SÍNTESE PROSÓDICA DA FALA EM PORTUGUÊS DO BRASIL

BRUNO F. DOS REIS[1], VANDER V. MARTINS[1],
MARCOS R. PEREIRA-BARRETTO[1], LUCAS A. MOSCATO

1. *Laboratório de Robôs Sociáveis, Depto de Eng. Mecatrônica, Escola Politécnica da USP*
Av. Prof. Melo Moraes 2231 – 05508-900 São Paulo-SP
E-mail para contato: marcos.barretto@poli.usp.br

Abstract— A TTS (text-to-speech) system, capable of reproduce speech with emotional nuances, with the accent of the Southeast region of Brazil, is described in this paper. Prosodic phonology has been used as the basis for system's parametric model.

Keywords— emotional speech, prosodic speech, TTS

Resumo— Um sistema TTS (*text-to-speech*), capaz de reproduzir a fala com nuances de emoção, é apresentado neste artigo. O modelo paramétrico foi baseado em fonologia prosódica e ajustado ao Português falado no sudeste do Brasil.

Palavras-chave— síntese de voz com emoções, síntese prosódica da fala, TTS.

1 Introdução

Emular a capacidade humana de contextualização de uma conversa e de reação flexível ao significado semântico é a forma mais natural e produtiva de lidar com o problema de interação homem-máquina: se o computador fosse capaz de utilizar-se dos recursos dos diversos sinais linguísticos e não linguísticos utilizados no cotidiano humano, seria possível ter-se máquinas capazes de compreender de forma mais adequada as necessidades e dificuldades do usuário, tornando mais próxima a possibilidade de uma interação genuinamente focada no indivíduo.

Dentre as dimensões que podem ser exploradas, este trabalho foca-se na adição das emoções na interação homem-máquina. Em particular, descreve um sistema capaz de reproduzir a voz humana (TTS: *text-to-speech*) adicionando-lhe conteúdo emocional. As nuances emocionais desempenham um papel importante na percepção da mensagem, de maneira que um ser humano classificaria o discurso como "não natural" ou "robótico" quando da ausência de emoções na fala [BURKHARDT, 2009].

Diversos trabalhos anteriores baseiam-se na alteração dos parâmetros sonoros diretamente por meio de tratamento de som, após a geração da fala; tais trabalhos consideram esta fala pré-sintetizada como a fala *neutra* [BULUT,2008], [DUTOIT,1993], [TAO,2006]. Outros trabalhos acoplam o modelo emocional diretamente na síntese da fala, considerando que até mesmo a fala *neutra* é indicação de um estado emocional [SCHRÖDER,2006]. Optou-se neste trabalho por seguir esta segunda vertente de pesquisadores.

Para determinar o contorno prosódico da fala, independente do estado emotivo, lançar-se-á mão da teoria da Fonologia Prosódica, tal como sugerida por Mateus [MATEUS,2004]. Ela servirá de base para a construção do modelo prosódico, o qual deve ser capaz de gerar uma fala verossímil. Seus parâmetros serão, no entanto, definidos por um modelo emocional, como explicado nas próximas seções

A camada emocional é construída sobre um sistema TTS existente, adaptada ao Português falado no sudeste do Brasil.

2 Fonologia Prosódica

Apresenta-se nesta seção uma breve revisão dos principais conceitos em Fonologia Prosódica que serão utilizados neste trabalho.

A fonologia prosódica constitui uma interface entre a gramática e a prosódia. Ela introduz os ditos *constituintes prosódicos*, entidades que surgem diretamente da gramática e da semântica da frase, mas que têm aspectos prosódicos bem definidos [MATEUS,2004]. Os constituintes prosódicos são:

a.Sílaba: constitui o elemento mínimo da hierarquia dos constituintes. Atende ao princípio da sonoridade, segundo o qual a sonoridade da sílaba aumenta do início até o seu centro, e decresce a partir de então até seu final.

