Análise comparativa do impacto da classe gramatical em sistemas TTS baseados em HMMs

Filipe Leandro de F. Barbosa[1], Ranniery da S. Maia[2] e Fernando Gil V. Resende Jr.[1,3]

***Resumo*— Este trabalho apresenta uma análise comparativa que tem como objetivo investigar a influência da classe gramatical das palavras em sistemas de síntese de fala para o português brasileiro baseados em modelos escondidos de Markov. Três versões do sintetizador foram comparadas: (1) com informação completa de classe gramatical; (2) somente com a diferenciação entre palavra de conteúdo e palavra de contexto e (3) sem qualquer informação de classe gramatical. Os resultados indicam que a presença da informação de classe gramatical traz pouca influência na qualidade da voz sintetizada. A versão com distinção entre palavra de conteúdo e de contexto leva a resultados similares aos obtidos quando utilizada a informação gramatical completa, com um menor custo computacional.**

***Palavras-Chave*— Processamento de voz, sistemas TTS, classe gramatical, HMM.**

***Abstract*— This paper presents a comparative test which investigates the influence of part-of-speech information on Brazilian Portuguese text-to-speech systems based on hidden Markov models. Three synthesizer versions were tested:(1) with full part-of-speech information; (2) only with the distinction between content and function words; and (3) without any part-of-speech information. The results indicate that part-of-speech information has a small influence on the quality of the synthesized speech. The synthesizer version where there is only the distinction between content and function words obtains results similar to the version with full information, although with lower computational effort.**

***Keywords*— Speech processing, TTS systems, part-of-speech, HMM.**

I. INTRODUÇÃO

A síntese de voz a partir de texto (*text-to-speech* - TTS) é uma área que vem despertando bastante interesse de pesquisadores da área de processamento de voz. A maior parte dos sintetizadores utiliza a concatenação de formas de onda. Contudo, quando se deseja sintetizar voz com vários estilos, tal como voz feminina, infantil, voz triste, alegre, etc, é necessária uma grande quantidade de dados de voz. Por outro lado, sistemas TTS baseados em modelos escondidos de Markov (*hidden Markov models* - HMMs) possuem a vantagem de não necessitar de grandes quantidades de dados de voz para realizar mudanças nos estilos da voz sintetizada. Segundo reportado em [1], dez frases podem, por exemplo, realizar a adaptação a um diferente locutor com uma boa fidelidade.

Neste trabalho, um sintetizador baseado em HMMs para o português brasileiro [2] é utilizado para investigação do impacto da utilização da classe gramatical das palavras (*part-of-speech* - POS) na qualidade da voz sintetizada. Em [3] investigou-se a influência de parâmetros como sílabas e POS. Foi verificado que a informação de POS apresentava uma pequena influência na avaliação qualitativa. Neste artigo, é apresentada uma alternativa à utilização de POS que mostrou resultados semelhantes em termos de avaliação subjetiva, mas cuja complexidade de implementação é muito reduzida.

Este artigo está organizado como segue. Na Seção 2 a técnica de síntese de voz baseada em HMMs é apresentada; na Seção 3 o sintetizador para o português brasileiro é descrito; na Seção 4 são apresentados os testes experimentais; na Seção 5 os resultados são discutidos; a conclusão e os trabalhos futuros são colocados na Seção 6.

[1]Departamento de Eletrônica e de Computação, Escola Politécnica, UFRJ. [2]Department of Computer Science and Engineering, Nagoya Institute of Technology, Japão. [3]Programa de Engenharia Elétrica, COPPE, UFRJ. E-mails: filipe@lps.ufrj.br, maia@ics.nitech.ac.jp, gil@lps.ufrj.br.

