ESCOLA POLITÉCNICA DA UNIVERSIDADE DE SÃO PAULO

BRUNO FRANÇA DOS REIS
VANDER VALENTE MARTINS

SÍNTESE DE VOZ COM EMOÇÕES

São Paulo
2010

BRUNO FRANÇA DOS REIS
VANDER VALENTE MARTINS

SÍNTESE DE VOZ COM EMOÇÕES

Monografia apresentada à Escola Politécnica da Universidade de São Paulo para a obtenção de graduação em Engenharia

São Paulo
2010

BRUNO FRANÇA DOS REIS
VANDER VALENTE MARTINS

SÍNTESE DE VOZ COM EMOÇÕES

Monografia apresentada à Escola Politécnica da Universidade de São Paulo para a obtenção de graduação em Engenharia

Área de Concentração: Engenharia Mecatrônica

Orientador: Prof. Doutor Marcos Ribeiro Pereira Barretto

São Paulo
2010

DEDICATÓRIA

Dedicamos este trabalho aos nossos queridos familiares.

FICHA CATALOGRÁFICA

Reis, Bruno França dos
Síntese de voz com emoções / B.F. dos Reis, V.V. Martins. -- São Paulo, 2010.
p.

Trabalho de Formatura - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia Mecatrônica e de Sistemas Mecânicos.

1. Síntese de voz 2. Emoções 3. Língua portuguesa I. Martins,
Vander Valente II. Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia Mecatrônica e de Sistemas Mecânicos III. t.

AGRADECIMENTOS

Ao professor Marcos Barretto pela proposta desafiadora do tema, e toda a orientação, incentivo e liberdade que nos deu na sua abordagem.

A Piero Cosi, do Istituto di Scienze e Tecnologie della Cognizione, Itália, por nos disponibilizar sua versão personalizada do MBROLA.

A Thierry Dutoit, da Faculté Polytechnique de Mons, Bélgica, criador original do MBROLA, por nos autorizar a utilização da versão estendida do MBROLA.

Ao Serviço Federal de Processamento de Dados em parceria com a Universidade Federal do Rio de Janeiro, por nos disponibilzar de antemão seu banco de dados de fonemas, antes de sua publicação.

A todos os amigos e familiares que contribuiram informalmente com ideias e inspirações para este projeto.

Seria de fato uma invenção considerável, a de uma máquina capaz de imitar nossa fala com seus sons e articulações.

Eu creio que isso não seja impossível.

(Leonhard Euler)

Resumo

Este trabalho de formatura ocupa-se da síntese de voz computacional com nuances emocionais (emotional text-to-speech), de maneira a reproduzir a fala natural humana com fidedignidade. Este estudo focará a língua portuguesa praticada na região sudeste do Brasil. Foi realizada uma pesquisa no campo da fonologia prosódica, fundamentando a criação de um modelo computacional dos mecanismos prosódicos. Tal modelo tem seus parâmetros determinados por um modelo emocional, que dará à fala características emocionais. O modelo emocional, por sua vez, foi baseado em modelos propostos na literatura, segundo os quais emoções podem ser representadas em função de três eixos.

Abstract

This thesis proposes an emotional text-to-speech system, focusing on the production of natural-sounding phrases in Brazilian Portuguese, specifically as spoken in the South-East. For this purpose, a study on Prosodic Phonology was the basis to creating a computational model of speech mechanisms. This model has its parameters defined by an emotional model, which is in turn based on a three-dimensional model of emotions. This approach allows the synthesis of natural-sounding sentences, in which one is able to recognize emotions.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Símbolos

TTS :	Text-To-Speech
PT:	Portugal (referente ao português europeu)
BR:	Brasil (referente ao português brasileiro)
pt1:	banco de dados para o português europeu
br1, **br2**, etc:	bancos de dados para o português brasileiro
fr1:	banco de dados para a língua francesa
U:	Enunciado
I:	Sintagma Entoacional
φ:	Sintagma Fonológico
C:	Grupo Clítico
ω:	Palavra Prosódica
Σ:	Pé
σ:	Sílaba

Introdução

A interface entre computador e usuário sofreu várias alterações no decorrer das últimas décadas. De maneira a diminuir a distância que separa usuários e máquinas, vêm-se empregando técnicas que buscam humanizar o computador, ou seja, dar-lhe características que facilitem ou possibilitem seu uso por uma maior quantidade de pessoas. Um dos grandes passos dados neste sentido é a apropriação da comunicação oral e sua implementação como interface homem e máquina. Esta acontece por parte computador por mecanismos de reconhecimento e síntese de voz.

Embora sistemas de reconhecimento e síntese de voz já estejam disponíveis há algum tempo no mercado, poucos conseguem lidar de maneira satisfatória com uma das características mais marcantes da comunicação oral: as variações emotivas que dão cor ao conteúdo textual. Ou seja, há softwares capazes de identificar o conteúdo textual da fala humana e reproduzi-la a partir de um texto pré-existente, mas poucos que conseguem reconhecer e sintetizar emoções.

Este trabalho de formatura tem por objetivo o estudo e a implementação de um modelo capaz de reproduzir nuances emocionais na geração de fala computacional, de forma a aproximá-la da fala natural humana. Para tanto, faz-se necessário um estudo preliminar da prosódia, em específico da língua portuguesa. A seção 2 será, portanto, dedicada a este estudo.

A seguir, na seção 3, serão apresentadas abordagens para modelagem de estados emocionais, assim como estratégias para síntese de fala, em especial da plataforma MBROLA, a qual será a base para a síntese de fala neste trabalho.

A seção 4 apresenta sucintamente a abstração computacional do modelo prosódico apresentado na seção 2, de forma viabilizar um software apto a gerar nuances emotivas. Por fim, seguem-se considerações finais sobre o trabalho realizado nesta primeira etapa do projeto, bem como uma perspectiva de trabalho para a segunda fase.

Estudo da Prosódia

Além da complexidade sintática e semântica que a língua escrita traz, a língua falada tem ainda as características inerentes ao som, que completam a constelação de aspectos que dão significado à fala. Ao estudo da melodia que compõe a voz falada se dá o nome de Prosódia, do grego *pros*, significando *junto*, e *odos*, que quer dizer *canto*. Estas mesmas palavras em latim podem ser traduzidas como *ad* e *cantus*, dando forma assim à palavra *acento* (ANDRADE, 1841).

O estudo da prosódia vem sendo desenvolvido há séculos, e os capítulos das gramáticas dedicados ao tema frequentemente se preocupam em explicar as características melódicas da fala ao mesmo tempo em que trazem instruções sobre a boa pronúncia das palavras. Com fins de separar estes dois domínios, foi introduzido o termo Ortoépia (BARBOSA, 1821), que compreende tanto as regras de boa pronúncia como a prosódia, ficando dentro deste subdomínio apenas o estudo das características sonoras da fala.

Muitas vezes comparada às características da música, a melodia da voz falada tem como elementos principais a duração, as inflexões e o acento que "tonaliza a voz" (COELHO DE CARVALHO, 1910).

Durante o século XX o termo prosódia caiu em desuso, sendo que a sílaba, o acento e a entoação passaram ser tratados em diferentes segmentos de estudo. Apenas há alguns anos o termo vem sendo novamente empregado pelos linguistas, incidindo sobre aspectos que eram referidos pelos primeiros gramáticos(MATEUS, 2004).

Segundo o Dicionário de Termos Linguísticos(Associação de Informação Terminológica), prosódia é "o estudo da natureza e funcionamento das variações de tom, intensidade e duração na cadeia falada". Já David Crystal (CRYSTAL, 1994 apud MATEUS, 2004) adiciona a estes três elementos o ritmo - "pitch, loudness, tempo and rythm". Estes quatro elementos constituem, portanto, os traços prosódicos.

1.1 Traços Prosódicos

Traços prosódicos são características inerentes ao som.

O tom (pitch) é correlato da frequência fundamental da onda sonora. Quanto mais ciclos completos a onda executar por unidade de tempo, maior é sua altura – mais aguda, portanto. A frequência fundamental está relacionada com as cordas vocais, do ponto de

vista articulatório. Quanto mais delgadas, maior o número de vibrações, e portanto mais alto o som.

A sequência de tons de um segmento de fala define sua melodia, ou entoação.

A intensidade é dada pela amplitude da onda sonora, ou seja, a diferença entre a pressão zero e a pressão máxima da onda. Quanto maior for a energia transportada pelas partículas, maior será a amplitude e a sensação auditiva da intensidade do som. A proeminência de certas sílabas, chamada de acento, ocorre por conta de um aumento na intensidade.

A duração está relacionada com o tempo de inflexão de um som, sílaba ou enunciado. A duração de cada unidade depende da velocidade de elocução. Se a velocidade de elocução é maior, a duração de cada elemento passa a ser menor.

Tom, intensidade e duração, juntos, são responsáveis pelo ritmo da língua, aspecto este mais complexo portanto que os anteriores. O ritmo da língua portuguesa praticada na região sudeste do Brasil será posteriormente tratado.

Infelizmente as afirmações sobre estas características não resultam numa descrição rigorosa, e nem possibilitam análises comparativas inter- e intra-línguas, ou seja, não servem para a determinação de padrões prosódicos de uma língua.

Sendo assim surgiu a fonologia prosódica, uma teoria de como o fluxo da fala é organizado num número finito de unidades fonológicas. A estas unidades é dado o nome de constituintes prosódicas, as quais serão brevemente explicadas a seguir.

1.2 Constituintes Prosódicos

A fonologia prosódica é a teoria das relações de interface entre a fonologia e as outras componentes da gramática, mediada pela prosódia. Como os traços prosódicos agrupam segmentos de diversos níveis, como o nível fonológico, morfológico, sintático e semântico, Nespor e Vogel (NESPOR, 1986 apud MATEUS, 2004) propuseram em sua obra A Fonologia Prosódica a existência de constituintes prosódicos, relacionados hierarquicamente, com o objetivo de poder estabelecer padrões prosódicos das línguas e analisá-los objetivamente. Os constituintes, apresentados na sequência de hierarquia, são:

Enunciado (U)
Sintagma Entoacional (I)
Sintagma Fonológico (φ)

Grupo Clítico (C)

Palavra Prosódica (ω)

Pé (Σ)

Sílaba (σ)

Conforme Pepperkamp (PEPPERKAMP, 1996 apud FRAGOSO, 2009) o grupo clítico não é relevante como constituinte prosódico, e será neste trabalho também ignorado. MATEUS sugere também que o pé não é de fundamental importância para a língua portuguesa. Além disso, o enunciado, por ser o último nível hierárquico, ultrapassa o nível de caracterização prosódica aqui pretendido. A seguir serão apresentados os constituintes relevantes para este trabalho: a sílaba, a palavra prosódica e os sintagmas fonológico e entoacional.

1.3 Sílaba (σ)

É considerado como o constituinte prosódico fundamental, a partir de qual se formam as palavras. A sílaba é a primeira unidade linguística a ser manipulada na produção da fala (HENRIQUES, 2009). A sílaba apresenta a seguinte estrutura interna: ataque, núcleo e coda (HENRIQUES, 2009), conforme mostra Figura 0.1 sendo que apenas o núcleo está necessariamente presente:

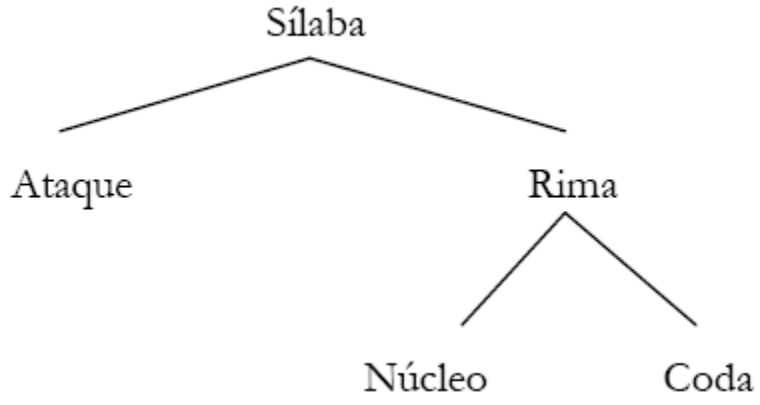


Figura 0.1 - Estrutura interna da sílaba. Fonte: (HENRIQUES, 2009)

O núcleo da sílaba tem por definição caráter vocálico, podendo ser um ditongo ou uma vogal simples. O som invocálico que antecede o núcleo é chamado ataque, e pode conter em alguns casos mais de uma consoante. Ao som invocálico que sucede o núcleo é dado o nome de coda.

Exemplo:

De-tec-tor

Analisemos o exemplo da palavra “detector”, pronunciada conforme a norma culta da língua, ou seja, de – tek – tor, e não de – te – ki – tor. Verifica-se a existência de duas sílabas completas (ataque + núcleo + coda), e uma que não contém coda.

A intensidade da sílaba varia durante sua execução, aumentando do início até seu núcleo, e decrescendo a partir de então até seu fim. Tal afirmação constitui o chamado Princípio de Sonoridade.

1.3.1 Ataque

Diferente de outras línguas, o português também possui sílabas sem ataque, como a primeira sílaba da própria palavra “ataque”. No alemão, por exemplo, todas as sílabas possuem ataque, sendo que, quando uma palavra não começa por consoante, uma parada glotal (“Knocklaut”) é executada para cumprir tal função.

