Um Framework para Desenvolvimento de Sistemas TTS Personalizados no Português do Brasil

Ericson Sarmento Costa, Anderson de Oliveira Monte, Nelson Neto, Aldebaro Klautau
Universidade Federal do Pará - UFPA
Rua Augusto Correa, 1 - 660750-110 - Belém, PA, Brasil
{ericson, aomonte, nelsonneto, aldebaro}@ufpa.br

***Resumo*— Recentemente, no campo da síntese de voz, muitos novos resultados tem sido alcançados através de técnicas inovadoras baseadas em aprendizado de máquina. Essas técnicas são interessantes do ponto de vista da facilidade para criar novos exemplares de vozes para os sistemas de síntese de voz. Com estas técnicas é necessário um esforço muito menor para obtenção de corpora de voz se comparado com outras técnicas como a síntese concatenativa. Um bom exemplo é o método baseado em HMM ("Hidden Markov Models"), que tem gerado bons resultados em várias línguas. Nesse sentido o estágio atual de pesquisa e aplicação para esta técnica é promissor. Este trabalho tem como objetivo apresentar um *framework* para desenvolvimento de sistemas TTS (Texto para Fala) personalizados no Português Brasileiro, com o principal objetivo de ser simples o suficiente para ser facilmente utilizado pela comunidade em geral, além de ser de uso livre.**

***Palavras-Chave*— Hidden Markov Models, HMM, Processamento de voz, TTS.**

***Abstract*— Recently in speech synthesis field many new results have been achieved using innovative techniques based on machine learning such as HMM-Based (Hidden Markov Models), which has generated good results in several languages by reducing the dificulty and time to build new voices for TTS (Text To Speech) systems. This paper aims to present a *framework* for building custom TTS systems in Brazilian Portuguese, with the main goal of being simple enough to be easily used by the community in general, besides being free to use.**

***Keywords*— Hidden Markov Models, HMM, Speech Processing, TTS.**

I. 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 [1], [2].

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 [3], [4], [5], [6], [7]. Atualmente o trabalho acadêmico considerado de mais alta qualidade, a partir testes auditivos subjetivos, é o trabalho [8].

Nos últimos anos, um método emergente, baseado em aprendizado de máquina, a síntese baseada em HMMs ("Hidden Markov Models") [9], 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 técnica, e de pior qualidade. 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 baseda em HMMs têm sido realizados objetivando desenvolver aplicações para diversas línguas [10], inclusive para o Português Brasileiro [11], [12], [13]. Porém, estes trabalhos ou não são de domínio público [12], ou mesmo, como em [13], 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, em [13], o *framework* utilizado não possui um cliente TTS *stand-alone*, 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, por exemplo.

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 [13], onde com o *framework* proposto:

- Seja possível criar novas vozes para os sistema de forma muito simples;
- Esteja disponível um módulo TTS *stand-alone*, pequeno o suficiente para ser embarcável;
- O mesmo seja independente de plataforma através de implementação na linguagem Java;
- O mesmo possua API ("Application Programming Interface") simples.

Desta forma, espera-se que a quantidade de usuários desse método cresça na comunidade brasileira e muitas novas aplicações surjam.

Para demonstrar resultados foi criada uma voz, neste trabalho, a partir de poucas amostras (221 sentenças, 5 a 6 segundos de gravação cada), e com gravação caseira, que obteve destaque em diversos quesitos subjetivos a frente de uma ferramenta comumente utilizada pela comunidade em geral [14], bem como da versão de demonstração disponibilizada

pelos desenvolvedores da técnica de síntese baseda em HMMs [15].

O trabalho está organizado da seguinte forma: Na seção II, tem-se a arquitetura básica de um sistema TTS, seus módulos, e algumas particularidades destes módulos no sistema TTS baseado em HMMs. Na seção III, demonstra-se as etapas que compõem a criação de um sistema TTS baseado em HMMs, bem como caraterísticas do sistema TTS desenvolvido nesse trabalho. Na seção IV, demonstra-se o funcionamento do *framework* desenvolvido neste trabalho, seus módulos, e suas características principais. Na seção V, Avalia-se resultados do trabalho realizado em comparação com outros sistemas TTS, utilizando-se de diversas métricas subjetivas.