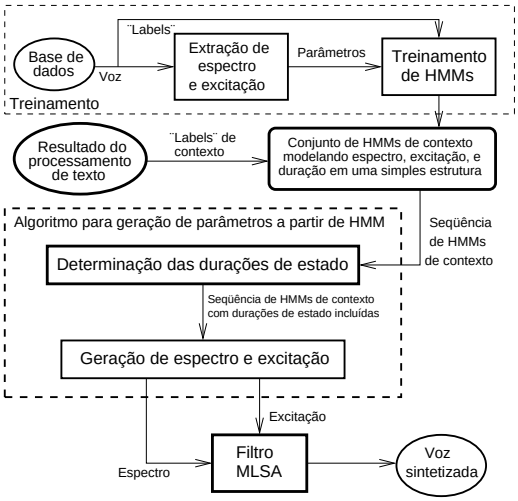

Fig. 1. Diagrama da técnica de síntese de voz baseada em HMM.

II. SINTETIZADOR BASEADO EM HMMS

A técnica de síntese de voz baseada em HMMs pode ser resumida pela Figura 1.

A. Treinamento de HMMs

Na parte de treinamento, os parâmetros de espectro e excitação são extraídos de uma base de voz, levando ao treinamento de HMMs. Na síntese de voz baseada em HMMs (*HMM-based speech synthesis* - HSS), cada HMM corresponde a uma máquina de estados finita, da esquerda para a direita, sem pulos, onde cada vetor de saída é composto da parte espectral, representada pelos coeficientes mel-cepstrais e suas características dinâmicas (delta e delta-delta),

TABELA I
VERSÕES UTILIZADAS E TESTES REALIZADOS

Versões utilizadas	
Versão 1	Sem informação de POS
Versão 2	Com informação de POS
Versão 3	Com distinção entre palavra de conteúdo e de contexto
Testes realizados	
Teste 1	Versão 1 x Versão 3
Teste 2	Versão 2 x Versão 3

TABELA II
SENTENÇ AS UTILIZADAS NO TESTE COMPARATIVO.

Sentenç as	
1	O consumo de laticínio pode proteger da obesidade.
2	O presidente deve passar o dia reunido.
3	As novas medidas seriam anunciadas.
4	O secretário-geral disse que será suspenso.
5	Não se falou em antecipar as eleições.
6	Ronaldinho foi destaque dos amistosos.
7	A Igreja vive uma onda de denúncias.
8	Eles possuem quarenta mil empregados.
9	As aves serão analisadas em um laboratório.
10	As preocupações se referem aos interceptadores.

e da excitação, representada por parâmetros relacionados à freqüência fundamental, e seus correspondentes coeficientes delta e delta-delta. Para reproduzir a estrutura temporal da fala, para cada HMM as durações dos estados ocorrem de acordo com distribuições de densidade de probabilidade, que são por sua vez re-estimadas durante o treinamento dos HMMs. Assim, espectro, excitação e duração de estado são modelados conjuntamente.

B. Modelamento do Espectro

O espectro é modelado por coeficientes mel-cepstrais que podem sintetizar voz de forma direta utilizando o filtro MLSA (Mel Log Spectrum Approximation) [4]. Os coeficientes mel-cepstrais são extraídos de uma base de dados de voz a partir de uma análise espectral de tempo curto. A probabilidade de saída de cada conjunto de coeficientes mel-cepstrais corresponde a uma distribuição gaussiana de vetores de variáveis [5].

C. Modelamento da Excitação

Os parâmetros de excitação são formados pelo logaritmo da freqüência fundamental F0, log(F0), e suas correspondentes características dinâmicas. Como a seqüência de observação de F0 é geralmente composta de valores contínuos e unidimensionais, e também de um símbolo que representa as regiões não vozeadas do contorno de F0, a abordagem por distribuição de probabilidade contínua não é capaz de modelar propriamente as seqüências de F0. Para contornar tal problema, a distribuição de probabilidade por soma de subespaços é aplicada [5].

D. Agrupamento de contextos através de árvores de decisões

Durante o treinamento de um sistema de HSS, a base de voz pode não incluir um número adequado de exemplos para todos os modelos de contexto existentes na base de dados, de forma a permitir uma boa estimação dos HMMs. Ademais, também durante a síntese, às vezes um *label* de contexto a ser sintetizado não tem o HMM correspondente no conjunto disponível de modelos treinados. Para resolver esses dois problemas, uma técnica de agrupamento de contextos através de árvores de decisões é aplicada nas distribuições de espectro, F0 e duração de estado [2]. Tais parâmetros são separadamente analisados por sofrerem influências distintas dependentes do contexto. Conseqüentemente, diferentes árvores de decisões são criadas para coeficientes mel-cepstrais, F0, e duração de estado [5].

