvolver aplicações para diversas línguas [Tokuda et al. 2002], inclusive para o Português Brasileiro [Maia et al. 2003], [Braga et al. 2010], [Couto et al. 2010]. Porém, estes trabalhos ou não são de domínio público [Braga et al. 2010], ou mesmo, como em [Couto et al. 2010], onde o *framework* utilizado é genérico demais, tentando atender a todas as línguas, gerando, assim, um ponto negativo, pois uma parte importante de um sistema TTS são seus módulos dependentes de linguagem, e este fator tem impacto direto na qualidade da síntese. Além disso, ainda, em [Couto et al. 2010], o *framework* utilizado não possui um cliente TTS *stand-alone*, e mesmo, este módulo cliente é acessível apenas através de interface-gráfica na forma como é distribuído, sendo necessário instalar toda a infra-estrutura do *framework* para que o cliente TTS possa funcionar, o que é um outro ponto negativo, e impossibilita a criação de aplicações embarcadas, e fácil integração com aplicações como no caso dos leitores de tela, por exemplo o leitor de tela ORCA.

Sendo assim, dada as vantagens do método de síntese baseada em HMMs, o objetivo desse trabalho é seguir a mesma linha e, ainda, estender o trabalho feito em [Couto et al. 2010], criando um sistema TTS *stand-alone*, puramente na plataforma Java, o qual possa ser integrado com facilidade a outros sistemas. E tenha um foco específico para o Português do Brasil, formando uma comunidade para dar manutenção e extensão.

Para demonstrar resultados foi criado um sistema TTS neste trabalho a partir de uma base de voz pobre, com poucas amostras (221 sentenças, 5 a 6 segundos de gravação cada), e gravação caseira, que obteve destaque em diversos quesitos subjetivos a frente de uma ferramenta comumente utilizada pela comunidade em geral [LIANE TTS 2012], bem como da versão de demonstração disponibilizada pelos desenvolvedores da técnica de síntese baseda em HMMs [HTS 2012]. O sistema TTS criado foi então integrado ao leitor de tela ORCA, afim de prover uma nova opção para este leitor de tela no Português do Brasil.

O trabalho está organizado da seguinte forma: Na seção 2, são apresentados os conceitos básicos relativos ao funcionamento de um sistema TTS genérico. Nas seção 3, são apresentadas as particularidades na criação de um sistema TTS baseado em HMMs. Na seção 4, tem-se uma pequena introdução ao leitor de tela ORCA. Na seção 5, são apresentadas algumas das atuais opções em sistemas TTS disponíveis para o Português do Brasil no sistema operacional Linux. Na seção 6, são apresentados alguns resultados: Primeiro, uma avaliação subjetiva do sistema TTS desenvolvido em relação ao sistema **LianeTTS**, e a versão demo para o Português do Brasil criada pelos desenvolvedores da técnica baseada em HMMs; Segundo, a integração do sistema TTS desenvolvido ao leitor de tela ORCA. Na seção 7, tem-se algumas conclusões acerca do trabalho realizado, e da experiência obtida. Na seção 8, propõe-se alguns dos trabalhos que podem vir a ser realizados futuramente.

2. Funcionamento Básico de um Sistema TTS

Um sistema TTS é comumente composto por duas partes:

- *Front-End*: Composto por módulos NLP (”Natural Language Processing”);
- *Back-End*: Composto por módulos de processamento de voz para a geração de voz sintetizada;

Pode-se ver na figura 1 um exemplo de um diagrama de bloco de um sistema TTS:

Figura 1. Diagrama de bloco de um sistema TTS.

2.1. Front-end

O *front-end* possui um conjunto de algoritmos que devem normalizar o texto [Kinoshita et al. 2006] , aplicar regras para conversão grafema-fonema [Siravenha et al. 2008], divisão silábica [Silva et al. 2006], marcação de sílaba tônica [Silva et al. 2008]. Estas informações são utilizadas para determinar características prosódicas da fala. No HTS (”HMM-based Speech Synthesis System”) [HTS 2012], ferramenta na qual este trabalho se baseia, as informações prosódicas são agrupadas em um arquivo chamado rótulo de contexto. Este arquivo determina informações de diversos níveis, como por exemplo: fonema, sílaba, palavra, frase. Em [Maia et al. 2006], pode-se encontrar a explicação detalhada de como são compostas as informações de contexto prosódico. Como exemplo, pode-se ver na figura 2 a informação prosódica referente apenas ao fone \p\ da palavra ”pesquisa”, no formato HTS:

```
y^sil-p+e=s/M2:1_3/
/S1:y_@y-0_@3+1_@2/S2:1_1/S3:1_16/S4:0_9/S5:0_2/S6:e
/W1:y_#y-function_#1+function_#1/W2:1_16/W3:0_0/W4:0_0
/P1:y_!y-16_!16+y_!y/P2:1_1
/U:16_$16_&1
```