II. Funcionamento Básico de um Sistema TTS

Um sistema TTS é comumente composto por duas partes:

- Front-end: Que é composto por módulos NLP ("Natural Language Processing");
- Back-end: Que é 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:

Fig. 1. Diagrama de bloco de um sistema TTS.

A. Front-end

O front-end possui um conjunto de algoritmos que devem normalizar o texto [16] , aplicar regras para conversão grafema-fonema [17], divisão silábica [18], marcação de sílaba tônica [19]. Estas informações são utilizadas para determinar características prosódicas da fala. No HTS ("HMM-based Speech Synthesis System") [15], 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 [20], 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
```

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

B. 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 front-end denominado hts_engine [21], com código original na linguagem C. Esse back-end foi portado para a linguagem Java [22] a algum tempo, e essa versão, distribuída com a plataforma Mary TTS [23], que foi utilizada para compor o TTS *stand-alone* desse trabalho.

III. Construção de Um Sistema TTS Baseado em HMMs

O processo de construção de um sistema TTS baseado em HMMs divide-se em duas partes:

- 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;
- 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) [9], gerando assim a voz sintetizada.

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

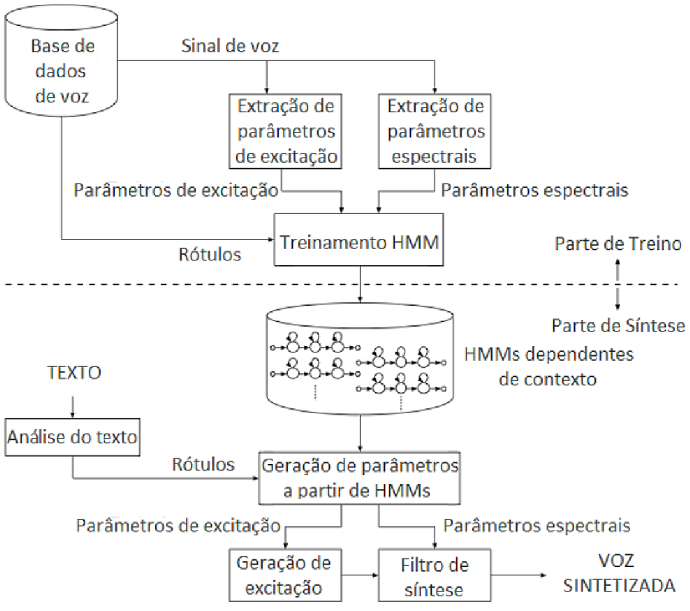

Fig. 3. Diagrama de bloco geral dos passos que compõem a geração de um sistema TTS baseado em HMMs.

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 [15]. 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.

Observa-se empiricamente que quanto maior a frequência de amostragem usada para as sentenças que compõem a base de treino, melhor é o resultado final [24]. 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 [25];
- 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 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.
- 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.

IV. Framework para Desenvolvimento de Sistemas TTS Personalizáveis no Português Brasileiro

O funcionamento do *framework* desenvolvido tem por finalidade dar liberdade ao usuário de se preocupar apenas com a aplicação que utiliza TTS. Portanto será disponibilizado modelos pré-treinados para que o usuário possa utiliza-lo de forma direta em suas aplicações. Serão disponibilizados dois modelos inicialmente, ambos de 22,05 kHz de frequência de amostragem, sendo um fruto de gravação caseira, e outro de gravação em estúdio.