E. O procedimento de síntese

O procedimento de síntese é descrito a seguir. Primeiramente, um dado texto a ser sintetizado é convertido em uma seqüência de *labels* de contexto [2]. Então, de acordo com esta seqüência, uma seqüência de HMMs é construída por concatenação dos correspondentes HMMs dependentes do contexto. Logo após, são determinadas as durações de estados para cada HMM de tal forma que a probabilidade de saída das durações de estado seja maximizada. A partir da seqüência de HMM com as durações de estado incluídas, uma seqüência de vetores de coeficientes mel-cepstrais e valores de log(F0), incluindo as decisões vozeado/não vozeado, são gerados utilizando o caso 1 do algoritmo apresentado em [6].

III. SINTETIZADOR PARA O PORTUGUÊS BRASILEIRO

A. Gravação, segmentação e etiquetagem da base de dados de voz

Uma base de dados de voz com 221 sentenças, incluindo 200 frases foneticamente balanceadas para o português brasileiro (PB) falado no Rio de Janeiro [7], foi gravada por um brasileiro, a uma taxa de 48 kHz, com 16 bits por amostra. A base de voz foi posteriormente decimada para 16 kHz.

O processo de etiquetagem foi realizado utilizando o HTK (*HMM toolkit*) [8], conforme descrito em [2]. O procedimento de etiquetagem foi feito em duas etapas. A primeira consistiu da transcrição automática dos grafemas em fones, realizada por um transcritor grafema-fone [9]. Na segunda etapa, foi feita a correção manual da etiquetagem realizada, devido a diferença existente entre o conjunto de fones descritos em [9] e o aplicado ao sintetizador utilizado. No total, são utilizados 38 unidades acústicas [2], excluindo os modelos de silêncio.

B. Informação contextual

Um dos fatores dependentes da língua, quando tratamos de sistemas TTS baseados em HMMs, é a entrada de informação contextual, utilizada para a determinação do HMM a ser escolhido no conjunto previamente treinado.

Tal informação contextual contribui para uma boa reprodução da prosódia. Contudo, quando a informação contextual obtida não corresponde a nenhum HMM da base treinada, as árvores de decisão geradas de acordo com a técnica de agrupamento de contextos, durante o treinamento

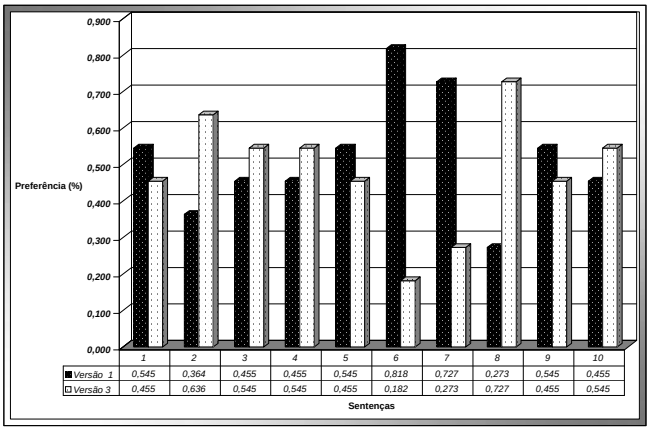

Fig. 2. Histograma do Teste 1 com relação às sentenç as. Versão 1 em preto e Versão 3 em branco.