Figura 2. Exemplo de informação prosódica referente ao fone \p\ na palavra ”pesquisa” no formato HTS

2.2. Back-end

O *back-end* possui um conjunto de filtros que recebem parâmetros amostrais de voz, juntamente com os rótulos de contexto prosódico para gerar a forma de onda que corresponde a pronúncia do texto. O HTS utiliza um módulo *back-end* denominado *hts_engine* [HTS_ENGINE 2012], com código original na linguagem C. Esse *back-end* foi portado para a linguagem Java a algum tempo [Schr et al. 2008], e essa versão, distribuída com a plataforma MARY TTS [MARY TTS 2012], que foi utilizada para compor o TTS *stand-alone* desse trabalho.

3. Construção de Um Sistema TTS Baseado em HMMs

O processo de construção de um sistema TTS baseado em HMMs divide-se em dois processos macro:

- Treinamento: No qual existe um conjunto de HMMs (uma para cada fonema) que serão treinadas com parâmetros amostrais da voz, e contextuais prosódicos, a fim de gerar um modelo que relaciona regras contextuais prosódicas, com parâmetros amostrais da voz. Este processo inclui alguns subprocessos, como:
    1. Geração dos rótulos de contexto para cada frase da base;
    2. Alinhamento forçado a nível de monofone para cada frase da base, o que é feito utilizando-se da ferramenta HVite presente no HTS;
    3. Re-sample dos arquivos de áudio, se necessário, e conversão para o formato RAW.
- Síntese : Em que módulos de NLP serão utilizados para gerar informações prosódicas de contexto, a fim de que as mesmas determinem a geração dos parâmetros amostrais da voz, que será a entrada para um filtro MLSA (filtro que gera aproximações de voz baseado em parâmetros amostrais) [Yoshimura et al. 1999], gerando assim a voz sintetizada.

Pode-se visualizar de forma geral os dois processos macro, e sua inter-relação, através da figura 3:

3.1. Particularidades do TTS baseado em HMMs Desenvolvido

Neste trabalho, para a etapa de treino, foi utilizado uma versão modificada dos *scripts* de treino de HMMs baseados na ferramenta HTS que vem no HTS-demo221 [HTS 2012]. Os scritps foram modificados porque vêm com parâmetros de treinamento para vozes de 16 kHz de frequência de amostragem. E no entanto, objetivou-se desenvolver uma voz de boa qualidade, assim foi escolhido criar um modelo de voz para 22,05 kHz.

Segundo [Yamagishi and Simon 2010], quanto maior for a frequência de amostragem usada para gravar as sentenças que compõem a base de treino, melhor é o resultado final. Isso se explica pelo fato de o modelo gerado pelo aprendizado de máquina conter mais informações, ser mais rico.

Os parâmetros que precisaram ser alterados foram os que segue:

- Fator *alpha*: Fator relacionado a distorção da fala. Este fator é diretamente dependente da frequência de amostragem, e em parte, dependente, também, de locutor [Tokuda et al. 1994];
- Ordem de análise *mel-cepstral*: A ordem de análise *mel-cepstral* define a quantidade de padrões que serão analisados por quadro, logo quanto maior a ordem, melhor será o resultado da análise. Porém, deve-se considerar que para baixas taxas de amostragem, como 8 kHz, pode ser até prejudicial uma análise muito grande, pois aumentando a ordem de análise não se estará acrescentando nenhuma riqueza nos padrões analisados. O ideal, advindo de determinação empírica, é uma ordem *mel-cepstral* de 12 a 16 para frequências de 8 kHz, de 20 a 24 para frequências de 16 kHz, e de 28 a 32 para frequências de 22,050 kHz. Ainda, sabe-se que o HTK (Toolkit para treinamento de HMMs, base do HTS) pode realizar análise *mel-cepstral* de sentenças de até 48 kHz, porém até o momento só foi analisado, neste trabalho, criação de modelos até 22,05 kHz.