O ataque pode também conter duas consoantes (ataque ramificado), como o caso de TRA – ba – lho. As letras “lh” juntas são um exemplo de dígrafo – duas letras que juntas têm um único som, como “qu”, “nh” – sendo portanto uma única consoante. Um ataque só pode ser ramificado se as duas consoantes compuserem um *grupo próprio*, compostos de uma consoante oclusiva (como [p], [b] ou [k]) e uma líquida (como [l] ou [r]). Tais ataques obedecem ao Princípio de Sonoridade, conforme mostra a Escala de Sonoridade a seguir:

Oclusivas < fricativas < nasais < líquidas (vibrantes, laterais) < semi-vogais e glides < vogais (altas, médias e baixas)

O princípio da sonoridade é violado frequentemente no português europeu, como na palavra “psicologia” (oclusiva seguida de fricativa sibilante, sonoridade decrescente antes do núcleo). No português brasileiro, tal violação é contornada com a inserção da vogal [i], \pi-si-ko-lo-jia\.

1.3.2 Núcleo

O núcleo é o único constituinte obrigatório de uma sílaba, e tem caráter exclusivamente vocálico. Dentro da sílaba, o núcleo representa o pico de intensidade, que ocorre sobre sua vogal principal. Mais de uma vogal pode estar contida no núcleo, como nos casos de ditongos e tritongos. Em casos de tritongo, a vogal principal é sempre a vogal central, como em pa-ra-GUAI. Em casos de ditongo, a vogal principal normalmente é a primeira, como em “pai”, “meu”, “tio”, e em poucos casos sobre a segunda, como em “índio”. Este segundo caso

ocorre principalmente quando da elisão de duas palavras, como em “se ajustar” \siá – jus – tar\.

No português, algumas consoantes são incorporadas às vogais, perdendo assim sua característica consonantal, como o caso de m e n, que servem para nasalizar a vogal precedente. Na palavra “amante”, por exemplo, o n da segunda sílaba não constitui coda, sendo assim parte integrante do núcleo. Neste caso, como em muitos outros, a sílaba gráfica não corresponde à sílaba fonética.

Amante \a.mã.tʃi\

Vale ressaltar que a sílaba nunca vem isolada, e a sílaba seguinte pode interferir na pronúncia da primeira, como no caso de “amanhã”, onde o dígrafo “nh” interfere na pronúncia do núcleo da sílaba anterior. Isto acontece principalmente quando a sílaba não está “protegida” por coda.

1.3.3 Coda

É o som de caráter consoante que encerra a sílaba. Tem intensidade reduzida, em comparação com o núcleo e não está necessariamente presente em todas as sílabas.

Na variante falada do português, a coda se restringe praticamente às consoantes r e s, pronunciadas das mais diferentes maneiras ao longo do país, sendo muitas vezes as principais características dos sotaques.

Não obstante, existem ocorrências de outras codas, como em ÓP-ti-co, e de-TEC-tor. Estas estão no entanto em pleno declínio, sendo dificilmente encontradas no português brasileiro falado, dando lugar a formas como “ótico” e “detetor”. Codas como “AD-vo-ga-do” são transformadas numa segunda sílaba, adicionando-se um i (em alguns casos um e), semelhante ao que ocorre com “pneu”

Tal tendência está em acordo com a forma típica da palavra portuguesa C-V-C-V (consoante, vogal, consoante, vogal). Quando da escansão, é comum a consoante final de uma sílaba abandonar sua função de coda para ser então ataque da sílaba seguinte, no processo de ressilabificação(MATEUS & RODRIGUES, 2003), como nos casos de

Os homens \o-‘zo-mẽs\

Cozinhar a comida \ko-zi-‘ŋa-ra-ko-‘mi-də\

1.4 Palavra Prosódica (ω)

É o próximo constituinte da hierarquia prosódica, quando descartado o Pé.

A palavra prosódica define-se pela existência de um acento principal, único. Assim como no caso da sílaba, a palavra prosódica não corresponde em todos os casos à palavra morfológica. As palavras prosódicas podem também possuir um ou mais acentos secundários.

Na tradição gramatical, costuma-se chamar de acento secundário o que ocorre por exemplo em diminutivos e advérbios derivados de adjetivo, como “infelizmente” e “cãozinho”, sendo colocado o acento secundário sobre o que, na palavra original, era o acento primário (infeliz, cão).

No entanto, sob o ponto de vista prosódico, acentos secundários são qualquer proeminência de sonoridade que ocorra durante o discurso, inerentes à cadeia sonora, reforçando o poder informativo do acento principal. Na fala, o acento principal e seus ecos – os acentos secundários – representam a alternância de batimentos fortes e fracos que caracterizam o ritmo.

De acordo com este ponto de vista, considera-se a palavra morfológica “infelizmente” como duas palavras prosódicas, pelo fato de ela possuir dois diferentes acentos principais. Acentos principais serão marcados com (‘) e secundários com (,).

$$[,\text{infe'liz}]_{\omega}\ [\text{'mente}]_{\omega}$$

Para o mesmo exemplo, pode-se observar a ocorrência de um acento secundário sobre a sílaba “in”, que é reforçada, mas não tão proeminente o acento principal. Acentos secundários ocorrem sempre em sílabas pré-tônicas

Estudos comparativos do português brasileiro e do europeu apontam para diferentes tendências no ritmo da fala, gerando diferentes posicionamentos do acento secundário. Enquanto no português europeu o acento secundário tende a ser posto sobre a primeira sílaba da palavra prosódica, no português brasileiro acentuam-se as sílabas pares a esquerda do acento principal (FROTA & VIGÁRIO, 2000). Analisemos as duas variantes rítmicas da palavra “temperatura”

,tem pe ra ‘tu ra (PE)

tem ,pe ra ‘tu ra (PB)

Vale lembrar que os acentos secundários são influenciados por fatores aleatórios, como por exemplo, estados emocionais.

Além de uma palavra morfológica poder ser representada por duas palavras prosódicas, o contrário também se verifica. É o caso de palavras que têm monossílabos átonos adjacentes, como artigos ou pronomes oblíquos átonos.

$$[\text{o homem}]_{\omega}$$

$$[\text{colocou-o}]_{\omega}$$

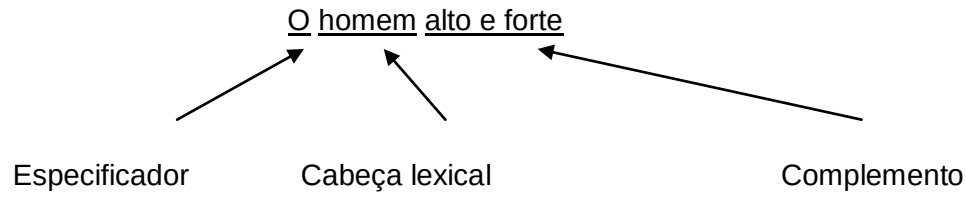
Desta maneira as palavras prosódicas não se restringem a oxítonas, paroxítonas e proparoxítonas, podendo o acento principal recair na quarta sílaba da direita para esquerda (embora estes casos sejam raros e de difícil pronúncia).

$$[\text{Colocávamo-nos}]_{\omega}$$

1.5 Sintagma Fonológico (φ)

Conforme sugere MATEUS, o sintagma fonológico é um domínio fraco na língua portuguesa, pois sua identificação não é evidente.

O sintagma fonológico se dá pelo agrupamento de palavras prosódicas. Tal constituinte baseia-se em conceitos gerais como cabeça lexical do sintagma sintático, a projeção máxima dessa cabeça lexical, e o seu lado recursivo. A cabeça lexical é a categoria lexical que pode ter complementos e um especificador. A projeção máxima da cabeça lexical é a própria cabeça, juntamente com seus complementos e especificador. O lado recursivo é o lado em que se encontram os complementos da cabeça lexical. No português, o lado recursivo é o direito. Vejamos o exemplo:



Adjetivos e advérbios só podem ser cabeça lexical se estes vierem à esquerda – ou seja, no lado não recursivo – de uma cabeça lexical, e forem dominados pela projeção máxima dessa cabeça lexical.

Para a formação do sintagma fonológico, incluem-se a cabeça lexical e seus complementos, desde que estes não sejam ramificados.

$$[\text{A casa velha}]_{\varphi}$$

$$[\text{A casa}]_{\varphi}\ [\text{muito velha}]_{\varphi}$$

FROTA (apud MATEUS, 2004) sugere a existência de acentos tonais, na palavra proeminente do sintagma fonológico. No português, a palavra proeminente é aquela mais à direita do sintagma fonológico. Vejamos o exemplo a seguir:

$$[\text{O jornaLISta}]_{\varphi}\ [\text{FEZ}]_{\varphi}\ [\text{uma entrevista interesSANte}]_{\varphi}$$

$$[\text{O jornaLISta}]_{\varphi}\ [\text{fez uma entreVISta}]_{\varphi}\ [\text{muito interesSANte}]_{\varphi}$$

1.6 Sintagma Entoacional (I)

O sintagma entoacional constitui-se de um ou mais sintagmas fonológicos, e caracteriza-se por um contorno identificável. Vejamos o exemplo a seguir:

As casas, assim por dizer, queimaram como papel.

Percebemos a existência de três contornos distintos de entoação. A coda de casas, neste caso, não se transforma em [z], e permanece como [s]. Este é um típico fenômeno que acontece na fronteira de um sintagma entoacional.

$$[\text{As casas}]_{I}\ [\text{assim por dizer}]_{I}\ [\text{queimaram como papel}]_{I}$$

Sintagmas entoacionais podem variar de acordo com o locutor. Frases longas tendem a ser quebradas em mais de um sintagma, quando da fala pausada. A quebra de sintagma entoacional vem muitas vezes associada à colocação de uma vírgula. Isso leva muitas vezes a colocação de vírgula em lugares inapropriados, como entre sujeito e predicado.

Do ponto de vista acústico, o sintagma entoacional possui um contorno de frequência tipicamente contínuo. No português, tal contorno frequentemente é o único recurso de que dispomos para diferenciar uma interrogação de uma afirmação.

1.7 Aspectos Emocionais

A fala é o principal meio de comunicação entre os humanos. Não obstante, ela tem sido estudada desde datas remotas. Com o surgimento e o desenvolvimento do computador, os meios de comunicação entre homem e máquina vêm se assemelhando cada vez mais com os meios existentes entre os homens. Um dos grandes passos dados nesta direção é o estudo da comunicação oral como possibilidade de interação homem-máquina.

No entanto, uma característica inerente ao mecanismo de fala torna esta interação, tão natural entre seres humanos, extremamente dificultosa para o computador: a variabilidade. Quando se dá a duas pessoas uma informação a ser transmitida, elas a transmitirão de forma diferente uma da outra, tanto em nível verbal quanto em nível vocal. Até mesmo diversos exemplos de uma mesma palavra não serão acusticamente idênticos. Isso se dá por conta de um grande número de razões, entre elas:

- **Estilo**: o estilo da fala age diretamente no vocabulário empregado. Qualidades acústicas, no entanto, não são impassíveis de alterações de estilo. Um mesmo texto pode ser pronunciado de maneira mais polida ou mais ríspida, por exemplo. O estilo da fala está intimamente ligado com o ambiente social do locutor. Ao mesmo tempo, o estilo leva em consideração também o receptor (o tom de voz que se emprega com crianças, com adultos e com idosos por exemplo). Trata-se portanto de um campo recheado de variáveis, de estudo bastante complexo.
- **Humor e emoção**: à parte o estilo empregado na fala, o estado emocional do locutor sujeita também a fala. Como será discutido mais adiante, os fatores acústicos da fala são diretamente influenciados por fatores emocionais. A título de distinção, entende-se na maioria das vezes por humor um estado emocional constante, enquanto por emoção entende-se uma qualidade emotiva pontual. Por exemplo, o locutor pode estar feliz – humor – e mesmo assim pronunciar uma frase com medo – emoção.
- **Estresse**: As condições físicas em que se encontra o locutor também têm ação direta sobre as características acústicas do som. A parte o estilo e a emoção, a fala também será afetada por estados de saúde, como rouquidão ou ação de remédios, por estado de embriaguez, entre outros. Fatores físicos, como vibrações e impactos, também alteram a qualidade da voz, por agirem fisicamente direto sobre o aparelho responsável pela fala.

A fala humana natural estará sempre sujeita a todos estes fatores de variabilidade. Mesmo a fala humana dita neutra não está livre de variabilidade. A fala computacional, portanto, muitas vezes erroneamente chamada de fala neutra, não leva em conta tal variabilidade, o que, ao invés de neutra, a torna uma voz estranha e maquinal.

A seguir, apresentaremos uma abordagem do ponto de vista computacional sobre as estratégias existentes para a síntese de voz que, embora não cubram todo o leque de variabilidade, preocupam-se em dar à fala características emotivas, aproximando-a portanto da fala humana natural.

1.7.1 Parâmetros acústicos

A síntese de voz carregando traços de emoções depende de uma difícil escolha de parâmetros acústicos, a serem habilmente manipulados, que influenciam na forma como um ser humano aprecia a fala.

Essa determinação de parâmetros de caráter exclusivamente acústico é bastante difícil, já que em um estudo com falantes humanos há muitos parâmetros que vão além da linguística, e que são muito difíceis de controlar, tais como idade ou estado de saúde do locutor (SCHRÖDER, 2006).

Há muito se preocupa com esse tema, tão essencial na síntese emocional de voz. Na literatura, encontra-se um conjunto de parâmetros bastante recorrentes, e que os estudos que os apresentam os escolheram a partir de métodos experimentais, em geral com seres humanos, que evidenciam sua importância(BULUT, 2008). Esses parâmetros principais se resumem aos traços prosódicos: **frequência fundamental (f0)**, **ritmo**, **intensidade** e **qualidade (timbre) da voz**.