Para a criação de novos modelos, o usuário deverá inserir um arquivo de áudio contínuo com a transcrição, que será segmentado automaticamente [26], ou ainda o usuário deverá gravar a voz, a partir de um conjunto de sentenças foneticamente balanceadas [27], [28] que é disponibilizado junto ao *framework*, ou não obstante, ainda poderá inserir a base já segmentada e com sua transcrição.

Pode-se ver na figura 4 o funcionamento em diagrama de blocos do *framework* proposto:

V. Resultados

Para avaliar o *framework* foi desenvolvida uma voz de 22,05 kHz, a partir da gravação caseira de um dos estudantes participantes do grupo de desenvolvimento. O motivo de se escolher uma gravação caseira é demonstrar a eficácia do *framework* 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** [11], o qual utiliza a mesma quantidade de sentenças. Porém as sentenças utilizadas foram retiradas de [28], 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) [14]. Este TTS é baseado no projeto **MBROLA** [29].

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 . 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.

A. Naturalidade da fala

Para naturalidade da fala, a voz criada neste trabalho, chamada aqui de **Anderson221**, obteve uma considerável

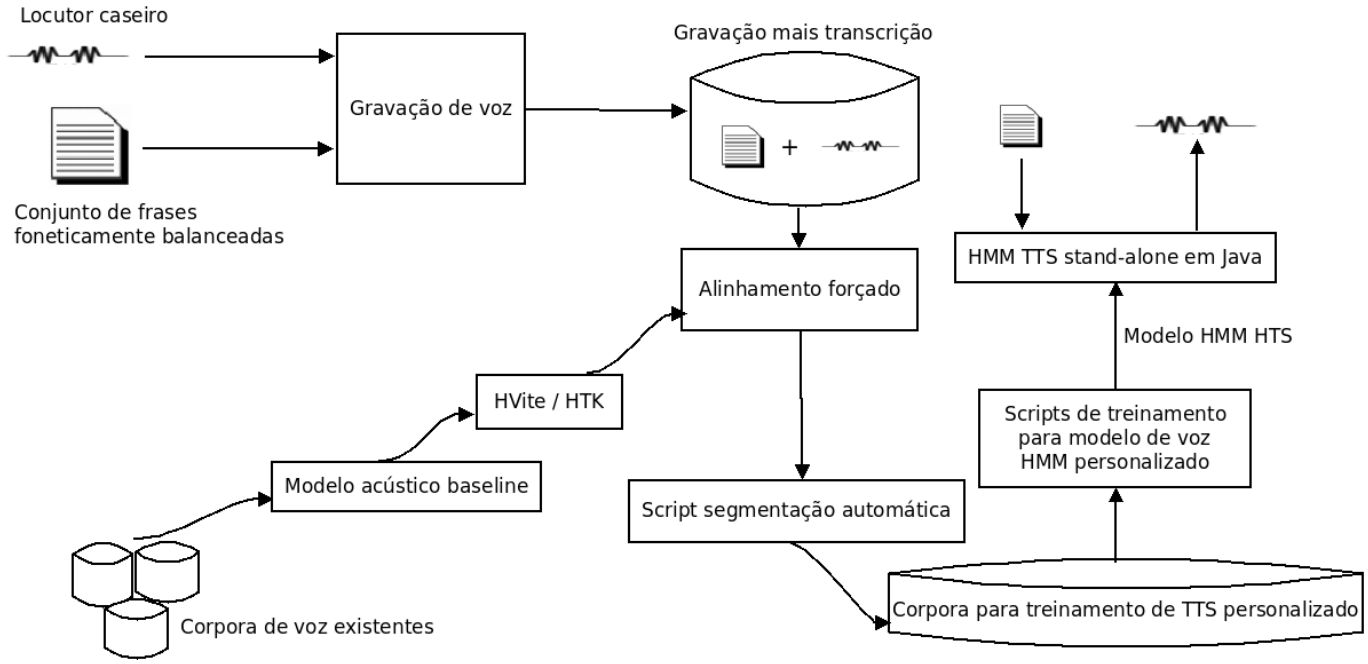

Fig. 4. Diagrama de bloco geral dos passos que compõem a geração de um sistema TTS baseado no framework proposto.

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 5.

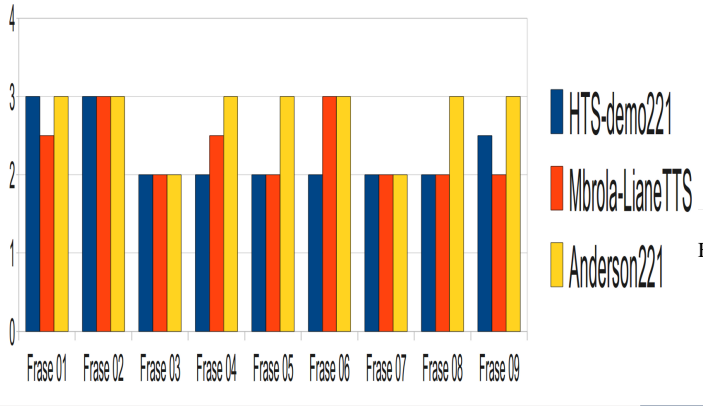

Fig. 5. Gráfico de comparação para o critério naturalidade da fala.

B. *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 6.

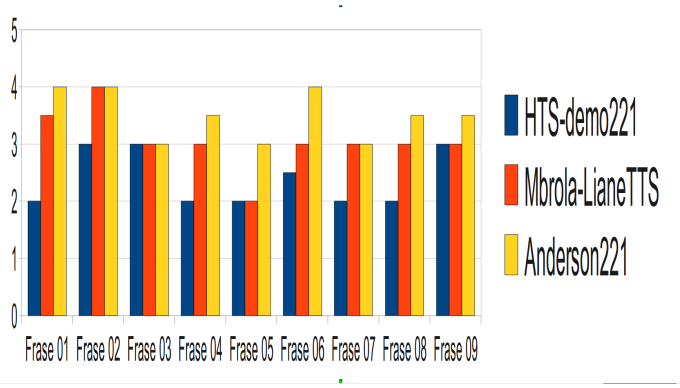

Fig. 6. Gráfico de comparação para o critério inteligibilidade da fala.

C. *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 7 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* .

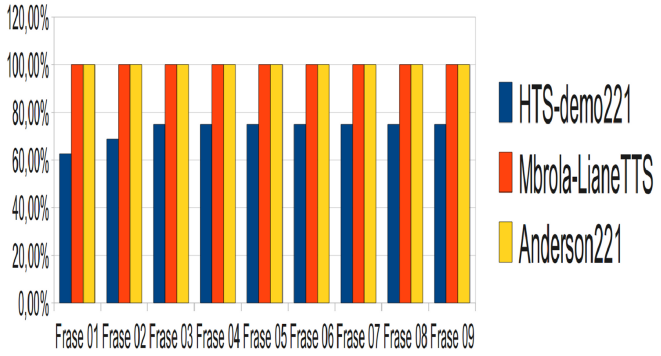

Fig. 7. Gráfico de comparação para o critério WAR.

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**.

VI. Conclusão

Acredita-se que o *framework*, quando estiver completamente disponibilizado para a comunidade em geral irá gerar uma onda de novos usuários e pesquisa relacionadas a síntese por HMMs no Português Brasileiro. Muitas novas aplicações úteis irão surgir. O *framework* é de domínio livre, portanto pode ser alterado ou expandido. Atualmente, este encontra-se em fase de teste *alpha*, e nos próximos meses deverá ser disponibilizado a comunidade em geral em sua versão *beta* na página do Grupo Fala Brasil *(http://www.laps.ufpa.br/falabrasil)*.