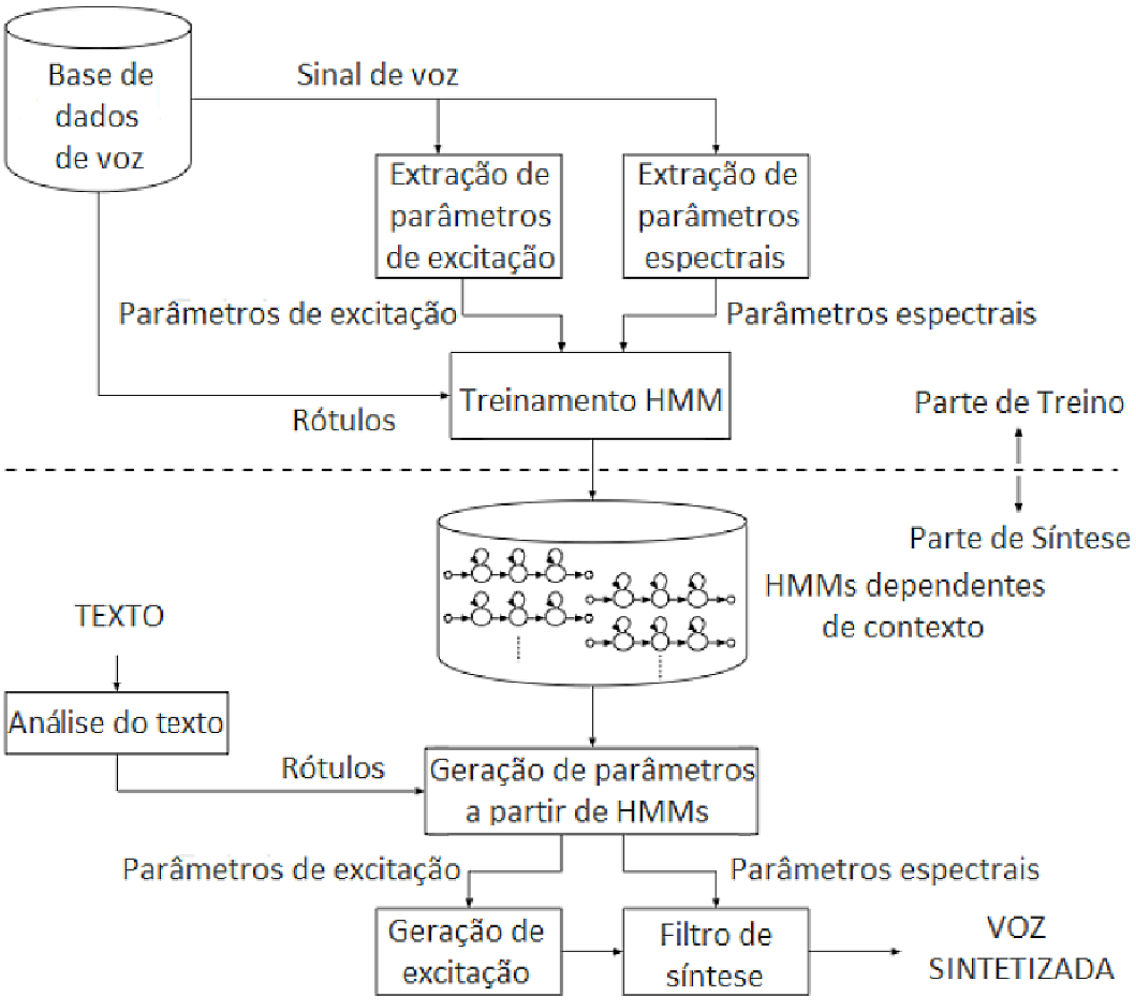

Figura 3. Diagrama de bloco geral dos passos que compõem a geração de um sistema TTS baseado em HMMs.

- Frame Shift: O *frame shift*, quando alterado na etapa de treino pode melhorar em parte o resultado do modelo gerado, ao exemplo da ordem de análise mel-cepstral. Na etapa de síntese, esse fator pode determinar uma fala mais rápida (apressada) ou mais lenta (preguiçosa). Estas observações foram feitas empiricamente neste trabalho.

4. Leitor de Tela Orca

Como aplicação do TTS gerado neste trabalho foi escolhido o leitor de tela ORCA por ser considerado pela comunidade, e por seus usuários um dos melhores leitores de tela livres para ambiente gráfico [ORCA 2012]. O projeto ORCA é bastante ativo, e recebe melhorias contínuamente. O ORCA é desenvolvido para o ambiente de janelas GNOME, suportante aplicações GTK2 e, mais recentemente, GTK+, portanto estando presente em diversas distribuições do sistema operacional Linux. Dentre as distribuições que possuem o leitor de tela ORCA disponível pode-se citar: Ubuntu, openSUSE, Fedora, Mandriva, Knoppix. Também pode-se considerar que algumas versões de sistemas UNIX possuem suporte ao ORCA, como por exemplo, o OpenSolares. Uma distribuição Linux voltada para acessibilidade que merece destaque é o LinuxAcessivel, baseado na distribuição

Ubuntu. A distribuição LinuxAcessível também utiliza o leitor de tela ORCA, como leitor de tela padrão. [LINUXACESSIVEL 2012]

5. Opções em sistemas TTS para o Português do Brasil com versão para o Ambiente Linux

A seguir, serão apresentadas algumas das opções para sistemas TTS no Português do Brasil que possuem versões para o ambiente Linux.