Os aspectos da frequência fundamental de maior importância são:

- o seu valor médio;
- o seu intervalo;
- a média e o intervalo das magnitudes de suas variações;
- a média e o intervalo da duração de suas variações; e
- a rapidez e a frequência de suas variações.

Quanto ao ritmo, destacam-se:

- a quantidade de pausas e suas durações; e
- a quantidade de picos de intensidade.

Relativamente à intensidade, os parâmetros mais importantes são:

- seus valores médios; e
- seus intervalos.

Finalmente, a qualidade da voz é definida por:

- esforço do falante; e
- a quantidade de ar presente na fala.

1.7.2 Modelos emocionais

Agrupar em categorias, abstrair matematicamente, encontrar relações de subordinação. Estas não são tarefas triviais quando os objetos a serem estudados são estados emocionais. Além de emoções serem subjetivas por natureza, a nomenclatura utilizada no dia-a-dia não é suficiente para cobrir a complexa gama de combinações e sutilezas emocionais. A definição de felicidade, por exemplo, não é evidente.

Alguns modelos vêm sendo propostos para simplificar os estados emocionais em “emoções principais”, de modo a reduzir drasticamente a quantidade de estados emotivos aos quais os homens estão sujeitos. Apesar de um modelo como este ficar extremamente aquém da realidade psicológica humana, ele já introduz no sistema de síntese de voz uma série de nuances emotivas, aproximando a fala computacional da fala humana.

Não obstante, a escolha das “emoções principais” não é consenso entre os pesquisadores. Bulut, por exemplo, vale-se de emoções como **neutro**, **felicidade**, **tristeza** e **raiva** como emoções principais. Já Tao acrescenta a este rol de emoções principais o **medo**, considerando o estado neutro como o estado inicial do discurso, o qual passará por uma ressíntese (TAO, 2006).

Tais modelos são genericamente chamados de “modelos de categorias” (BURKHARDT, 2006). Neste caso, os estados emocionais traduzem-se diretamente em parâmetros acústicos. Na literatura encontram-se algumas sugestões de relação entre os estados emocionais e as características acústicas. Vale lembrar, no entanto, que tais estudos foram feitos para línguas diversas, e muitas vezes não entram em consenso sobre determinados parâmetros. Nenhum estudo para o português foi encontrado.

Alguns modelos menos subjetivos introduzem outro nível de informação entre o estado emocional e seus parâmetros acústicos. Bulut, por exemplo, utiliza-se de um modelo bidimensional, onde as quatro emoções principais (neutro, felicidade, tristeza e raiva), se distribuem entre os quatro quadrantes do plano cartesiano, de forma que um estado

emocional possa se encontrar em um quadrante, mas na iminência de pertencer ao outro (BULUT, 2008). O significado direto dos eixos do plano cartesiano não fica explícito, e serve neste caso apenas para o treinamento de uma rede neural artificial.

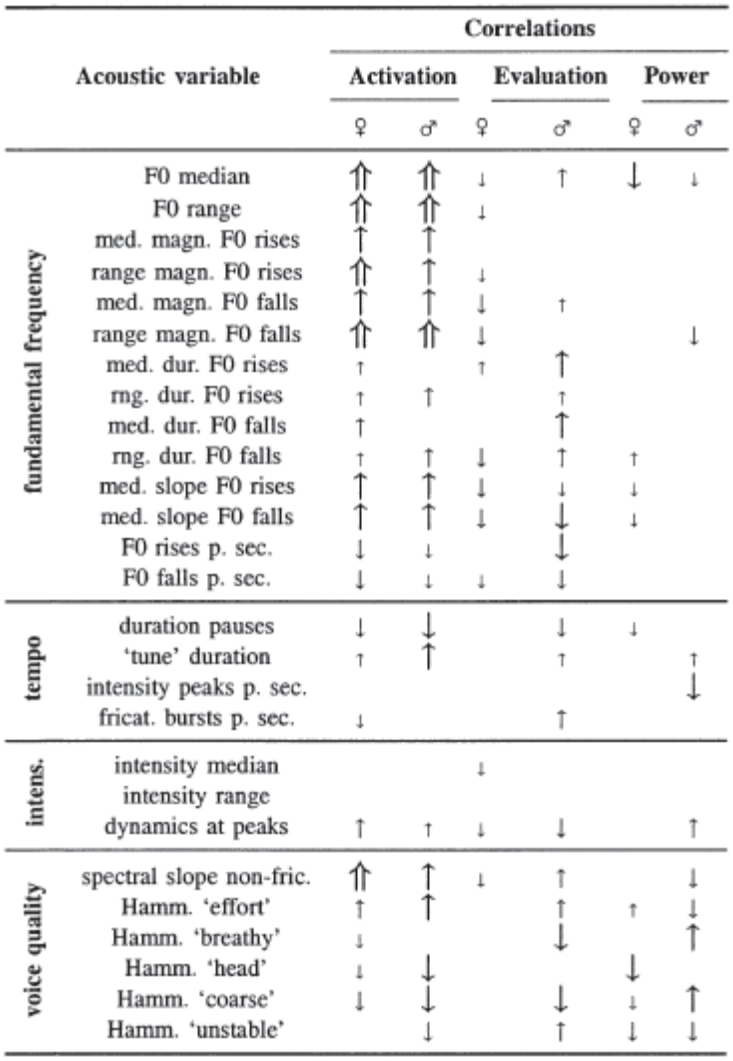
Por outro lado, Schröder utiliza-se de um espaço tridimensional, onde cada eixo tem uma significação bastante própria (SCHRÖDER, 2006). Estes são, em inglês: *activation*, *evaluation* e *power*. Embora o mesmo modelo seja amplamente utilizado por outros estudiosos, tal nomenclatura varia fortemente de autor para autor. Neste trabalho, utilizaremos uma tradução livre, com base nestes três termos apresentados Schröder, comparados com termos utilizados por outros autores:

- Excitação
- Contentamento
- Dominação

Os estados emocionais passam portanto a estar distribuídos neste espaço tridimensional. Pode-se associar, assim, um estado **entediado** a baixos graus de excitação, contentamento e dominação. Um estado raivoso, por outro lado, terá altos graus de dominação e excitação, porém pouco contentamento etc.

Tal modelo permite distinguir estados emocionais como **alegria eufórica** de uma **felicidade amena e constante**. Ambas os estados, que seriam agrupados sob a emoção principal **felicidade**, passam a ser estados distintos. Desta maneira, não faz mais sentido tentar agrupar e localizar os diferentes estados dentro deste espaço tridimensional. Em seu software de síntese emocional de voz, o OpenMary, Schröder deixa a critério do usuário a livre utilização dos três eixos.

Os eixos, por seu lado, passam a se correlacionar com os parâmetros acústicos de maneira difusa. Abaixo, vemos as relações entre parâmetros acústicos, e os respectivos graus de excitação, contentamento e dominação. Observemos que tal correlação não se dá com valores exatos. Pelo contrário, a relação entre parâmetros acústicos e os eixos se dá por regras baseadas em valores linguísticos (como “pouco”, “muito”, “muito pouco”, etc.), os quais apontam para a necessidade do uso de lógica difusa no processamento destas regras.



	Acoustic variable	Correlations					
		Activation		Evaluation		Power	
		♀	♂	♀	♂	♀	♂
fundamental frequency	F0 median	⇑	⇑	↓	↑	↓	↓
	F0 range	⇑	⇑	↓			
	med. magn. F0 rises	↑	↑				
	range magn. F0 rises	⇑	↑	↓			
	med. magn. F0 falls	↑	↑	↓	↑		
	range magn. F0 falls	⇑	⇑	↓			↓
	med. dur. F0 rises	↑		↑	↑		
	rng. dur. F0 rises	↑	↑		↑		
	med. dur. F0 falls	↑			↑		
	rng. dur. F0 falls	↑	↑	↓	↑	↑	
	med. slope F0 rises	↑	↑	↓	↓	↓	
	med. slope F0 falls	↑	↑	↓	↓	↓	
	F0 rises p. sec.	↓	↓		↓		
	F0 falls p. sec.	↓	↓	↓	↓		
tempo	duration pauses	↓	↓		↓	↓	
	'tune' duration	↑	↑		↑		↑
	intensity peaks p. sec.						↓
	fricat. bursts p. sec.	↓			↑		
intens.	intensity median			↓			
	intensity range						
	dynamics at peaks	↑	↑	↓	↓		↑
voice quality	spectral slope non-fric.	⇑	↑	↓	↑		↓
	Hamm. 'effort'	↑	↑		↑	↑	↓
	Hamm. 'breathy'	↓			↓		↑
	Hamm. 'head'	↓	↓			↓	
	Hamm. 'coarse'	↓	↓		↓	↓	↑
	Hamm. 'unstable'		↓		↑	↓	↓

Figura 0.1 – Correlação das variáveis acústicas com os eixos emotivos *Activation* (excitação), *Evalutaion* (contentamento) e *Power* (dominação), tanto para voz masculina como para a femina. Estas regras são apenas qualitativas. Fonte:(SCHRÖDER, 2006)

1.8 Estratégias de síntese

Sistemas de síntese de fala podem ser classificados em três grupos distintos(LEMMETTY, 1999), os **sistemas articulatórios**, os sistemas de **síntese por componentes** e os sistemas de **síntese por concatenação**.

1.8.1 Sistemas articulatórios

Os sistemas articulatórios tentam modelar diretamente o aparelho fonético humano, o que é uma tarefa extremamente complicada. Deve-se modelar a movimentação de massas e a deformação de corpos (língua, pulmão), propriedades vibratórias das cordas vocais, dos ossos, etc. Tal modelagem, de altíssima complexidade, embora possa produzir um resultado de alta qualidade, exige estudos e processamento computacional muito intensos.

1.8.2 Sistemas de síntese por componentes

Os sistemas de síntese por componentes são bastante flexíveis pois envolvem a geração de som exclusivamente a partir da manipulação de componentes acústicos. Esses sistemas utilizam um conjunto de regras que determinam como certos parâmetros (tais como a frequência natural, sua amplitude, os harmônicos e suas amplitudes, etc.) devem ser ajustados para se produzir um determinado som. Em geral, esses sistemas requerem muito menos processamento que um sistema articulatório, e não necessitam de uma grande quantidade de dados para funcionarem.

Este sistema, que já foi bastante usado anteriormente, tem dado lugar aos sistemas de síntese por concatenação(LEMMETTY, 1999), descritos a seguir.

1.8.3 Sistemas de síntese por concatenação

Finalmente, os sistemas de síntese por concatenação são sistemas que requerem um banco de dados de unidades, em geral gravadas por um ser humano, que são concatenadas para se obter o som desejado.

Deve-se determinar o tamanho da unidade a ser concatenada para a criação de um banco de dados de unidades. Quanto maiores forem as unidades, maior será o grau de naturalidade da fala e mais memória será usada para o armazenamento. As unidades usadas hoje em dia são geralmente palavras, sílabas, fonemas ou dífonos. Há também sistemas que envolvem frases, que são bastante simples mas de uso severamente limitado – sistemas de atendimento telefônico automático –, pois se restringem às frases previamente gravadas.

Os bancos de dados de palavras ou sílabas não são convenientes para sistemas de síntese de fala genéricos, pois há a necessidade de uma quantidade gigantesca – da ordem de centenas de milhares para palavras, e da ordem de dezenas de milhares para sílabas – de unidades para se poder transformar em som qualquer frase válida na língua do sistema.

As unidades mais comumente usadas são os fonemas, ou para maior qualidade do resultado, os dífonos. Uma língua tem da ordem de algumas dezenas de fonemas, o que torna o banco de dados bastante conciso. Entretanto, a concatenação de fonemas pode apresentar problemas como distorções nos pontos de concatenação, assim como baixa naturalidade, pois a sonoridade de um fonema depende muito frequentemente dos fonemas ao seu redor. Para resolver tal problema, é comum o uso dos dífonos, que são mais numerosos que os fonemas, mas ainda sim suficientemente restritos para que seja possível a criação de um banco de dados de tamanho razoável.

Um algoritmo para a concatenação de unidades sonoras, capaz de modificar tom e duração de forma independente e eficaz (pois trabalha do domínio do tempo), é conhecido como TD-PSOLA(LEMMETTY, 1999). Uma evolução deste algoritmo, cujo objetivo é a síntese de fala com maior qualidade, dependente de um banco de dados de difonos com certos requisitos de qualidade discutidos mais adiante neste trabalho, é hoje conhecido por MBROLA(DUTOIT & LEICH, 1993), e deu origem ao Projeto MBROLA.

1.9 Projeto MBROLA

O Projeto MBROLA tem por objetivo a disponibilização gratuita de um sintetizador de fala de alta qualidade para a maior quantidade de línguas possível.

O sistema de base é composto por um executável, o MBROLA, disponível para aproximadamente 30 plataformas (dentre elas Windows, Linux e Mac) e por um conjunto de bancos de dados de difonos que contempla, atualmente, mais de 30 línguas.

Esse software não é um sistema completo de síntese de voz, pois ele não é capaz de tratar uma entrada textual arbitrária. Os formatos de entrada e saída serão descritos em 1.9.3.

1.9.1 Bancos de dados

Os bancos de dados de difonos do MBROLA são compostos cada um por um arquivo único, cujo tamanho é da ordem de no máximo dezenas de megabytes.