b.Palavra Prosódica: assemelha-se a palavra gramatical, embora muitas vezes não sejam idênticas. Tem como característica essencial um único acento

principal e pode ter diversos acentos secundários. No Português brasileiro é comum que sílabas pares à esquerda do acento principal recebam acento secundário [FROTA,2000].

c.Sintagma Fonológico: domínio fraco no Português brasileiro. Envolve a cabeça lexical, seu especificador e seu lado recursivo, se este não for ramificado. O último acento principal é mais acentuado que os demais. Tal constituinte não será empregado no modelo prosódico apresentado neste trabalho.

d.Sintagma Entoacional: tem um contorno melódico identificável, chamado curva entoacional. É intuitivamente associado com a posição das vírgulas.

3 Modelo prosódico proposto

O modelo proposto admite que a identificação dos fonemas foi feita anteriormente, por algum outro componente. A partir dos fonemas, o modelo proposto determina as características acústicas de acordo com os seguintes procedimentos:

a) Fonema
- *Duração;*
- Um ponto de *pitch* definido, localizado a 50% de sua duração total;
- Um ponto de *intensidade* definido, localizado a 50% de sua duração total.
- O princípio da sonoridade interno à sílaba é garantido já na criação do fonema. Estes são agrupados em categorias que recebem diferentes durações. Em ordem decrescente, a escala de sonoridade utilizada é: *vogais* > *semi-vogais* > *fricativas* > *líquidas e vibrantes* > *oclusivas*. A primeira categoria tem duração padrão de 100 ms e a última de 40 ms.

b) Sílaba
- Um conjunto de *fonemas*, em determinada ordem;
- Uma *função*, que pode assumir valores de: acento principal, acento secundário, não acentuado ou após acento principal.

c) Palavra prosódica
- Conjunto de *sílabas*, em determinada ordem.

d) Sintagma Entoacional
- Conjunto de *palavras prosódicas,* em determinada ordem;
- *Curva entoacional*: neste trabalho utiliza-se apenas a curva entoacional afirmativa, mostrada na Figura 1, em frases com um único verbo, sem subordinações nem apostos.

Estes tratamentos podem ser representados pelos algoritmos a seguir:

1) Processa *palavra prosódica:*

I) Ajusta a função das *sílabas* de acordo com a posição que ocupam na *palavra prosódica*. *Sílabas* à esquerda e a um número par de distância do acento principal são consideradas acentos secundários.
II) *Fonemas* têm sua intensidade alterada, conforme a função da *sílaba* a que pertencem. Sílabas acentuadas têm intensidade maior.
III) *Fonemas* têm sua duração alterada, conforme a função da *sílaba* a que pertencem. Sílabas acentuadas têm maior duração.
IV) Define ponto de pitch p_s para cada *sílaba*, associado ao ponto t_s, na escala de tempo referenciada na *palavra prosódica*. O ponto de pitch é definido conforme a função da *sílaba* na *palavra prosódica* (sílabas acentuadas são mais agudas); em t_s, a *sílaba* está a 50% de sua execução.
V) Interpola, para cada fonema, o valor de pitch entre os pares ordenados (p_s, t_s) mais próximos.

2) Processa o *sintagma entoacional:*

I) Intensifica e alonga ligeiramente a última *sílaba* com acento principal do *sintagma*.
II) Encontra os seguintes tempos característicos, referenciados na escala de tempo do sintagma: t_1, referente à primeira ocorrência de acento principal; t_2, referente à última ocorrência de acento principal; e t_f, duração total do sintagma.
III) Determina, para cada fonema, um valor de *pitch* interpolado segundo a curva da Figura 1, que se soma ao seu *pitch* original.

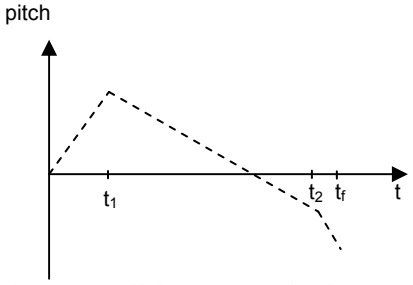

Figura 1 – perfil da curva entoacional para uma sentança afirmativa simples.