Os resultados alcançados com este trabalho mostraram-se satisfatórios, através da geração de um modelo de voz com qualidade de razoável para boa, tendo nível de Inteligibilidade e Naturalidade suficientes para ser utilizada como ferramenta, mesmo sendo apenas uma demonstração. Outras vozes ainda melhores devem ser produzidas na continuação do trabalho.

VII. Trabalhos Futuros

Como trabalhos futuros espera-se alcançar um nível ainda melhor de qualidade nos modelos de voz gerados de forma a nivelar com os sistemas TTS comerciais, através do treinamento de modelos utilizando gravações de estúdio, alta taxa de amostragem, e grande número de sentenças foneticamente balanceadas. Espera-se, também, desenvolver aplicações que utilizem síntese de voz, e ainda, posteriormente desenvolver sistemas de síntese de voz emotiva.

Referências

[1] (2012) ORCA HOME. [Online]. Available: http://live.gnome.org/Orca
[2] (2012) DOSVOX HOME. [Online]. Available: http://intervox.nce.ufrj.br/dosvox/
[3] L. De C.T. Gomes, E. Nagle, and J. Chiquito, "Text-to-speech conversion system for brazilian portuguese using a formant-based synthesis technique," *SBT/IEEE International Telecommunications Symposium*, pp. 219–224, 1998.
[4] J. Solewicz, A. Alcaim, and J. Moraes, "Text-to-speech system for brazilian portuguese using a reduced set of synthesis unit," *ISSIPNN*, pp. 579–582, 1994.
[5] F. Egashira and F. Violaro, "Conversor texto-fala para a língua portuguesa," *13th Simpósio Brasileiro de Telecomunicações*, pp. 71–76, 1995.
[6] E. Albano and P. Aquino, "Linguistic criteria for building and recording units for concatenative speech synthesis in brazilian portuguese," *Proceedings EuroSpeech, Rhodes, Grecia*, pp. 725–728, 1997.
[7] P. Barbosa, F. Violaro, E. Albano, F. Simes, P. Aquino, S. Madureira, and E. Franozo, "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*, pp. 2059–2062, 1999.
[8] I. Seara, M. Nicodem, R. Seara, and R. S. Junior, "Classificação sintagmática focalizando a síntese de fala: Regras para o português brasileiro," *SBrT*, pp. 1–6, 2007.
[9] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis," *European Conf. on Speech Communication and Technology (EUROSPEECH)*, 1999.
[10] K. Tokuda, H. Zen, and A. Black, "An hmm-based speech synthesis applied to english," *IEEE Workshop in Speech Synthesis*, 2002.
[11] H. Z. R. Maia, K. Tokuda, T. Kitamura, F. G. Resende, and H. Zen, "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)*, 2003.
[12] D. Braga, P. Silva, M. Ribeiro, M. S. Dias, F. Campillo, and C. Garc´a-Mateo, "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*, 2010.
[13] I. Couto, N. Neto, V. Tadaiesky, A. Klautau, and R. Maia, "An open source hmm-based text-to-speech system for brazilian portuguese," *7th international telecommunications symposium*, 2010.