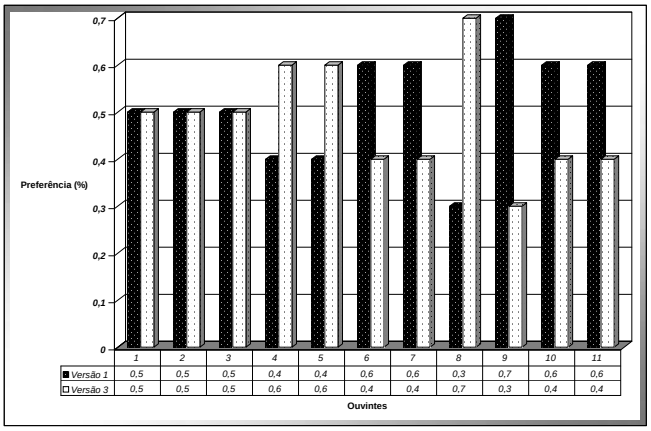

Fig. 4. Histograma do Teste 1 com relação aos ouvintes. Versão 1 em preto e Versão 3 em branco.

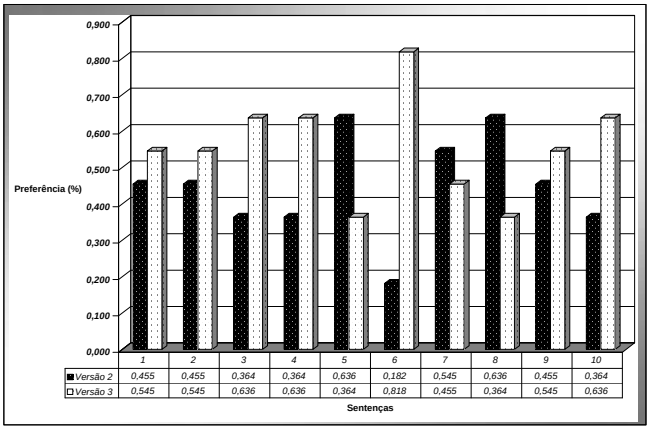

Fig. 3. Histograma do Teste 2 com relação às sentenç as. Versão 2 em preto e Versão 3 em branco.

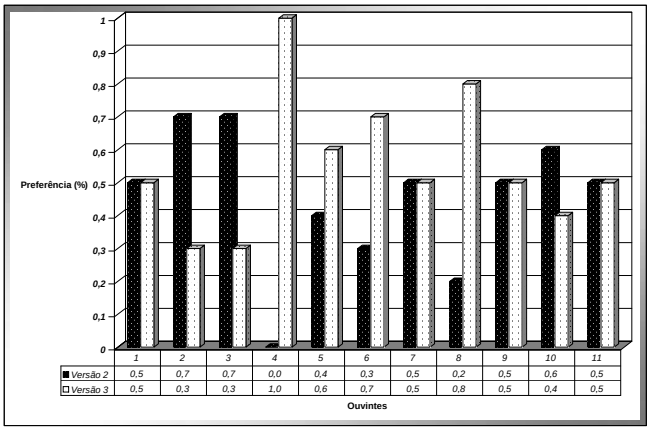

Fig. 5. Histograma do Teste 2 com relação aos ouvintes. Versão 2 em preto e Versão 3 em branco.

do sintetizador, são aplicadas para gerar o respectivo modelo não visto. Dentre as informações contextuais que são levadas em consideração para o sintetizador utilizado, está a informação referente à classe gramatical das palavras. Para que tal informação seja obtida automaticamente é necessária a implementação de um analisador morfológico. Buscando verificar a necessidade de tal implementação, a investigação do impacto da POS em isolado se faz necessária.

IV. Testes Experimentais

Para os testes experimentais foram utilizadas três versões do sintetizador. A primeira versão (Versão 1), sem informação de POS. A segunda versão (Versão 2), com informação completa de POS inserida manualmente. Na última versão (Versão 3), foi implementada uma identificação de palavras de conteúdo e de palavras de contexto como uma alternativa à POS. As palavras de contexto são identificadas como: artigos, preposições, pronomes, numerais, conjunções, interjeições e contrações (preposições com artigos). As palavras de conteúdo são dadas por: verbos, substantivos, adjetivos e advérbios. Como o conjunto de palavras de conteúdo é muito maior, foi desenvolvido um programa para automaticamente distinguir entre palavras de contexto e de conteúdo, baseado na listagem das palavras de contexto. Assim, uma alternativa simplificada à segunda versão do sintetizador foi criada e automatizada.