4 Modelo Emocional

O modelo emocional deve agir sobre os parâmetros do modelo prosódico, de modo a dar-lhe nuances que caracterizem emoções.

Cada estado emocional é organizado em um espaço tri-dimensional e pode ser considerada como a

combinação de três variáveis: *excitação*, *satisfação* e *dominação* [BULUT,2008], [SCHRÖDER,2006].

A *excitação* do locutor reflete o quanto ele está ativo durante o discurso. Já a *satisfação* aponta para seu estado de contentamento com a situação ou conteúdo da fala. Finalmente, a *dominação*, indica o quanto o locutor se esforça para convencer ou dominar o receptor.

Com ajuda desta representação, é possível distinguir estados emocionais complexos, como uma empolgação violenta, de uma felicidade mais amena. Enquanto ambos os estados têm alto grau de *satisfação*, diferem fortemente no eixo de *excitação.*

Neste trabalho, foram considerados apenas os estados emocionais **neutro**, **feliz**, **triste** e **bravo**. Estes estados serão caracterizados como mostrado na Tabela 1.

Tabela 1 - Estados emocionais nos eixos do modelo tri-dimensional de emoções

	Excitação	**Satisfação**	**Dominação**
Neutro	Baixo	Baixo	baixo
Feliz	Médio	Alto	baixo
Triste	baixo	muito baixo	baixo
Bravo	alto	muito baixo	alto

Os eixos do modelo tri-dimensional de emoções correlacionam-se com os parâmetros do modelo prosódico, determnando seus parâmetros: Duração das consoantes e vogais; intensidade, duração e variação de *pitch*, relacionados aos diferentes tipos de acentuação; pontos de pitch característicos da curva entoacional; freqüência média; parâmetros de qualidade da voz – voz trêmula, ar na voz.

Além disso, as correlações qualitativas dos eixos emotivos com os parâmetros acústicos propostos por Schröder [2006] são adaptados para os parâmetros disponíveis no modelo prosódico. Os parâmetros propostos em [SCHRÖDER,2006], bem como suas correlações com os eixos, não serão reproduzidos aqui, por brevidade.

Compondo os eixos e atribuindo valores para os parâmetros, a caracterização de cada emoção em relação ao estado neutro fica como descrito a seguir:

Feliz
- f0 ligeiramente elevado
- Curva entoacional bem expressiva
- Desenho melódico da palavra moderado
- Variação perceptível dos picos de intensidade nos acentos
- Variação bem moderada de duração das sílabas

Triste
- f0 ligeiramente elevado
- Curva entoacional bem pouco expressiva
- Desenho melódico da palavra bastante expressivo
- Pouca variação nos picos de intensidade dos acentos
- Fala ligeiramente lenta
- Variação audível da duração das sílabas (conforme acentos)

Bravo
- f0 menor
- Curva entoacional bastante expressiva
- Desenho melódico da palavra bastante expressivo
- Grandes picos de intensidade nos acentos
- Grande variação na duração das sílabas
- Fala ligeiramente acelerada

5 Resultados

Como TTS, foi utilizado o software MBROLA. Trata-se de um sintetizador de voz com base em difonos previamente gravados em um banco de dados. Tal software está disponível gratuitamente na Internet. No entanto, a versão disponível não permite trabalhar com variações na intensidade. Por esse motivo, foi usada uma versão estendida do MBROLA, gentilmente cedida por Piero Cosi, do *Istituto di Scienze e Tecnologie della Cognizione*, e com consentimento de Thiery Dudoit, líder e fundador do projeto MBROLA. Tal versão permite não apenas trabalhar com a intensidade, como também com parâmetros de qualidade da voz, como rouquidão, trêmulo, entre outros. O banco de dados utilizado foi o br4, ainda não disponibilizado na rede, mas gentilmente cedido pelo Serviço Federal de Processamento de Dados, em parceria com a Universidade Federal do Rio de Janeiro.