5.1. Voxin

Dada a falta de opções para sistemas TTS livres, de boa para ótima qualidade, que possam ser utilizados em aplicações, como os leitores de tela, uma opção foi recorrer a sistemas TTS não livres. Esse é o objetivo do projeto Voxin [VOXIN 2012]. Uma parceria com a IBM, para aquisição do sistema TTS IBM ViaVoice. O projeto Voxin reúne grupos de usuários que necessitem adquirir um sistema TTS, de boa para ótima qualidade, e cria vantagens para a aquisição destes sistemas. Assim atendendo a demanda. Estes sistemas então são integrados a ferramentas livres, e a sistemas operacionais livres, como é o caso do ORCA, no sistema Linux. Esta iniciativa por si só mostra uma grande deficiência que precisa ser resolvida, e uma demanda que precisa ser atendida, no mundo do software livre.

5.2. Sistemas TTS baseados na Engine MBROLA

O projeto MBROLA é um projeto inicialmente desenvolvido pela faculdade politécnica de Mons, na bélgica, e tem como objetivo criar uma engine de síntese de boa qualidade, baseada na técnica concatenativa e com aplicações em diversas línguas, inclusive o Português do Brasil [MBROLA 2012]. Oficialmente o projeto MBROLA disponibiliza 3 vozes diferentes para o Português do Brasil, são: br1, br2, br3. Estas todas são vozes masculinas. Abaixo tem-se alguns dos *Front-End* disponíveis para a *engine* MBROLA:

LianeTTS: Desenvolvido em uma parceria entre SERPRO e UFRJ para criar uma solução definitiva para acessibilidade em sistemas Linux, o qual passou a ser utilizado em larga escala em info-centros, através de projetos governamentais de inclusão digital [LIANE TTS 2012]. O projeto LianeTTS forneceu uma nova voz para o Português do Brasil, feminina, compatível com a engine MBROLA. Esta foi chamada de br4. LianeTTS consiste em um *Front-End* para a *engine* de síntese MBROLA, e scripts para integração ao leitor de tela ORCA, no sistema Linux, através do *driver* de voz *speech-dispatcher*.

Furb-Speech: O TTS Fub-Speech é um *Front-End* para a *engine* de síntese MBROLA. Desenvolvido na Faculdade de Blumenau, possui implementação na linguagem Java, e API (”Application Programing Interface”) simples [FURBSPEECH 2012]. Possui atualmente distribuição pública contendo as vozes para Português do Brasil br1, br2, br3, do projeto MBROLA.

Festival: O TTS Festival foi um TTS desenvolvido inicialmente pela Universidade de Edinburgh [FESTIVAL 2012], também é um *Front-End* para a *engine* de síntese MBROLA, bem como para muitas outras, sendo um TTS que utiliza diversas técnicas mas sempre com função principal de *Front-End*. Tem arquitetura voltada para ser um *framework* integrado, portanto não possuindo cliente TTS *stand-alone*. Possui uma versão de voz para o Português do Brasil baseada na síntese de formantes, porém esta é de domínio não livre. Comumente é utilizado em conjunto com as vozes disponibilizadas pelo projeto MBROLA para o Português do Brasil.

E-Speak: O TTS E-Speak é um TTS baseado em formantes, que abrange muitas línguas, dentre elas o Português do Brasil [ESPEAK 2012]. Alternativamente também pode ser utilizado como um *Front-End* para a *engine* de síntese MBROLA.

6. Resultados

6.1. Qualidade Subjetiva do TTS Desenvolvido

Para avaliar o sistema TTS deste trabalho foi desenvolvida uma voz de 22,05 kHz, a partir da gravação caseira da voz de um dos estudantes participantes do grupo de desenvolvimento. O motivo de se escolher uma gravação caseira é demonstrar a eficácia da técnica baseada em HMMs mesmo partindo de uma base de treino para TTS longe do ideal. Foram escolhidas 221 sentenças apenas, para que fosse comparado com o **HTS-demo221** [Maia et al. 2003], o qual utiliza a mesma quantidade de sentenças. Porém as sentenças utilizadas foram retiradas de [Cirigliano et al. 2005], utilizando-se das primeiras 221 sentenças listadas no trabalho. Ainda, foi incluído na avaliação para fins de comparação o TTS, baseado em técnica concatenativa, **LianeTTS**, que é suportado pela **SERPRO** (Empresa Federal Brasileira de Processamento de Dados) [LIANE TTS 2012]. Este TTS é baseado no projeto **MBROLA** [Dutoit et al. 1996].