Um banco de dados é composto por um conjunto Φ de fonemas e uma associação $s: D \subset \Phi^2 \rightarrow \Omega$ entre um subconjunto D dos difonos possíveis de se formar a partir dos fonemas, e as formas de onda Ω que representam o som do difono.

Para se produzir um banco de dados, é necessário gravar uma quantidade de texto que contenha todos os difonos desejados, para em seguida realizar, em geral de forma manual, uma quebra do som em partes pequenas associadas cada uma a um único fonema. Finalmente, um difono é a forma de onda presente entre os centros das formas de onda do primeiro e do segundo fonema.

Para garantir a qualidade do resultado produzido pelo MBROLA, exige-se que cada forma de onda do banco de dados tenha o mesmo tom constante, isto é, a mesma frequência fundamental, e a mesma fase.

Os bancos de dados são nomeados segundo o padrão de duas letras maiúsculas, que simbolizam a língua representada, seguidas de um número que identifica a ordem de

produção do banco de dados. Assim, o primeiro banco de dados para a língua francesa se chama FR1, o segundo FR2, e assim por diante.

1.9.2 Língua portuguesa

O projeto MBROLA inclui diversos bancos de dados para a língua portuguesa. Deve-se fazer uma importante distinção entre os bancos de dados para a língua portuguesa tal qual praticada na Europa e no Brasil. A primeira usa o símbolo PT, enquanto que a segunda usa o símbolo BR.

Hoje, no website do Projeto MBROLA, encontram-se um banco de dados para o português europeu, PT1, e três bancos de dados para o português brasileiro, BR1, BR2 e BR3. Estes três bancos de dados foram produzidos pelo ex-estudante de Ciência da Computação na Universidade de São Paulo e atual diretor de tecnologia da empresa MicroPower, Denis R. Costa.

Os três bancos de dados do português brasileiro citados no parágrafo anterior contêm exatamente a mesma voz masculina, na frequência fundamental de 110 Hz, porém cada um com uma qualidade ligeiramente superior à versão anterior. Ainda assim, os resultados que se obtém com o uso destes bancos de dados é de qualidade bastante inferior aos obtidos, por exemplo, com o banco de dados FR1, no quesito compreensibilidade da fala produzida.

Encontrou-se, então, um quarto banco de dados para o português brasileiro, ainda não disponível no site do Projeto MBROLA, produzido pelo Serviço Federal de Processamento de Dados (SERPRO) em parceria com a Universidade Federal do Rio de Janeiro, sob coordenação do analista de sistemas José Antônio Borges. Trata-se de um banco de dados com uma voz feminina, na frequência fundamental de 230 Hz. A locutora original das palavras é a Liane Paixão Borges, que batizou o banco de dados, **Liane TTS**.

O BR4 apresenta uma série de vantagens em relação as versões anteriores. A qualidade do banco de dados é sensivelmente superior, além de incluir difonos que o primeiro não incluía, como o *d* final da palavra “saudade”, e o *t* da palavra “tia”. A voz feminina também apresenta uma vantagem significativa em relação à voz masculina por ser mais clara e, portanto, de mais fácil compreensão. Portanto, este trabalho basear-se-á exclusivamente no banco de dados BR4.

1.9.3 Formatos de entrada e saída

O MBROLA recebe como entrada um arquivo textual do tipo .PHO, processado linha por linha, cujo formato é descrito a seguir.

Um **comentário** começa com um ponto-e-vírgula e se estende até o final da linha. Linhas começando com dois pontos-e-vírgulas podem representar **definições de parâmetros de inicialização** que são aplicados ao processo de síntese do som de forma global, tais como a multiplicação de todos os tons por uma constante, ou a intensidade sonora do arquivo de som resultante como um todo.

Uma **linha vazia** é ignorada pelo sistema.

Finalmente, uma linha não vazia e que não comece com um comentário, é uma **linha de definição de um fonema**. Tal linha é composta por duas partes obrigatórias e uma parte opcional, todas separadas por um espaço. A primeira parte, obrigatória, é uma sequência de caracteres que define o fonema. A segunda parte, também obrigatória, é um número que define a duração, em milissegundos, daquele fonema. Finalmente, a terceira parte, opcional, define uma curva de evolução do tom no tempo, através de uma sequência de pares de números, todos separados por espaços: o primeiro dos números representa uma porcentagem do intervalo de duração do fonema, e o segundo representa o tom, em Hz, daquele instante. A curva é definida por segmentos de retas que ligam os pontos assim definidos.

Um arquivo de entrada é válido quando estiver conforme o formato descrito acima.

Para que o software consiga produzir o arquivo de som de saída, é necessário ainda que todos os pares de fonemas na sequência descrita no arquivo .PHO de entrada formem dífonos contidos no banco de dados utilizado.

Satisfeitas essas condições, o programa será capaz de produzir um arquivo de som, em um dos formatos WAV, AU, AIFF ou RAW.

1.9.4 Limitações do MBROLA

O software disponibilizado pelo Projeto MBROLA apresenta algumas limitações quanto a seu uso, sem tratamento adicional, para a síntese de voz natural para o português, em especial para a voz com emoções.

O português, como vimos anteriormente, é uma de prosódia bastante rica, que, diferentemente do francês, contém sílabas tônicas. Estas são sílabas que, acusticamente, têm maior duração, maior altura, mas, principalmente, maior intensidade. O português requer, então, a possibilidade de se controlar a evolução da intensidade do som no tempo, o que o MBROLA original não possibilita.

A adição de emoções à fala é também fortemente dependente da possibilidade de controle da qualidade da voz, conforme explicado anteriormente.

Frente a essas dificuldades, um grupo italiano de pesquisa de síntese de voz com emoções, do Istituto di Scienze e Tecnologie della Cognizione, sob coordenação de Piero COSI, realizou uma modificação do software MBROLA, de forma a torná-lo muito mais poderoso e flexível, oferecendo um grande leque de parâmetros que se pode controlar.

Essa versão estendida do MBROLA, produzida com o acordo do autor da versão original do software Thiery Dudoit, foi disponibilizada para a realização deste trabalho de formatura, sujeita às mesmas condições de uso da versão original e à sua não distribuição.

1.9.5 MBROLA estendido

Para se poder controlar os novos parâmetros disponibilizados pelo software estendido, o formato do arquivo .PHO de entrada foi alterado.

Nas linhas de definição de fonema, além das partes obrigatórias, que definem o fonema e sua duração, e da parte opcional que define o contorno de f0, pode-se encontrar um grande número de outras partes opcionais. Estas partes devem ser especificadas numa determinada ordem. Isto possibilita o controle dos parâmetros adicionais da versão estendida do MBROLA.

A definição se dá por uma palavra correspondente ao parâmetro a ser modificado, seguida de uma sequência de pares de números, descrevendo a evolução do parâmetro no intervalo de duração do fonema, da mesma forma como se faz a variação do tom.

Panorama do Projeto

Um sistema completo de síntese de voz a partir de texto (TTS, *text-to-speech*) é aquele capaz de produzir som a partir de uma entrada exclusivamente textual. Um tal sistema envolve no mínimo duas fases bastante distintas, a primeira chamada de processamento de linguagem natural e a segunda, processamento digital de som(BURKHARDT, 2006).

A fase de **processamento de linguagem natural (PLN)** é responsável por transformar uma entrada textual arbitrária numa representação abstrata baseada nos conceitos de domínio da prosódia, estudados anteriormente. Trata-se de uma representação intermediária entre os domínios textual e sonoro, e que contém aspectos morfológicos e prosódicos.

A fase seguinte é o **processamento digital de som (PDS)**, que é responsável por transformar a representação prosódica do texto, obtida da primeira fase, no som que corresponde à entrada textual original. As técnicas usadas nesta etapa são aquelas descritas na seção 1.7.2.

A síntese emocional de voz adiciona uma fase de processamento intermediária, que modifica parâmetros da representação prosódica do texto conforme a emoção que se deseja obter na saída sonora do sistema. Além disso, a síntese emocional de voz impõe a necessidade de um maior controle na fase final.

O diagrama a seguir esquematiza um sistema TTS completo capaz de tratar emoções, e outro sem essa possibilidade.

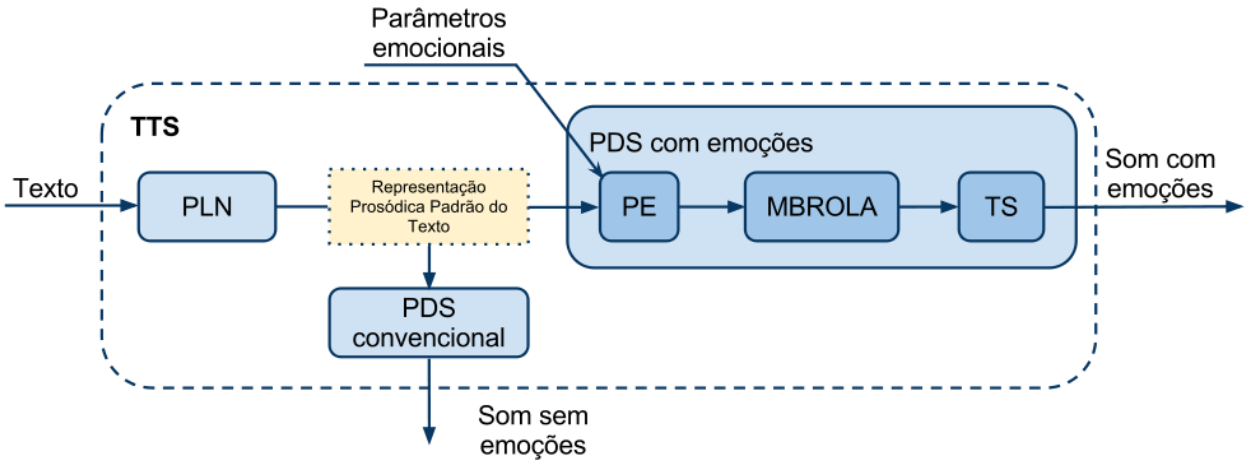


Figura 0.1 Esquema geral de um TTS completo, com ou sem emoções.

Neste diagrama, dá-se o nome **PDS convencional** a um PDS incapaz de trabalhar com emoções, que simplesmente toma o produto da fase PLN e o transforma em som, valendo-se de uma das estratégias descritas na seção 1.8, e que pode muito bem ser implementado

com base no MBROLA, por exemplo. O bloco **PE** corresponde a um processamento de emoções, que recebe como entrada, além do produto da fase PLN, um conjunto de parâmetros que definem os traços emocionais que se quer dar ao som produzido. O bloco **TS** representa um eventual pós-tratamento de som para aumento de sua qualidade.

1.10 Linguagem de programação

Para se construir um modelo computacional qualquer, uma vez definido de forma clara o problema, deve-se escolher inicialmente um paradigma de programação adequado, e em seguida uma linguagem de programação que será usada no projeto.

Para o problema foco deste trabalho, a representação prosódica de um texto para síntese emocional de voz, dois paradigmas são inicialmente considerados: programação funcional e programação orientada a objetos.

A programação funcional tem como principais pontos fortes a simplicidade de representação de algoritmos complexos e de estruturas recursivas, e que utiliza em geral funções puras, sem estado, e no qual se evita o uso de mutabilidade de variáveis. É um paradigma que combina bastante bem com a primeira fase de um sistema de síntese de voz, o processamento de linguagem natural, graças às suas ferramentas poderosas para a descrição de regras de processamento, como o casamento de padrões (*pattern matching*), e o código conciso, bastante expressivo, que se pode obter com muitas linguagens deste paradigma. Como exemplos tradicionais de linguagens funcionais têm-se Haskell ou Lisp.

A programação orientada a objetos, por outro lado, é muito bem adaptada à representação de estruturas composicionais, de natureza hierárquica. A mutabilidade e a manutenção de estado nas operações fazem parte da natureza do paradigma. Essencialmente, o que se faz é utilizar o conceito de *objeto* para representar uma estrutura de dados e seu estado juntamente com as operações que se pode realizar nesta estrutura. A vantagem principal da representação por objetos reside na facilidade de controle e redução da complexidade de um sistema. Mais conhecido, este paradigma tem como exemplos de linguagem C++, Java ou C#.

Hoje em dia, o que se observa é uma convergência entre estes dois paradigmas, com linguagens como OCaml, F# ou Scala sendo grandes representantes deste movimento. Busca-se combinar as vantagens das linguagens funcionais para a criação de algoritmos adaptados à computação paralela com a facilidade do controle da complexidade oferecida pelo paradigma da orientação a objetos.

1.10.1 A máquina virtual Java

Programas escritos em Java tem seu código fonte compilado no que se chama *bytecode*, uma representação intermediária entre aquela da linguagem Java e aquela entendida por microprocessadores. Torna-se necessário então o uso de uma máquina virtual (JVM), que é responsável por transformar o *bytecode* em código de máquina.

Hoje em dia, as máquinas virtuais são extremamente eficientes, e possibilitam certos tipos de otimização em tempo de execução não possíveis antes.

Finalmente, uma grande vantagem do uso da linguagem Java é a interoperabilidade com outras linguagens, que também são compiladas para o mesmo *bytecode* definido pela JVM. Temos como exemplo a linguagem Scala, que pode usar diretamente as estruturas definidas no Java, e vice-versa. Assim, pode-se construir um módulo do programa no paradigma funcional, caso este se mostre mais vantajoso.

1.10.2 Linguagem Scala

Para o desenvolvimento do projeto, foi escolhida a linguagem de programação Scala, na sua versão 2.8.1, de 9 de novembro de 2010.