[14] (2012) LIANE TTS HOME. [Online]. Available: http://intervox.nce.ufrj.br/~serpro/home.htm
[15] (2012) HTS HOME. [Online]. Available: http://hts.ics.nitech.ac.jp/
[16] J. Kinoshita, L. N. Salvador, and C. E. D. Menezes, "Cogroo: a brazilian-portuguese grammar checker based on the cetenfolha corpus," *The fifth international conference on Language Resources and Evaluation*, 2006.
[17] A. Siravenha, N. Neto, V. Macedo, and A. Klautau, "Uso de regras fonológicas com de terminação de vogal tônica para conversão grafema-fone em português brasileiro," *7th International Information and Telecommunication Technologies Symposium*, 2008.
[18] C. D. Silva, A. Lima, R. Maia, D. Braga, J. F. Morais, J. A. Morais, and F. G. V. R. Jr., "A rule-based grapheme-phone converter and stress determination for brazilian portuguese natural language processing," *IEEE Int. Telecomm. Symposium (ITS)*, 2006.
[19] D. C. Silva, D. Braga, and F. G. V. R. Jr, "Separação das silabas e determinação da tonicidade no português brasileiro," *XXVI Simpósio Brasileiro de Telecomunicações (SBrT'08)*, 2008.
[20] R. Maia, H. Zen, K. Tokuda, T. Kitamura, J. Resende, and F. G. V. Jr., "An hmm-based brazilian portuguese speech synthesizer and its characteristics," *IEEE Journal of Communication and Information Systems*, 2006.
[21] (2012) HTS_ENGINE HOME. [Online]. Available: http://sourceforge.net/projects/hts-engine/
[22] M. Schr, M. Charfuelan, S. Pammi, and O. Türk, "The mary tts entry in the blizzard challenge 2008," *Proc. of the Blizzard Challenge 2008*, 2008.
[23] (2012) MARY TTS Home. [Online]. Available: http://mary.opendfki.de/
[24] J. Yamagishi and K. Simon, "Simple methods for improving speaker-similarity of hmm-based speech synthesis," *Proc. ICASSP 2010*, 2010.
[25] K. Tokuda, T. Kobayashi, and S. Imai, "Recursive calculation of mel-cepstrum from lp coefficients," *Technical Report of Nagoya Institute of Technology*, 1994.
[26] (2012) AUTOMATIC SEGMENTATION. [Online]. Available: http://www.voxforge.org/home/dev/autoaudioseg
[27] A. Alcaim, J. A. Solewicz, and J. A. de Morais, "Frequência de ocorrência dos fones e listas de frases foneticamente balanceadas para o português falado no rio de janeiro," *Revista da Sociedade Brasileira de Telecomunicacoes*, vol. 7, no. 1, pp. 23–41, 1992.
[28] R. J. R. Cirigliano, C. Monteiro, F. L. de L. Barbosa, F. G. V. R. Jr., L. R. Couto, and J. A. de Morais, "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)*, 2005.
[29] T. Dutoit, V. Pagel, N. Pierret, F. Bataille, and O. V. D. VRECKEN, "The mbrola project: Towards a set of high-quality speech synthesizers free of use for non-commercial purposes," *Proc. ICSLP'96, Philadelphia*, vol. 3, pp. 1393–1396, 1996.