Deve-se considerar que avaliação de vozes, e falas humanas é difícil de se fazer, porque entra o critério subjetivo do ouvinte. A opinião de quem ouve, portanto, é sempre o melhor critério de avaliação [ITU MOS 2012]. Nesse sentido, foram realizados diversos testes subjetivos, que utilizam notas de opinião direta de vários ouvintes, obedecendo uma escala, onde:

- A nota 1 representa a opinião ”Muito Ruim”;
- A nota 2 representa a opinião ”Ruim”;
- A nota 3 representa a opinião ”Razoável”
- A nota 4 representa a opinião ”Bom”;
- A nota 5 representa a opinião ”Excelente”;

Posteriormente é calculada a média dessas notas e então tem-se uma métrica conhecida como MOS (”Mean Opinion Score”), que representa a média das notas dadas como opinião. Esta métrica de base pode sofrer variações para se testar fatores específicos da comunicação, como foi feito nos critérios de avaliação deste trabalho.

Os critérios de avaliação utilizados foram os que segue:

- MOS para Naturalidade da fala: O ouvinte é convidado a ouvir uma fala, e tentar responder as seguintes perguntas, conforme a escala MOS: A voz é natural? É

produzida por um ser humano? É artificial? Quanto mais ela chega perto de ser natural?

- MOS para Inteligibilidade da fala: O ouvinte é convidado a ouvir uma fala, e tentar responder as seguintes perguntas, conforme a escala MOS: É possível entender o que está sendo dito? A mensagem está clara? Está difícil de compreender?
- WER (”Word Error Rate”) e WAR (”Word Accuracy Rate”) baseado em opinião: O ouvinte é convidado a expressar quantas palavras não consegue entender, ou estão muito difíceis de entender. Apesar de não utilizar a escala MOS, este teste leva em consideração que o ouvinte pode indicar no mínimo 0 (zero) palavras não entendidas, ou no máximo a quantidade de palavras total da frase.

Foram utilizadas ao todo 9 frases no teste, para que os participantes não ficassem muito cansados, ou se acostumassem com as vozes, o que alteraria muito o resultado do teste. No total participaram do teste 30 pessoas de idade variando de 17 a 48 anos, e de número equilibrado de sexos.

6.2. Naturalidade da fala

Para naturalidade da fala, a voz criada neste trabalho, chamada aqui de **Anderson221**, obteve uma considerável vantagem em relação ao **LianeTTS**, chamado aqui de **Mbrola-LianeTTS**, e ao **HTS-demo**, chamado aqui de **HTS-demo221**, sendo considerada quase uma voz humana.

Pode-se ver o resultado com mais facilidade na figura 4.

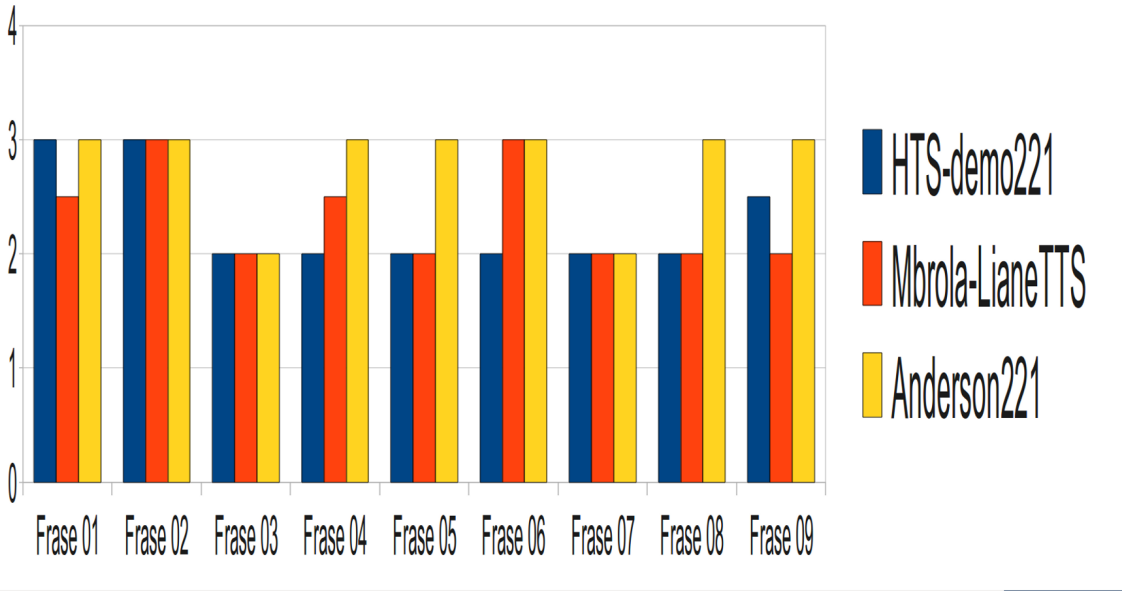

Figura 4. Gráfico de comparação para o critério naturalidade da fala.