Trata-se de uma linguagem moderna, multi-paradigma, que propõe uma forma de se aproveitar ao máximo as vantagens dos paradigmas da orientação a objetos e da programação funcional. O código gerado executa dentro de uma JVM (Java Virtual Machine), e é totalmente compatível com Java (isto é, pode-se usar classes Java em Scala e vice-versa), o que promove a extensibilidade e o reuso do código.

Graças a um sistema de tipos bastante sofisticado e poderoso, tendo funções como objetos de mesma categoria que outros tipos de dados, a biblioteca padrão da linguagem inclui uma grande quantidade de classes de coleções (listas, conjuntos, sequências, etc). Estas coleções possuem métodos (funções) de ordem superior, tais como map, foldLeft, reduce, filter, entre muitos outros, comuns em linguagens funcionais, que simplificam tremendamente a escrita de algoritmos.

Dessa forma, pode-se trabalhar ao mesmo tempo com o paradigma da orientação a objetos, que é uma forma natural de definir os tipos de dados usados no projeto, e com o paradigma da programação funcional, que simplifica bastante a escrita dos algoritmos descritos adiante, de forma robusta e mais produtiva.

1.11 Processamento digital do som

A etapa de processamento digital do som será feita com o uso da versão estendida do MBROLA, e com o banco de dados BR4. A execução do MBROLA, e todo pré- e pós-tratamento necessário para a obtenção da fala, a partir da representação na estrutura definida na seção anterior, são encapsulados dentro do software em Java criado para acompanhar este trabalho de conclusão de curso.

1.12 Confecção do Protótipo

Com o objetivo de testar a capacidade do modelo desenvolvido e da plataforma Mbrola, confeccionou-se um pequeno protótipo. Este foi desenvolvido de maneira manual, ou seja, os parâmetros do MBROLA foram escolhidos um a um. Outrossim, a escolha dos parâmetros foi guiada exclusivamente pelos princípios prosódicos apresentados na seção 0. Os dados obtidos da literatura sobre modelos emocionais, em especial a tabela da Figura 0.1, serviu de base para criação das nuances emocionais.

Ao final, obtiveram-se duas amostras, ambas com características de estados emocionais bem marcadas.

Ambos os protótipos baseaim-se sobre o mesmo texto, do poema de Manuel Bandeira

> “Eu quero a estrela da manhã”

Tal frase foi escolhida pelos seguintes motivos:

- Curta o bastante, reduz o trabalho árduo de ajustar os parâmetros de cada fonema.
- Não carrega consigo, a princípio, nenhum conteúdo emotivo específico.
- O caráter poético da frase reforça a nuance emocional que lhe for dada.

O primeiro passo é a identificação dos fonemas da base de dados BR4 na frase. Logo em seguida, durações idênticas foram atribuídas a cada fonema, no valor de 100 ms. Para fins de comparação, esta primeira etapa do protótipo foi processada pelo MBROLA, sem consideração nenhuma sobre traços prosódicos, gerando um arquivo de som completamente monótono e pouco natural.

Em seguida tomou-se o cuidado de se aplicar valores de duração para cada fonema, de acordo com a escala de sonoridade, da seção 1.3. No entanto, este ação garante apenas que a sílaba sozinha soe mais natural, mas, quando colocadas uma ao lado da outra, elas

permanecem monótonas e artificiais. A partir de então, passou-se a fazer alterações em nível de sílaba, com a preocupação constante de que o princípio da sonoridade fosse respeitado no interior de cada sílaba.

Quando considerada prosodicamente, a distribuição das palavras e os acentos fica:

['eu] ['que ro] [a es 'tre la] [,da ma 'nhã]

Com base então nos acentos principais e secundários, as sílabas acentuadas foram proporcionalmente alongadas, e as sílabas que sucedem a tônica foram fortemente encurtadas.

As primeiras atribuições de altura foram também feitas a este nível. Cada palavra foi tratada separadamete, com um pico ligeiro de frequência sobre os acentos, e queda após a tônica. Além disso, os acentos principais tiveram sua intensidade de som aumentados, característica principal dos acentos. No entanto, chegou-se a conclusão de que este aumento de intensidade deve ser bastante pequeno, pois pequenas variações já são o suficiente para gerar efeito de acentuação.

Em seguida foi aplicada a curva entoacional, com o seguinte formato:

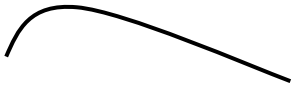


Figura 0.2 - Curva entaocional típica de uma afirmação

Subida até a segunda tônica da frase, e queda a partir de então. A curva entoacional se mostrou a mais difícil de se implementar manualmente, pois ela traça apenas uma tendência na altura durante a frase, mas não obriga as palavras a terem o mesmo contorno. Ou seja, pode-se ter altura ascendente em meio a uma queda da curva entoacional. Como se vê a seguir, a decisão sobre os parâmetros da curva foram de fundamental importância para a determinação dos estados emocionais.

1.13 Tratamento emocional

O primeiro estado emocional produzido foi, no sistema de eixos da seção 1.7.2, com baixa excitação, baixa dominação, assim como baixo contentamento. Tal estado vem associado ao tédio e à melancolia. Características finais da prosódia são

- menor diferença entre as durações das sílabas;
- pouca alteração de volume nas tônicas;
- ritmo menos marcado;
- curva entoacional sem grande pico de altura; e
- queda constante da altura.

A parte tais fatores acústicos, a qualidade da voz foi alterada, adicionando-se um pequena flutuação na frequência fundamental, o que, junto com um aumento geral de f0, proporciona o tom ligeiramente choroso do discurso, sem que ele se pareça entediado.

O segundo estado emocional produzido foi o de alegria. Neste caso, optou-se por um tom imperativo, ou seja, alto grau de dominação e excitação. O estado emotivo foi garantido principalmente pela curva entoacional. Em resumo, a amostra apresenta as seguintes características:

- sílabas com grande variação de duração;.
- pouca variação de frequência no interio das palavras;
- dominância da curva entoacional;
- curva entoacional bem acentuada;
- valores de altura muito baixo para a primeira sílaba, subida brusca da curva até a segunda sílaba acentuada, e queda bem marcada, com ênfase no final da frase;
- nenhum tratamento de qualidade de voz e
- f0 neutro.

O resultado do discurso tratado com emoções, posto ao lado da primeira versão monótona do texto, é de qualidade incomparável.

Desenvolvimento do Modelo Prosódico

1.14 As Classes

A implementação do código consiste basicamente em transcrever a teoria da fonologia prosódica em regras computacionais. Como sugerido pela teoria da fonologia prosódica, utilizam-se neste código entidades de interface entre a gramática e a prosódia, os ditos constituintes prosódicos. No entanto serão feitas simplificações no modelo, eliminando o sintagma fonológico, por exemplo. A seguir estão listados os principais objetos:

- Fonema;
- Sílaba;
- Palavra Prosódica e
- Sintagma Entoacional.

Como não é objetivo deste trabalho a transformação de texto em constituintes prosódicos – o que caberia a um projeto de análise gramatical e semântica –, partiremos do pressuposto de que já é sabido quais são as palavras prosódicas de um sintagma, quais suas respectivas sílabas e assim por diante. O papel de cada um destes elementos será discutido detalhadamente a seguir.

1.14.1 Fonemas

O fonema é, nessa hierarquia, o elemento atômico, ou seja, aquele que carrega consigo o menor nível de informação para caracterizar um som. Além disso, são os únicos que têm importância para o processamento digital do som. Eles se traduzem no input do programa MBROLA, com suas características acústicas individuais, enquanto todos outros níveis de informação são relevantes apenas para caracterizar o fonema. No entanto, estes devem estar tão bem caracterizados, ou seja, suas informações prosódicas devem ser tão consistentes entre si, que o resultado da fala como um todo seja próximo à fala natural humana.

O formato de entrada do programa MBROLA se resume em:

[Código do fonema] [Duração] [Pitch] Vol [Volume]

Ou seja, estão contidas nesta linha todas as informações necessárias para caracterizar o átomo da fala, com suas características prosódicas essenciais.

Como este trabalho não se ocupa da transcrição do texto em fonemas, temos o código do fonema já definido por input. Já os parâmetros restantes devem ser analisados com cuidado.

A **duração** de um fonema está intrinsecamente ligada à sua sonoridade. Fonemas mais sonoros têm, mais do que uma maior intensidade, uma duração mais expressiva. Este parâmetro variará de acordo com os objetos de hierarquia mais alta: a sílaba, a palavra prosódica e o sintagma entoacional. Apesar disso, os valores iniciais da duração de cada fonema devem ser ajustados conscientemente.

A sonoridade dos fonemas terá um papel importante no nível silábico, no cumprimento do princípio da sonoridade, de acordo com o qual a sonoridade é crescente do ataque até o núcleo e decrescente a partir de então. Mas, embora esta seja uma informação inerente à sílaba, ela é intrínseca aos fonemas, já que as sílabas “escolhem” os fonemas de acordo com o princípio da sonoridade. Ou, por outro lado, este princípio define as sílabas.

Assim, os valores iniciais – antes do processamento dos níveis mais altos da hierarquia – serão ajustados de acordo com o código do fonema. Mais tarde, quando os outros níveis interferirem na duração dos fonemas, estes o farão apenas alterando a sílaba como um todo. Assim, cada sílaba permanece coerente internamente, obedecendo ao princípio da sonoridade.

Para determinar os valores iniciais da duração, os fonemas foram agrupados nas seguintes categorias, em ordem decrescente de sonoridade: vogais, semivogais, vibrantes, laterais, nasais, fricativas e oclusivas, sendo que as primeiras tem duração de 120 ms e as últimas de 40 ms.

Já os outros parâmetros do fonema têm uma inicialização mais simples. O pitch é definido por dois valores: um ponto temporal em porcentagem e um ponto de freqüência, em Hz. Neste modelo desconsiderou-se a priori a necessidade de se definir mais de um ponto para o pitch, já que os fonemas têm em geral duração curta. Além disso, os fonemas mais longos e com possivelmente maior riqueza de variações na altura são as vogais, e estas vêm em geral circundadas de consoantes. Como o MBROLA trabalha com interpolação para definir a freqüência, a informação de pitch das consoantes que emolduram a vogal é o bastante para criar um desenho sensível na execução da mesma. Assim, embora o fonema esteja definido

por apenas um ponto de pitch, ele não tem freqüência média constante, e pode sim ter um desenho prosódico mais complexo, de acordo com os fonemas que o circundam.

A mesma lógica é usada para o volume, que vai também ser definido por um ponto único.

Em resumo, todo fonema passa a ter um só ponto de pitch e de volume, localizado no centro do fonema (ponto 50% do MBROLA). O valor de pitch deste ponto é simplesmente a freqüência média da fala. No caso da voz feminina do banco de dados br4, 230 Hz. Como o volume é dado relativamente em dB, este será inicializado em zero.

1.14.2 Sílabas

Sílabas são formadas exclusivamente de fonemas. Tal objeto existe simplesmente para agrupar tais elementos em uma unidade mínima de sentido. Embora os fonemas sejam os átomos da fala, as sílabas são os menores elementos percebidos. Na analogia, elas seriam as moléculas.

A coerência interna da sílaba é dada pelo princípio da sonoridade. Este, no entanto, já vem garantido dos fonemas que a constitui. Assim, a sílaba é, num primeiro momento, apenas um conjunto de fonemas. Tal conjunto, no entanto, tem a capacidade de assumir uma característica prosódica: o fato de ser acentuada ou não. Os valores possíveis para essa característica são:

- Acento Principal;
- Acento Secundário;
- Não Acentuado e
- Após acento principal.

Características de “acentos secundários”, ou sílabas “após acento principal” não precisam ser definidas na criação de uma sílaba, já que, no próximo nível de hierarquia, na palavra prosódica, tais inflexões serão definidas de forma simples, como se verá mais adiante. Desta forma, todas as sílabas têm como valor default “não acentuado”. Na criação de uma sílaba, o usuário deve definir se esta é um acento principal. Posteriormente, os valores “não-acentuados” serão automaticamente atualizados para “acento secundário” ou “após acento principal”.

Além disso, este objeto possui métodos que possibilitam manipular os fonemas nele contidos, alterando durações, pitch e volume de todos seus componentes de forma uniforme.

1.14.3 Palavra Prosódica

A palavra prosódica é um conjunto de sílabas. Sua caracterização se dá por uma série de inflexões de acentuação, que a caracterizam para o ouvinte. Sua principal inflexão é um único acento principal. Do ponto de vista de programação, tal característica de acentuação está inerente a sílaba, constituinte da palavra prosódica.

Desta feita, a classe Palavra Prosódica não tem nenhuma outra característica do que o próprio conjunto de sílabas. Tal agrupamento será importante durante o primeiro processamento, no qual as sílabas terão suas funções atualizadas de acordo com a posição que elas ocupam dentro da palavra.

1.14.4 Sintagma Entoacional

O sintagma entoacional é um conjunto de palavras prosódicas. Este não é, no entanto, apenas um aglomerado de palavras, e possui uma segunda característica: a curva entoacional. Isto, pois se prevê a programação de diferentes curvas entoacionais para um sintagma, de modo que ele possa assumir características afirmativas, interrogativas ou de aposto.

A curva entoacional seria chamada de uma biblioteca de curvas. No entanto, por ora, apenas uma curva entoacional afirmativa será considerada, de maneira que, à rigor, o sintagma entoacional se restringe aos seus constituintes, as palavras.