Foram realizados dois testes experimentais. O primeiro teste (Teste 1) foi feito comparando-se a primeira e a terceira versões do sintetizador. O segundo teste (Teste 2) foi realizado pela comparação da segunda e da terceira versões do sintetizador. A Tabela I identifica as versões do sintetizador utilizadas, bem como os testes realizados.

Cada um dos testes foi realizado com um conjunto de 10 frases, mostradas na Tabela II, que não foram utilizadas no treinamento do sintetizador. Para cada teste cada a pessoa ouviu dez conjuntos de duas elocuções cada. Para cada conjunto a pessoa foi instruída para apontar qual das duas elocuções apresentava melhor qualidade. No Teste 1 foi reproduzido em cada conjunto uma mesma frase nas versões 1 e 3, enquanto que no Teste 2 foi feita a reprodução de uma frase nas versões 2 e 3 do sintetizador. O conjunto poderia ser reproduzido até que a pessoa se sentisse confortável para indicar qual a elocução preferida. Como havia uma alternância aleatória entre as diferentes versões do sintetizador nos dois testes, caso a pessoa declarasse impossibilidade de detectar qual a frase escolhida, a primeira opção foi sempre a apontada.

Durante os testes foi utilizado um fone de ouvido. Além disso, todos os testes foram executados em um mesmo computador. Um total de 11 pessoas, 10 homens e uma mulher, realizaram os dois testes, ouvindo 10 conjuntos de duas sentenças por teste, totalizando 40 sentenças por pessoa.

TABELA III
RESULTADO GLOBAL DO TESTE COMPARATIVO

Versão 1	Versão 3
52%	48%
Versão 2	Versão 3
44%	56%

Dentre os ouvintes, dois trabalham na área de síntese de voz. O restante é formado por alunos de graduação e pós-graduação familiarizados com a área de processamento de sinais.

V. RESULTADOS

Com relação aos resultados globais dos Testes 1 e 2, a Tabela III mostra que houve uma preferência de 52% para Versão 1, contra 48% para Versão 3. Na comparação das Versões 2 e 3 do sintetizador, houve 44% de preferência para a Versão 2 e 56% para a versão proposta como alternativa. Tais resultados ratificam a pouca influência da POS como já verificado em [3], além de confirmar ainda que a Versão 3 é uma alternativa, com baixa complexidade de implementação, à Versão 2. O resultado global confirma o fato da maioria das pessoas não ter percebido a diferença entre as versões 2 e 3, já que a Versão 3 foi escolhida em 56% das vezes, apesar de não conter toda a informação de POS.

Mais detalhadamente, os resultados obtidos com os dois testes realizados podem ser observados nas Figuras 2 e 3. Nessas figuras, os resultados foram apresentados por sentenças, mostradas na Tabela II. Na maioria dos casos, a preferência ficou bem dividida entre as versões, em ambos os testes.

Para o teste da Figura 2 a maioria das preferências ficou situada em 50% ou no seu entorno. Apenas as sentenças 6 e 7 apresentaram um desequilíbrio maior, tendendo para a Versão 1, que não possui informação contextual de POS. Isso mostra a pouca influência exercida pela Versão 3 para essas sentenças, o que levou os ouvintes a não perceberem a diferença entre as versões, acabando por optar pela versão com menos informação de contexto. Por outro lado, a preferência dos ouvintes na sentença 8 foi favorável à Versão 3.

A Figura 3 mostra os resultados do Teste 2. Nesse caso houve uma maior preferência para a Versão 3 em detrimento da Versão 2, apesar de em alguns casos haver pouca distinção entre a preferência das versões, como nos casos 1, 2, 7 e 9. Tal resultado corrobora novamente a pouca influência da POS, visto que os ouvintes, em sua maioria, escolheram a versão mais simplificada do sintetizador. Além disso, indica que o impacto da substituição da Versão 2 pela Versão 3 não é percebido pela maior parte dos ouvintes.