6.3. Inteligibilidade da fala

Para o critério de inteligibilidade, a voz criada neste trabalho obteve um resultado, ainda melhor, em relação ao **Mbrola-LianeTTS**, e ao **HTS-demo221**, como se pode ver na figura 5.

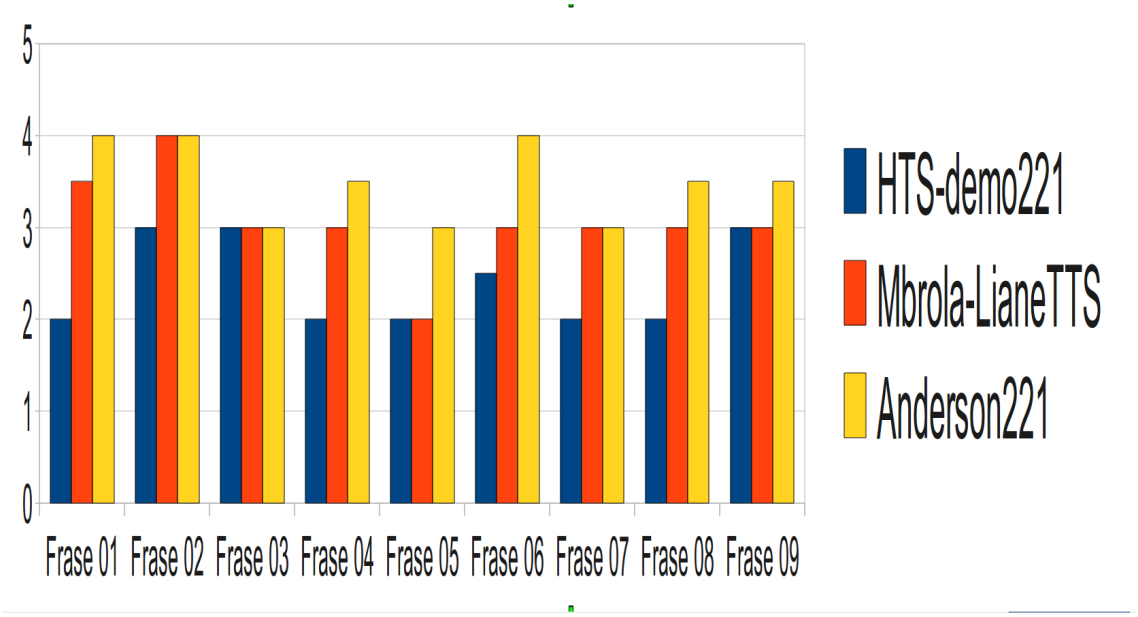

Figura 5. Gráfico de comparação para o critério inteligibilidade da fala.

6.4. WER e WAR

O WER representa o número de palavras não entendidas em relação ao total de palavras da frase, no teste subjetivo. O WAR representa o número total de palavras entendidas em relação ao total de palavras da frase.

O calculo da WER foi feito da seguinte forma:

$$WER = \frac{PE}{TP} * 100$$

Onde PE representa a quantidade de palavras entendidas subjetivamente como erradas, e TP representa a quantidade total de palavras da frase. Para o WAR foi utilizada a seguinte fórmula:

$$WAR = 100 - WER$$

Para todas as sentenças testadas, foi calculado o WAR. E pode-se ver na figura 6 que no resultado dessa métrica deu empate entre a voz gerada neste trabalho, **Anderson221**, e a **Mbrola-LianeTTS**, sendo ambas consideradas pelos ouvintes de fácil entendimento, de forma que todas as palavras foram entendidas por todos os candidatos. O **HTS-demo221**, teve apenas um pouco mais de 78% das palavras entendidas pelos ouvintes, na maioria das frases.

Uma amostra das vozes, bem como o teste que foi realizado pode ser encontrado neste endereço: *http://goo.gl/qwusP* . Onde, o locutor denominado **A** é o **HTS-demo221**, o locutor denominado **B** é a **Mbrola-LianeTTS**, e o locutor denominado **C** é o **Anderson221**.