As características acústicas do sintagma entoacional serão diretamente determinadas durante o segundo processamento com base na posição das sílabas e dos fonemas dentro do sintagma. Como já levantado, a linguagem Scala permite fazer este trânsito de informações entre níveis hierárquicos de forma rápida e pouco burocrática.

1.15 Processadores

A função dos processadores é transitar as informações de um constituinte prosódico para o outro, tendo sempre como fim os fonemas, que são o ponto de partida para o processamento digital do som feito pelo MBROLA. O trânsito das informações é, no entanto, intrincado e não-linear.

Desta maneira os valores de pitch, duração e intensidade dos fonemas serão alterados diversas vezes pelos diferentes processadores a que serão submetidos, os quais levarão em conta, a cada momento, informações de cada nível de constituinte prosódico, sobrepondo os efeitos de todos eles sobre o fonema.

1.15.1 Primeiro Processamento: As Palavras Prosódicas

Antes de se ter a sensação de que uma frase completa foi dita, é necessário que o ouvinte consiga identificar quais são as palavras dentro de um emaranhado de sons, e isso é dado por uma série de inflexões rítmicas, os acentos. Estes se constituem de uma distribuição lógica de sonoridade entre as sílabas, a qual está ligada simultaneamente aos parâmetros acústicos de intensidade, duração e altura.

Na criação das sílabas, o usuário já deve ter definido se esta é um acento principal ou não. Cabe ao programa, a partir desta semente de informação, definir os parâmetros acústicos das outras sílabas da palavra prosódica, de maneira que o ouvinte tenha a sensação de palavra.

1.15.1.1 Pré-Processamento

Antes de agir sobre os parâmetros acústicos, é preciso distribuir os acentos secundários e definir sílabas menos expressivas: aquelas que sucedem o acento principal.

Embora isto não constitua uma regra, existe uma tendência do brasileiro a acentuar secundariamente todas as sílabas pares a esquerda do acento principal. Não são raros, no entanto, casos em que o locutor quer sublinhar determinada palavra, e acaba por colocar apenas um acento secundário na primeira sílaba da palavra. Na busca pela fala natural, é de bom tamanho contentar-se com o primeiro cenário, mais recorrente.

O pré-processamento consiste, portanto, em alterar a função de cada sílaba da palavra prosódica para:

- “acento secundário”, se estiver à esquerda, a um número par de sílabas de distância do acento principal.
- “após acento principal”, se estiver à direita do acento principal.

Estão assim caracterizadas todas as sílabas. Resta, portanto, retrabalhar os parâmetros acústicos dos fonemas com base nestas informações.

1.15.1.2 Intensidade

O parâmetro acústico mais evidente e simples de se alterar é a intensidade. Sílabas com acento principal terão volume mais expressivo que sílabas não acentuadas, etc.

Para simplificar o processamento, os fonemas muitas vezes não serão modificados individualmente, mas sim as sílabas como um todo. Ao se aumentar a duração ou volume de uma sílaba, todos seus fonemas serão alongados ou intensificados proporcionalmente.

De quantos decibéis uma sílaba deve ter seu volume aumentado para simular a fala humana é uma pergunta difícil de responder. Mais que isso, este é um parâmetro que será definido pelo estado emocional do locutor. Nesta etapa do trabalho, considerou-se que um aumento de 2 dB para acentos principais e 1 dB para acentos secundários criavam um ritmo sutil, mas claro o bastante para haver sensação de acentuação. Sílabas após acento principal tiveram sua intensidade reduzida por -1 dB.

1.15.1.3 Duração

Além da intensidade, sílabas acentuadas são mais longas do que as outras. Assim, com base nas informações da função da sílaba, sua duração será alterada de forma multiplicativa conforme a Tabela 0.1:

Sílabas não acentuadas	100%
Acentos principais	120%
Acentos secundários	110%
Após acento principal	80%

Tabela 0.1 - Exemplo de duração das sílabas conforme o tipo de acento.

Tais valores serão alterados dinamicamente mais tarde, em função do estado emocional.

1.15.1.4 Altura

Uma vez que as sílabas já têm uma duração temporal bem definida, é possível alterar os valores de pitch com base um uma curva entoacional simplificada, que dará o contorno entoacional à palavra. Esta curva não deve ser confundida com curva entoacional típica do sintagma entoacional. Esta curva é mais primitiva e está ligada ao aparelho sonoro humano, que tende a utilizar a subir o pitch quando se aumenta a intensidade. Assim, sílabas acentuadas têm pitch mais elevado que as outras. Apenas são mais sensíveis que as variações de duração as variações de altura dentro da palavra prosódica, sendo da mesma ordem de grandeza das variações que serão causadas mais tarde pela curva entoacional do sintagma.

A tabela a seguir apresenta as variações de pitch, segundo cada função da sílaba. Tais valores serão revistos mais tarde, em função do estado emocional.

Sílabas não acentuadas	90%
Acentos principais	110%
Acentos secundários	100%
Após acento principal	80%

Tabela 0.2 - Exemplo de variação de pitch das sílabas conforme o tipo de acento.

No entanto, estes valores devem ser interpolados para cada fonema, de maneira que as inflexões sejam bem precisas. Por isso, o valor de pitch da sílaba não será alterado como um todo, e sim os fonemas serão acessados diretamente para terem seus valores alterados. A linguagem Scala promove tal flexibilidade de maneira simples, sem necessidade de se criar um sem-número de métodos de acesso a classes.

Conhecem-se os valores no tempo de cada sílaba (considerando o centro da mesma), e os valores em pitch que o som deve assumir. Uma função recebe um número n de pares ordenados tempo e pitch e uma posição t no tempo, e devolve o valor em pitch interpolado linearmente entre os pares ordenados mais próximos.

Assim, cada fonema, com sua posição no tempo definida na palavra, recebe um valor de pitch.

1.15.1.5 Resumo do Primeiro Processamento

O pseudocódigo a seguir ilustra a dinâmica deste primeiro processamento.

Pré-Processamento:

Para cada Sílaba, Sílaba.Função =

"acento secundário", **se** à esquerda, a um número par de sílabas do acento principal

"após acento principal", **se** à direita do acento principal

Fim Para

Processa Intensidade:

Para cada Sílaba

Se "acento principal" Sílaba.Volume← 2

Se "acento secundário" Sílaba.Volume← 1

Se "não acentuado" Sílaba.Volume← 0

Se "após acento principal" Sílaba.Volume← -1

Fim Para

Processa Duração:

Para cada Sílaba

Se "acento principal", Sílaba.Duração ← Sílaba.Duração * 1.2

Se "acento secundário", Sílaba.Duração ← Sílaba.Duração * 1.1

Se "não acentuado", Sílaba.Duração ← Sílaba.Duração * 1.0

Se "após acento principal", Sílaba.Duração ← Sílaba.Duração * 0.8

Fim Para

Processa Altura:

Para cada Sílaba

t_i ← posição na palavra prosódica

Se “acento principal”, pitch = 1.1

Se “acento secundário”, pitch = 1.0

Se “não acentuado”, pitch = 0.9

Se “após acento principal”, pitch = 0.8

f_i ← pitch

Lista_Pares **inclui novo** Par (t_i, f_i)

Fim Para

Para cada Fonema

t ← posição na palavra prosódica

Fonema.Pitch ← interpolação (Lista_Pares, *t*)

Fim Para

1.15.2 Segundo Processamento: O Sintagma Entoacional

O primeiro processamento se preocupa basicamente com as inflexões que dão ao ouvinte a sensação de palavra. Não obstante, o som produzido até então não passa de uma seqüência de palavras, e ainda não há nada que as conecte. Cabe então aqui utilizar os conceitos do sintagma entoacional e, sobretudo, aplicar a curva entoacional, para que a sensação de frase se complete.

Este processamento se restringe basicamente à **altura**. Os valores de duração dos fonemas serão mantidos como definidos no processamento anterior, e serão a base para definir a curva de variação do pitch no tempo. Apenas uma pequena variação de volume e intensidade acontecerá, para marcar a tônica do sintagma.

A mesma estratégia para definir o desenho de pitch da palavra prosódica será empregada para confeccionar a curva entoacional. Esta não será, no entanto, constituída de um número arbitrário de pares ordenados. Pelo contrário, serão utilizados poucos pares, de significado bem preciso.

Por ora nos preocuparemos apenas com a curva entoacional afirmativa.

São parâmetros para a curva entoacional:

t_1 : posição do primeiro acento principal

t_2 : posição do último acento principal

t_f : duração completa do sintagma

Uma vez definidos estes pontos no tempo, cabe à função de curva entoacional definir valores de pitch para cada um deles, e devolver o valor do pitch para uma posição no tempo qualquer.

Conforme diversos exemplos da literatura, a curva entoacional afirmativa possui pitch crescente até o primeiro acento principal do sintagma. A partir de então, possui uma queda suave até a tônica, que corresponde ao último acento principal do sintagma. As últimas sílabas do sintagma têm então um decréscimo acentuado do pitch. A Figura 5.1 mostra o desenho padrão que foi adotado para o sintagma entoacional afirmativo.

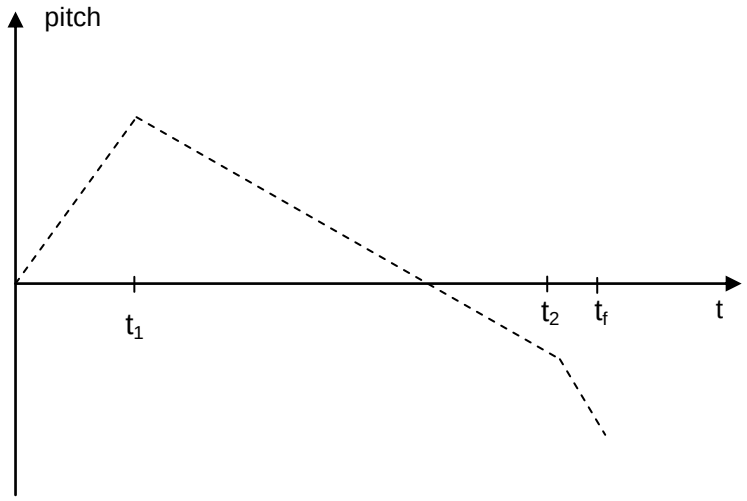


Figura 0.1 – Perfil da curva entoacional para uma sentença simples (com um único verbo e sem apostos ou subordinações)

s pelo primeiro processamento. Já as alturas de pitch foram escolhidas arbitrariamente neste momento, e serão mais tarde determinadas pelo modelo emocional.

O pseudocódigo a seguir resume o funcionamento deste segundo processamento.

Segundo Processamento:

Encontra *última* Palavra **no** Sintagma

Encontra Sílaba *com acento principal* **na** Palavra

Sílaba.Duração ← 1,1 * Sílaba.Duração

Sílaba.Volume ← 1,1 * Sílaba.Volume

t_2 ← Sílaba.Posição_no_sintagma

t_f ← Sintagma.Duração

Encontra *primeira* Palavra **no** Sintagma

Encontra Sílaba *com acento principal* **na** Palavra

t_1 ← Sílaba.Posição_no_sintagma

Para cada Fonema

t ← Fonema. Posição_no_sintagma

Fonema.Pitch ← Fonema.Pitch + curva_entoacional(t, t_1, t_2, t_f)

Fim Para

Note-se que o valor da curva entoacional é aditivo, ou seja, ela se sobrepõe ao desenho prosódico da palavra. Os contornos tanto da palavra quanto da curva entoacional devem ser ambos audíveis, e não diferem muito na ordem de grandeza. Ou seja, em alguns casos de estados emocionais o contorno da palavra pode ser mais expressivo que o da curva entoacional.

1.15.3 Terceiro Processamento: Tratamento do foco

O foco é mais um elemento da fonologia que, quando tratado adequadamente, nos abre inúmeras portas para obtenção de características emocionais na fala.

Formalmente o foco é a parte da frase que responde a uma pergunta específica, mesmo se esta for implícita ou mesmo se for inconsciente. Por exemplo, a frase “Eu fui à escola ontem de bicicleta” pode ser resposta para perguntas como “como você foi à escola ontem”, “quem

foi de bicicleta para a escola", "qual foi a última vez que você foi de bicicleta à escola", entre inúmeras outras. Para cada caso, a intenção da primeira afirmação é diferente, e isto se traduz acusticamente, muitas vezes de maneira bastante sutil.

Neste trabalho, no entanto, o foco será estudado apenas no seu caso mais extremo, ou seja, naquele em que o locutor deliberademente sublinha uma palavra e destaca-a da frase toda. Numa frase de um texto escrito, sem alguma mudança nítida da formatação da palavra com foco, é impossível dar-se conta de sua presença.

Essa dificuldade de representação textual do foco, assim como sua importância para a semântica de um frase, ficam bastante evidentes quando se consideram textos de "chats", conversas, na internet, por exemplo. Quando um dos interlocutores deseja que uma determinada palavra tenha foco, costuma-se diferenciá-la através de sua escrita em letras maiúsculas, ou com negrito, ou finalmente com caracteres tais como "*" ou "_" ao redor da palavra, como nos exemplos a seguir:

> Como foi que VOCÊ conseguiu ganhar aquele jogo?
>
> Eu preciso daquele arquivo *agora*.

Para se obter semântica idêntica à das frases acima, sem o uso de tais diferenciações na forma do texto escrito, seriam necessárias adição e modificação de palavras, o que não é o que se faz na prática num diálogo.