As Figuras 4 e 5 mostram as preferências de cada um dos 11 ouvintes, para os testes 1 e 2. Na Figura 4 podemos notar que os ouvintes 1, 2 e 3, apresentaram resultados idênticos, não apontando tendência para nenhuma das versões. As escolhas dos outros ouvintes se compensaram, justificando o resultado global apresentado na Tabela III, para o Teste 1. A Figura 5 mostra que para o Teste 2, os ouvintes deram preferência à Versão 3, apesar dos ouvintes 1, 7, 9 e 11 não apresentarem tendência a preferir nenhuma das versões. Tal distribuição justifica o resultado da Tabela III, para o Teste 2, além de enfatizar a possibilidade de se aplicar a alternativa à POS proposta.

Um fato que merece destaque na Figura 5 é que apenas dois ouvintes (2 e 3) aparentemente detectaram diferença considerável a favor da Versão 2. Em contrapartida, três ouvintes (4, 6 e 8) optaram pela Versão 3, sendo que o ouvinte 4 escolheu a Versão 3 para todas as frases que ouviu e o ouvinte 8 apontou para a Versão 3 em 8 dos 10 conjuntos de sentenças.

Contudo, é importante notar que os testes realizados neste trabalho foram feitos com frases isoladas e relativamente curtas. É possível que a informação de POS tenha mais influência na síntese da fala a partir de textos com frases mais longas ou com diversas frases em seqüência.

VI. CONCLUSÕES

Este trabalho apresenta uma alternativa à utilização da classe gramatical das palavras em um sintetizador de voz baseado em HMMs. Resultados mostram que a abordagem adotada, que distingue palavras de conteúdo de palavras de contexto, não traz prejuízo à qualidade da voz sintetizada, podendo substituir a informação de POS, que é mais difícil de ser automatizada e tem maior custo computacional.

O estudo da influência da informação de POS em textos longos, assim como o impacto de questões relativas à semântica, são objetos de pesquisa em andamento.

REFERÊNCIAS

[1] T. Masuko, K. Tokuda, T. Kobayashi, e S. Imai. *Voice characteristics conversion for HMM-based speech synthesis system*. In Proc. ICASSP, 1997.

[2] R. da S. Maia, H. Zen, K. Tokuda, T. Kitamura e F. G. V. Resende Jr., *Towards the Development of a Brazilian Portuguese Text-to-Speech System Based on HMM*, In Proc. Eurospeech, 2003.

[3] R. Maia, H. Zen, K. Tokuda, T. Kitamura e F. G. Resende, *Influence of part-of-speech tagging, syllabication, and stress on HMM-based Brazilian Portuguese synthesis*, In Proc. Annual Spring Meeting of the Acoustical Society of Japan, 2004.

[4] T. Fukada, K. Tokuda, T. Kobayashi, e S. Imai, *An adaptive algorithm for mel-cepstral analysis of speech*, In Proc. ICASSP, 1992.

[5] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, e T. Kitamura, *Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis*, In Proc. EUROSPEECH, 1999.

[6] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, e T. Kitamura, *Speech synthesis generation algorithms for HMM-based speech synthesis*, In Proc. ICASSP, 2000.

[7] A. Alcaim, J. A. Solemicz, e J. A. de Morais, *Freqüência de ocorrência dos fones e listas de frases foneticamente balanceadas para o português falado no Rio de Janeiro*, Revista da Sociedade Brasileira de Telecomunicações, vol. 7, Dec. 1992.

[8] The Hidden Markov Model Toolkit. Disponível em: http://htk.eng.cam.ac.uk. Acesso em 10/10/2003.

[9] F. L. de F. Barbosa; G.O. Pinto; F. G. V. Resende Jr.; C. A. Gonç alves; R. Monserrat e M. Carlota Rosa. *Grapheme-phone transcription algorithm for a Brazilian Portuguese TTS*. In: INTERNATIONAL WORKSHOP, PROPOR, 6TH, 2003, Faro, Portugal. Computational processing of Portuguese language: proceedings. Nuno J. Mamede, Jorge Baptista, Isabel Trancoso, Maria das Graç as Volpe Nunes (Eds.) New York: Springer, 2003. p.23-30. ISBN: 3-540-40436-8. ISSN: 0302-9743.