6.5. Adaptação do TTS Desenvolvido ao Leitor de Tela Orca

Para a adaptação do TTS desenvolvido ao leitor de tela ORCA, foi utilizado o driver de voz *speech-dispacher* [SPEECH-DISPATCHER 2012]. A integração do *speech-distpacher* com qualquer TTS dá-se através de comando do *shell* especificado no arquivo

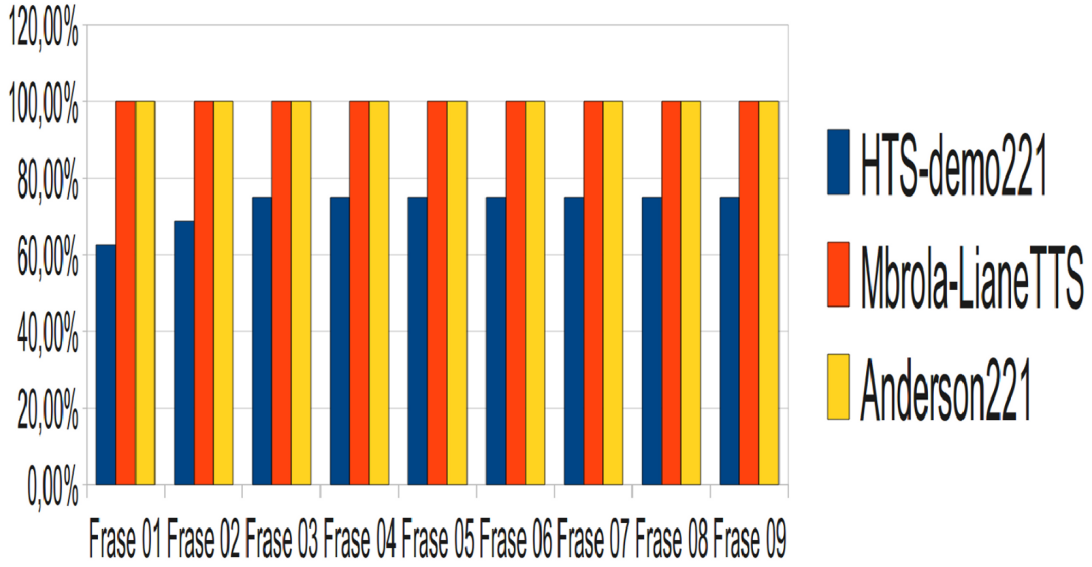

Figura 6. Gráfico de comparação para o critério WAR.

de configuração do driver, sendo a passagem da frase a ser sintetizada feita através de *pipe* entre comandos. Portanto, qualquer sintetizador que tenha interface em linha de comando, e possa receber um argumento via *pipe* pode ser integrado ao *speech-distpacher*.

7. Conclusão

Acredita-se que o desenvolvimento de vozes utilizando a técnica baseada em HMMs pode vir a resolver o problema de demanda por sistemas TTS de qualidade para aplicações livres no Português do Brasil. Os resultados demonstrados nesse trabalho foram satisfatórios, pois possibilitou a geração de um sistema TTS de boa qualidade, facilmente expansível por ser baseado em aprendizado de máquina, e portanto necessitando apenas de melhorias no *Front-End* e bases de voz de melhor qualidade. As bases que deveram ser utilizadas para compor os sistemas gerados a partir da técnica baseada em HMMs são de baixo custo, uma vez que seu tamanho é reduzido, evidenciando ainda mais o poder das técnicas de síntese de voz baseadas em aprendizado de máquina, e em destaque a técnica baseada em HMMs. Os diversas versões do TTS, e sua versão integrada ao ORCA durante seu desenvolvimento futuro devem ser disponibilizadas através do site do GRUPO FALABRASIL (*http://www.laps.ufpa.br/falabrasil*)

8. Trabalhos Futuros

Como trabalhos futuros espera-se gerar um modelo de voz baseado em HMMs de melhor qualidade, de forma a nivelar com os sistemas TTS comerciais. Também espera-se compor melhorias no *Front-End*, integrar o TTS com outras aplicações que utilizem interface de voz, principalmente na área de acessibilidade.

Referências

DOSVOX (2012). http://intervox.nce.ufrj.br/dosvox/.

ESPEAK (2012). http://espeak.sourceforge.net/.

FESTIVAL (2012). http://festvox.org/festival/.

FURBSPEECH (2012). http://code.google.com/p/furbspeech/.

HTS (2012). http://hts.ics.nitech.ac.jp/.

HTS_ENGINE (2012). http://sourceforge.net/projects/hts-engine/.

ITU MOS (2012). http://www.itu.int/rec/T-REC-P.800-199608-I/en.

LIANE TTS (2012). http://intervox.nce.ufrj.br/ serpro/home.htm.

LINUXACESSIVEL (2012). http://www.linuxacessivel.org/.

MARY TTS (2012). http://mary.opendfki.de/.

MBROLA (2012). http://tcts.fpms.ac.be/synthesis/.

ORCA (2012). http://live.gnome.org/Orca.

SPEECH-DISPACHER (2012). http://devel.freebsoft.org/speechd.

VOXIN (2012). http://voxin.oralux.net/.

Albano, E. and Aquino, P. (1997). Linguistic criteria for building and recording units for concatenative speech synthesis in brazilian portuguese. *Proceedings EuroSpeech, Rhodes, Grecia*, pages 725–728.