Temos, então, que o foco é uma característica essencialmente acústica da comunicação. Este trabalho propõe um modelo para o tratamento desse elemento na geração de falas.

As principais características prosódicas levantadas para foco são:

- elevação do tom,
- elevação da intensidade,
- maior duração das sílabas, e
- pausas antes e depois da palavra em questão.

No modelo implementado, considerou-se apenas variações constantes no tom, na intensidade e na duração. Seu tratamento é feito depois de todo o processamento das intensidades e durações, bem como das curvas entoacionais da palavra e do sintagma, e se traduz por uma adição de dois fonemas de silêncio, antes e depois da palavra com foco, uma soma de um valor constante ao pitch da palavra, o produto da duração dos fonemas por outra constante, e finalmente a soma de uma última constante à intensidades dos fonemas.

Para o tratamento de foco em palavras, este modelo se mostra satisfatório. Como a palavra à qual se aplica o foco costuma ser curta, não se faz necessário o tratamento mais detalhado das variações acústicas, como a aplicação de uma nova curva entoacional.

1.16 **Generalização do Modelo**

O código descrito anteriormente cumpre a função de tornar um texto extremamente monótono e duro em um discurso compreensível e com menos “sotaque” robótico. No entanto, algumas questões permanecem abertas sobre como expressar o estado emocional deste locutor virtual. E mais, como fazer com que a entrada textual atenda aos moldes de constituintes prosódicos – ou seja, divisão em sílabas, palavras prosódicas e sintagmas – de maneira simples e intuitiva? A resposta para tais questões se resume na parametrização do *modelo prosódico*. Assim, tanto o texto quanto os parâmetros necessários para seu processamento se encontram em arquivos externos ao software em si, podendo ser modificados ao gosto do usuário.

1.16.1 **Parametrização**

Todos os valores numéricos utilizados na descrição dos algoritmos anteriores serão de importância fundamental para a caracterização da emoção na voz. Embora, para fins de simplicidade, tenham sido dados valores fixos para duração de acentos principais, etc, estes devem, na realidade, permanecer em aberto, para serem determinados pelo *modelo emocional*. Estes parâmetros são:

Parâmetros gerais:

- Frequência fundamental média
- Duração média da sílaba

Parâmetros de acentuação da palavra:

- Fator multiplicativo de duração
- Fator multiplicativo de pitch
- Fator aditivo de intensidade

para cada um destes tipos de acento:

- Acento principal

- Acento secundário
- Sem acento
- Após acento principal

Parâmetros aditivos da curva entoacional:

- Pitch inicial
- Pitch primeiro acento principal (pico)
- Pitch do último acento principal
- Pitch final

Parâmetros da tônica do sintagma (último acento principal)

- Fator multiplicativo de duração
- Fator aditivo de intensidade

Parâmetros do foco:

- Fator aditivo de pitch
- Fator multiplicativo de duração
- Fator aditivo de intensidade
- Duração do silêncio anterior
- Duração do silêncio posterior

Totalizando 27 parâmetros.

Abaixo pode-se ver um arquivo típico de parâmetros, caracterizando um estado emocional neutro:

```
pitch.average=230
syllable.duration=190

word.intensity.primary=-2
word.intensity.secondary=-3
word.intensity.nostress=-3
word.intensity.post=-3

word.duration.primary=1.1
word.duration.secondary=1.0
word.duration.nostress=1.0
word.duration.post=0.95

word.pitch.primary=1.05
word.pitch.secondary=1.0
word.pitch.nostress=0.95
```

```
word.pitch.post=0.8

intonationcurve.begin=-40
intonationcurve.firststress=40
intonationcurve.laststress=-100
intonationcurve.end=-160

foco.pre.silence.duration=50
foco.post.silence.duration=30
foco.pitch=30
foco.duration=1.3
foco.intensity=6
```

1.16.2 **Formato de Entrada**

Além do objetivo de estudar a prosódia e a influência do estado emocional sobre fala, este trabalho se propõe a criar um software capaz de sintetizar fala com nuances emocionais. Para tanto, deve haver uma interface simples o bastante de se utilizar para a entrada do texto, de modo que, tanto para pesquisadores como para usuários leigos, seja possível entrar com diversos textos, testando diversas combinações de emoções. Além disso, é interessante que futuros usuários possam desenvolver novas combinações de parâmetros, refinando e gerando mais estados emocionais.

Este formato de entrada deve cumprir algumas características:

- Deve conter o texto e a emoção que lhe será dada;
- O texto deve estar escrito na linguagem fonética do banco de dados do Mbrola, já que a transcrição fonética não é objetivo deste trabalho;
- Devem estar discriminadas as estruturas de sílabas, palavras prosódicas e sintagma, bem como marcações de sílaba tônica e foco do sintagma (opcional);
- Deve ser rápido e amigável para que o usuário, com algum treino, possa ser capaz de escrever o texto no devido formato de maneira rápida;
- Simples o bastante, do ponto de vista de desenvolvimento, pois esta interface não é o objetivo principal deste trabalho.

A maneira mais simples encontrada é trabalhar diretamente em um editor de texto, e enviar o arquivo como parâmetro para a execução do software, chamado de pho-player. O software deve então ler dinamicamente o arquivo de entrada e gerar a estrutura de dados necessária para o processamento dos dados acústicos.

O formato de entrada pode ser exemplificado da seguinte maneira, para a frase “Estou muito contente que o você veio!”:

```
{feliz}
e s
‘t o w

!‘m un y
t w

k on
‘t en
t y

k y
v o
‘s e

‘v e y
w
```

Como pode se ver, todo sintagma é iniciado com a declaração de qual aspecto emocional este sintagma terá. Em seguida, cada sílaba é escrita em uma linha, sendo o espaço a delimitação entre os fonemas desta sílaba e a linha em branco a distinção entre palavras prosódicas. Além disso, as aspas simples marcam os acentos principais, e são obrigatórias em cada palavra.

Opcionalmente, pode-se dar foco para certas palavras, inciando-as com um ponto de exclamação. No exemplo anterior, a palavra “muito” foi destacada.

Diversos sintagmas podem estar descritos em um mesmo arquivo de entrada, desde que comecem com uma declaração de emoção.

1.16.3 **Definição das emoções**

Para uma maior riqueza de estados emocionais possíveis de serem desenvolvidos, decidiu-se por manter os parâmetros em arquivos externos ao código principal. Assim é possível ao usuário:

- Modificar os parâmetros de uma emoção já existente e
- Criar novos estados emocionais, conforme lhe convir.

O nome dos arquivos contendo os parâmetros obedece à seguinte sintaxe:

`parameters.emocao.properties`

Onde *emocao* é o nome do estado emocional, conforme determinado no arquivo de entrada do sintagma. Ou seja, para o exemplo da frase citado acima, deve haver uma arquivo com o nome *parameters.feliz.properties*. Caso o usuário solicite uma emoção para a qual não existe um conjunto de parâmetros correspondente, uma mensagem de aviso será mostrada na tela e o arquivo *parameters.default.properties* será carregado.

Em suma, tem-se já um modelo completo para síntese de voz e definição de emoções através de parâmetros prosódico-acústicos. Resta, para a conclusão deste trabalho de formatura, apenas a definição de conjuntos de parâmetros que possam dar o colorido necessário à fala, aproximando-a da fala emotiva. Não obstante, serão desenvolvidos apenas alguns conjuntos de parâmetros, cabendo aos interessados desenvolverem e refinarem outros estados emotivos, completando este modelo, conforme a disponibilidade de parâmetros.

Desenvolvimento do Modelo Emocional

Com um modelo prosódico parametrizado, resta ainda definir quais serão seus parâmetros. Estes se traduzirão em aspectos emotivos na fala.

Conforme explicado anteriormente, a estratégia adotada é a definição de um número finito de estados emotivos, cada um relacionado a um conjunto de parâmetros prosódico-acústicos. Ou seja, no campo das emoções, só existem valores discretos, cada um com um rótulo condizente, como “feliz”, “eufórico”, “cansado” etc. Tal rótulo servirá ao usuário quando este quiser dar tons emocionais ao texto a ser sintetizado.

No entanto, o campo das emoções está longe de ser bem definido e fácil de rotular. Emoções como “feliz” e “muito-feliz” podem ter diversas gradações, assim como podem se manifestar acusticamente de forma muito diferente do estado “eufórico”, que nosso senso comum consideraria como emoção vizinha.

O modelo tri-dimensional, como apresentado por Schröder (SCHRÖDER, 2006), propõe-se a reduzir estes muitos rótulos, considerando cada estado emotivo como uma combinação de três eixos independentes: excitação, dominação e satisfação.

Esta abordagem será útilizada neste trabalho para a compreensão dos estados emocionais, na tentativa de entender como o estado “triste” se aproxima acusticamente do estado “feliz”, ou o que este tem em comum com estado “bravo”. No entanto, para fins de usuário, não faz sentido utilizar a abordagem tri-dimensional. Assim sendo, ela será apenas uma referência na construção de estados emocionais discretos.

Também não é pretensão deste trabalho cobrir a infinita gama de estados emotivos que podem surgir da variação dos parâmetros prosódico-acústicos. Contentar-nos-emos com a definição de alguns poucos estados emotivos contrastantes entre si, e que possam ser identificados como tais.

1.17 Compreensão do modelo tri-dimensional

Com fins de auxiliar a criação dos conjuntos de parâmetros para algumas emoções, vale a pena compreender como o modelo tri-dimensional de emoções se relaciona com os parâmetros disponíveis.

Conforme mostrado na Figura 0.1, existem diferentes correlações entre parâmetros acústicos e eixos emotivos para voz masculina e feminina. Somente a voz feminina será tratada a partir deste momento.

1.17.1 **Eixo de Excitação**

O eixo de excitação é aquele que tem relações mais claras com os parâmetros acústicos. De maneira geral, está ligado com aumentos expressivos de quase todos os parâmetros, tanto os de valor médio quanto os de variação. Destacam-se as seguintes relações:

- Aumento expressivo de f0 médio
- Aumento expressivo do espectro de f0
- Grande variação média de magnitude durante subidas e descidas de f0
- Grande variabilidade destas variações (irregularidade)
- Pequeno aumento nos tempos de subida e descida de f0
- Queda na duração das pausas
- Picos de intensidade bem definidos

1.17.2 **Eixo de Satisfação**

Um aumento no eixo de contentamento significa, em geral, um abrandamento de todos os parâmetros acústicos, destacando-se:

- Pequena queda em f0 médio
- Pequena queda no espectro de f0
- Pequenos gradientes na variação de f0
- Menor intensidade média
- Picos um pouco menos acentuados

1.17.3 **Eixo de Dominação**

Um aumento no eixo de dominação está relacionado com:

- Queda significativa de f0 médio
- Pequeno aumento na variabilidade da duração da queda de f0
- Pequena queda no gradiente das quedas e subidas de f0
- Poucos picos de intensidade por segundo
- Picos de intensidade mais acentuados
- Esforço na voz

1.18 Confecção das Emoções

O entendimento de como os diversos eixos se relacionam com os parâmetros acústicos viabiliza um ponto de partida para criar diferentes emoções. Tentarão ser criados quatro diferentes estados emocionais: “feliz”, “triste”, “bravo” e “neutro”.

Estes quatro estados podem se traduzir no modelo tri-dimensional conforme a seguinte tabela abaixo:

Tabela 0.1 - Posicionamento dos estados emocionais neutro, feliz, triste e bravo no sistema tri-dimensional

	Excitação	**Satisfação**	**Dominação**
Neutro	baixa	baixa	baixa
Feliz	média	**alta**	baixa
Triste	baixa	muito baixa	baixa
bravo	**alta**	muito baixa	**alta**

O estado feliz se compõe, por exemplo, basicamente do eixo da *satisfação*, portanto os parâmetros que se relacionam com este eixo devem ser bastante expressivos. Isto significa um abrandamento da variação de todos os parâmetros.

Por outro lado, o eixo da *excitação* deste estado “feliz” se correlaciona com os parâmetros de forma bem mais intensa que o eixo da *satisfação*, de modo uma excitação média já seja o bastante para alterar significativamente os parâmetros sobre o qual o eixo tem efeito. A *excitação*, no entanto, vai contra o abrandamento das variações.

Desta maneira, cabe analizar quais os parâmetros se relacionam de maneira mais forte. Para tanto, na figura mostrada em Figura 0.1, as flechas para cima serão representadas por valores positivos, as flechas para baixo por valores negativos, e sua magnitude será expressa em inteiros de -4 a +4. A seguir as emoções serão compostas combinando os valores dos eixos. Os estados serão calculados de acordo com a seguinte fórmula:

feliz = 5 x **excitação** + 10 x **satisfação** + 0 x **dominação**

triste = 0 x **excitação** – 5 x **satisfação** + 0 x **dominação**

bravo = 5 x **excitação** - 5 x **satisfação** + 10 x **dominação**

Deve-se atentar para o fato de que utilizaram-se valores negativos de satisfação, para ressaltar o efeito contrário do que expressa o eixo. Além disso, não pode ser esquecido que estes multiplicadores foram escolhidos arbitrariamente e as relações encontradas a seguir são apenas um ponto de partida para afinação dos parâmetros.

Tabela 0.2 - Correlação qualitativa entre os eixos do modelo tri-dimensional, e a determinação dos estados emocionais.