O MBROLA tem como entrada um arquivo de texto, no qual cada linha contém: um código, simbolizando um fonema; o valor de sua duração, em ms; seu contorno de *pitch*, representado por uma posição em porcentagem no tempo e um valor em freqüência em Hz; e seu contorno de intensidade, em dB.

O modelo prosódio aqui descrito,, com adição de emoções, foi implementado em JAVA e SCALA, cujas classes representam basicamente os constituintes prosódicos apresentados anteriormente.

Vários exemplos de frase foram compostos, utilizando-se os quatro estados emocionais desenvolvidos. Tais frases foram apresentadas a ouvintes. Os estados emocionais mais expressivos foram o estado *bravo* e *triste*, sendo o primeiro de identificação imediata. O estado *feliz*, dada sua sutileza

e pouca diferença do estado *neutro* – já que não se trata de *empolgação*, mas simplesmente de *felicidade* –, foi identificado com maior dificuldade.

Quando colocadas lado a lado, para uma mesma frase, todas as emoções apresentaram contraste perceptível, apontando para a almejada variabilidade típica da fala natural humana.

Ressalvas devem ser feitas, no entanto, quanto à naturalidade do discurso, já que para todas as emoções ,um "sotaque" robótico, aliado a um timbre metálico típico do sintetizador do MBROLA, comprometeram ligeiramente a percepção da fala. Refinamentos do modelo prosódico são uma possível solução para a atenuação do "sotaque".

No entanto, foi unânime entre os ouvintes que os exemplos apresentados soam mais naturais que a versão monotônica, com frequência central constante e fonemas de duração e intensidade homogêneas.

6 Conclusão

Teorias lingüísticas foram transpostas para linguagem computacional, criando um modelo prosódico coerente, capaz de gerar falas próximas à fala natural. Os parâmetros deste modelo foram determinados com base no modelo tri-dimensional de emoções, possibilitando assim que nuances emotivas sejam adicionadas ao discurso computacional.

O modelo prosódico mostrou-se eficaz para sentenças afirmativas simples, ou seja, para um único tipo de curva entoacional. Outras naturezas de curva devem ser estudadas no futuro, de maneira a compor uma biblioteca de diferentes tipos de sintagma. Estas curvas podem, eventualmente, levar em conta outros parâmetros, os quais também deverão ser correlacionados com o modelo emocional.

Agradecimentos

Piero Cosi e *Thiery Dudoit*, por ceder e autorizar o usa da versão estendida do MBROLA.

Servição Federal de Processamento de Dados e *Universidade Federal do Rio de Janeiro*, por autorizar o uso do banco de dados br4, antes de sua divulgação.

Referências Bibliográficas

BURKHARDT, F. (2009). "Emotional Speech Synthesis: applications, history and possible future". Proceedings of ESSV, Dresden, Alemanha.

BULUT, M. (2008). Recognition for Synthesis: Automatic parameter selection for resynthesis of emotional speech from neutral speech. Proceedings of ICASSP, Las Vegas, EUA.

DUTOIT, T., & LEICH, H. (1993). MBR-PSOLA: Text to Speech synthesis based on a MBE re-synthesis of the segments database. Speech Communication, vol.13 no.3-4.

TAO. J.; KANG,Y.; LI, A. (2006). Prosody Conversion From Neutral Speech. IEEE Transactions on audio, speech, and languague processing, vol. 14, no. 4 .

SCHRÖDER, M. (2006). Expressing Degree of Activation in Synthetic Speech. IEEE Transactions on audio, speech and language processing, vol. 14, no. 4, Julho.

MATEUS, M. H. (2004). Estudando a melodia da fala - traços prosódicos e constituintes prosódicos. Palavras - Revista da Associação de Professores de Português, n.º 28 , 79-98.

FROTA, S., & VIGÁRIO, M. (2000). Aspectos da Prosódia Comparada: Ritmo e entoação no PE e no PB. Actas do XV Encontro da Associação Portuguesa de Linguística, Braga, Portugal.