Barbosa, P., Violaro, F., Albano, E., Simes, F., Aquino, P., Madureira, S., and Franozo, E. (1999). Aiuruete: a high-quality concatenative text-to-speech system for brazilian portuguese with demisyllabic analysis-based units and hierarchical model of rhythm production. *Proceedings of the Eurospeechâ99, Budapest, Hungary*, pages 2059–2062.

Braga, D., Silva, P., Ribeiro, M., Dias, M. S., Campillo, F., and Garc´a-Mateo, C. (2010). Hélia, heloisa and helena: new hts systems in european portuguese, brazilian portuguese and galician. *PROPOR 2010 - International Conference on Computational Processing of the Portuguese Language*.

Cirigliano, R. J. R., Monteiro, C., de L. Barbosa, F. L., Resende, J. F. G. V., Couto, L. R., and Morais, J. A. (2005). Um conjunto de 1000 frases foneticamente balanceadas para o português brasileiro obtido utilizando e a abordagem de algoritmos genéticos. *Anais do Simpósio Brasileiro de Telecomunicações (SBrT)*.

Couto, I., Neto, N., Tadaiesky, V., Klautau, A., and Maia, R. (2010). An open source hmm-based text-to-speech system for brazilian portuguese. *7th international telecommunications symposium*.

Dutoit, T., Pagel, V., Pierret, N., Bataille, F., and VRECKEN, O. V. D. (1996). The mbrola project: Towards a set of high-quality speech synthesizers free of use for non-commercial purposes. *Proc. ICSLP'96, Philadelphia*, 3:1393–1396.

Egashira, F. and Violaro, F. (1995). Conversor texto-fala para a língua portuguesa. *13th Simpósio Brasileiro de Telecomunicações*, pages 71–76.

Gomes, L. D. C., Nagle, E., and Chiquito, J. (1998). Text-to-speech conversion system for brazilian portuguese using a formant-based synthesis technique. *SBT/IEEE International Telecommunications Symposium*, pages 219–224.

Kinoshita, J., Salvador, L. N., and Menezes, C. E. D. (2006). Cogroo: a brazilian-portuguese grammar checker based on the cetenfolha corpus. *The fifth international conference on Language Resources and Evaluation*.

Maia, R., Tokuda, K., Kitamura, T., Resende, F. G., and Zen, H. (2003). Towards the development of a brazilian portuguese text-to-speech system based on hmm. *Proc. of the European Conf. on Speech Communication and Technology (EUROSPEECH)*.

Maia, R., Zen, H., Tokuda, K., Kitamura, T., and and, J. F. G. V. R. (2006). An hmm-based brazilian portuguese speech synthesizer and its characteristics. *IEEE Journal of Communication and Information Systems*.

Schr, M., Charfuelan, M., Pammi, S., and Türk, O. (2008). The mary tts entry in the blizzard challenge 2008. *Proc. of the Blizzard Challenge 2008*.

Seara, I., Nicodem, M., Seara, R., and Junior, R. S. (2007). Classificação sintagmática focalizando a síntese de fala: Regras para o português brasileiro. *SBrT*, pages 1–6.

Silva, C. D., Lima, A., Maia, R., Braga, D., Morais, J. F., Morais, J. A., and Resende, J. F. G. V. (2006). A rule-based grapheme-phone converter and stress determination for brazilian portuguese natural language processing. *IEEE Int. Telecomm. Symposium (ITS)*.

Silva, D. C., Braga, D., and Resende, J. F. G. V. (2008). Separação das silabas e determinação da tonicidade no português brasileiro. *XXVI Simpósio Brasileiro de Telecomunicações (SBrT'08)*.

Siravenha, A., Neto, N., Macedo, V., and Klautau, A. (2008). Uso de regras fonológicas com de terminação de vogal tônica para conversão grafema-fone em português brasileiro. *7th International Information and Telecommunication Technologies Symposium*.

Solewicz, J., Alcaim, A., and Moraes, J. (1994). Text-to-speech system for brazilian portuguese using a reduced set of synthesis unit. *ISSIPNN*, pages 579–582.

Tokuda, K., Kobayashi, T., and Imai, S. (1994). Recursive calculation of mel-cepstrum from lp coefficients. *Technical Report of Nagoya Institute of Technology*.

Tokuda, K., Zen, H., and Black, A. (2002). An hmm-based speech synthesis applied to english. *IEEE Workshop in Speech Synthesis*.

Yamagishi, J. and Simon, K. (2010). Simple methods for improving speaker-similarity of hmm-based speech synthesis. *Proc. ICASSP 2010*.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (1999). Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis. *European Conf. on Speech Communication and Technology (EUROSPEECH)*.