	Excitação	**Satisfação**	**Dominação**	*Feliz*	*Triste*	*Bravo*
f0 médio	4	-1	-3	10	5	-5
Espectro de f0	4	-1	0	10	5	25
Magn. Média das subidas de f0	3	0	0	15	0	15
Variação das Magn. das subidas de f0	4	-1	0	10	5	25
Magn. Média das quedas de f0	3	-2	0	- 5	10	25
Variação das Magn. das quedas de f0	4	1	0	10	-5	15
Duração média das subidas de f0	1	0	0	5	0	5
Variação da duração das subidas de f0	1	0	0	5	0	5
Duração média das quedas de f0	2	0	0	10	0	10

Variação da duração das quedas de f0	1	-2	1	- 15	10	30
Gradiente médio das subidas de f0	3	-2	-1	- 5	10	15
Gradiente médio das quedas de f0	3	-2	-1	- 5	10	15
Dinâmica nos picos de intensidade	2	-1	1	0	5	20

Alguns parâmetros mostrados na Tabela 0.2 podem ser diretamente alterados no modelo prosódico construído, como a frequência média f0. No entanto, é necessária uma certa interpretação das outras variáveis, para que possam se traduzir nos parâmetros disponíveis no modelo prosódico.

Por exemplo, o espectro de f0 pode ser facilmente interpretado como a imagem da função da curva entoacional. Ou seja, para um espectro mais amplo de f0, pode-se contar com uma curva entoacional bem expressiva.

Já a magnitude das subidas e descidas de f0 pode ser traduzida no desenho melódico da palavra prosódica. As durações médias de subida e descida estão também diretamente ligadas ao tamanho médio das sílabas, já que a alternância entre sílabas fracas e fortes se reflete em subidas e descidas de f0, e assim por diante.

Vale lembrar que as alterações se dão com base em um referencial *neutro*. Este foi alcançado sem grande esforço com a variação manual dos parâmetros, e não será tratado mais a fundo.

1.18.1 Estado Feliz

Depois de analisar os resultados da tabela Tabela 0.2, podemos deduzir os parâmetros que servem para caracterizar o estado *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

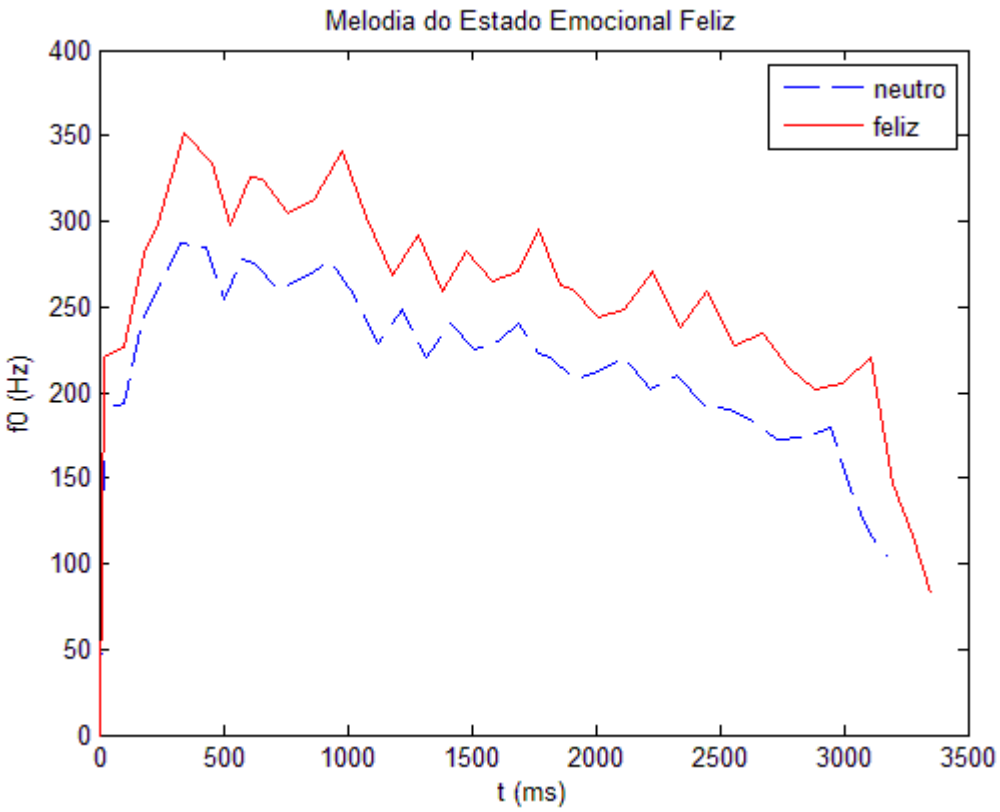


Figura 0.1 – Variação da frequência fundamental f0 ao longo do tempo, para a frase "Sistemas de síntese de voz também têm emoções", com os estados emocionais neutro e feliz.

De todos os estados emocionais, o estado feliz é aquele que menos se distancia do estado neutro. Vemos, na Figura 0.1, que a curva melódica da frase se difere da curva neutra apenas por um certo exagero dos parâmetros. Isto é compreensível se compreendermos que o estado feliz aqui proposto não é um estado eufórico (de altíssima excitação e inclusive um certo grau de dominação). É apenas o estado de alguém satisfeito e ligeiramente ativo.

O resultado auditivo não fica longe do esperado: o aspecto de felicidade surge como uma sutileza. Poderia bem ser considerado neutro, mas ganha mais expressão quando posto diretamente ao lado de uma fala dita neutra.

Entende-se que este resultado, embora sutil, faça bastante sentido, quando comparado com relações pessoais típicas: é bastante complicado saber se um desconhecido está feliz ou triste, mas o dizemos de pronto quando se trata de um amigo ou familiar. Ou seja, boa parte do julgamento do estado emocional é dependente de um referencial.

Um exemplo de frase com tratamento neutro e feliz se encontram nos arquivos anexos, com nome *Exemplo1-Neutro.wav* e *Exemplo1-Feliz.wav*. Estes podem também ser confrontados com um exemplo sem nenhum tratamento prosódico em *Exemplo1-Monótono.wav*.

1.18.2 Estado Triste

Como referência para a criação do estado *tristes*, estão os parâmetros:

- 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)

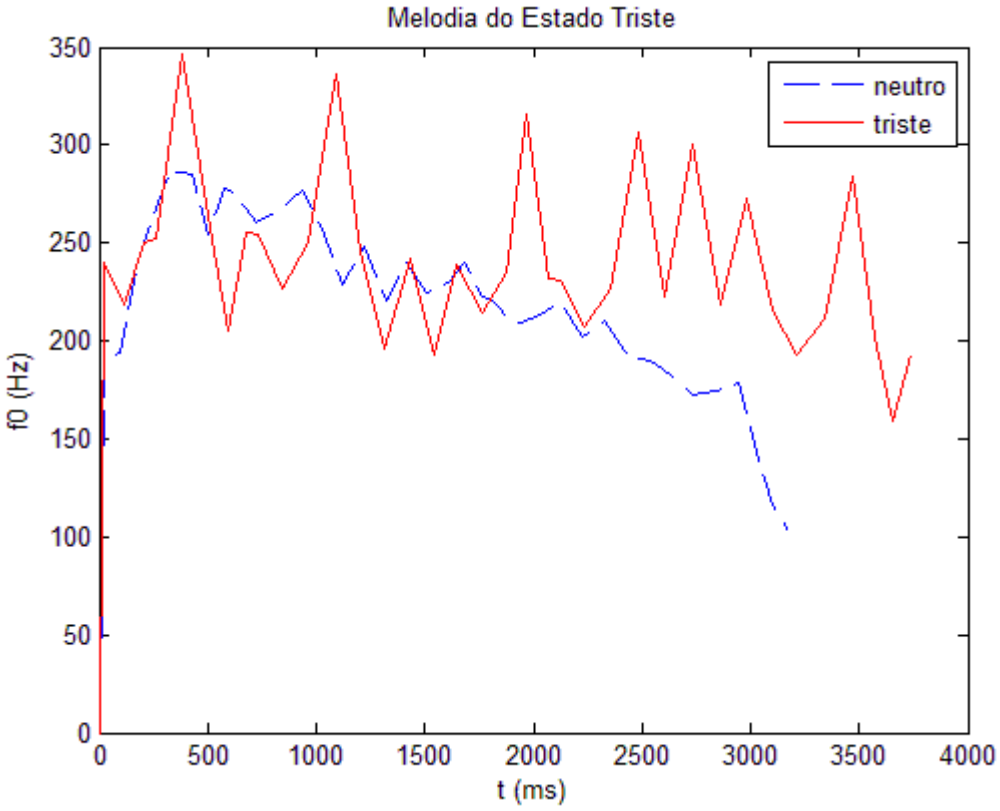


Figura 0.2 - Variação da frequência fundamental f0 ao longo do tempo, para a frase "Sistemas de síntese de voz também têm emoções", com os estados emocionais neutro e triste.

No estado triste, com a ativação baixa, a melodia das palavras se sobrepõem fortemente à curva entoacional, como se vê na Figura 0.2. A marca da acentuação se dá muito mais por esta melodia e pela variação das durações do que pelos golpes de intensidade.

Quando ouvido, o resultado obtido é mais evidente do que o estado feliz. Pode, no entanto, ser facilmente confundido com tédio, ou medo. De qualquer maneira, o aspecto “negativo” deste estado emocional fica evidente na flutuação de f0 e na curva entoacional típica.

Exemplo deste estado pode ser encontrado no arquivo *Exemplo1-Triste.wav*.

1.18.3 Estado Bravo

Utilizando como referência a tabela Tabela 0.2, os parâmetros que caracterizam o estado *bravo* são:

- 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

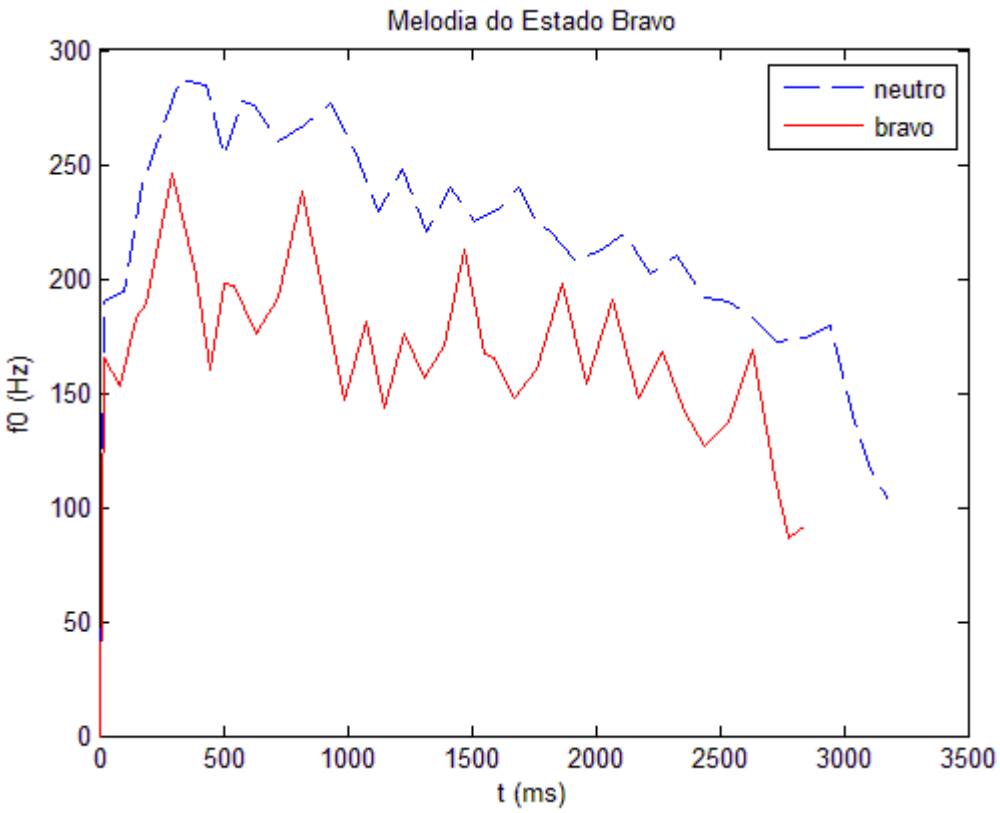


Figura 0.3 - Variação da frequência fundamental f0 ao longo do tempo, para a frase "Sistemas de síntese de voz também têm emoções", com os estados emocionais neutro e bravo.

Picos incisivos de f0 e de intensidade, caracterizam uma acentuação exagerada de cada palavra. A frequência média diminuída, bem como o ritmo acelerado, como se vê na Figura 0.3, denotam uma certa afetação que, juntamente com o aspecto de esforço na voz, dão o tom necessário para a identificação do estado *bravo*.

De todos os estados emocionais produzidos, este é o que mais se faz evidente na audição. Tal resultado não é exatamente supreendente: diferentemente do estado *feliz* e *triste*, que são íntimos do locutor, o estado *bravo* é naturalmente voltado ao receptor, e é de fácil percepção para qualquer um. Desta maneira, o estado referencial *neutro* não precisa ser conhecido para que o discurso soe irritado, já que estamos naturalmente condicionados a reconher este estado emocional no nosso cotidiano, independente da intimidade que temos com o locutor.

Em anexo, no arquivo *Exemplo1-Bravo.wav*, pode-se verificar o resultado do tratemento acústico realizado.

1.19 Avaliação dos Resultados

Foram criados três estados emocionais, mais um estado neutro, alterando parâmetros do modelo prosódico desenvolvido. Os modelos foram aplicados a diversas frases. Embora alguns estados emocionais sejam mais facilmente identificados – no caso, o estado bravo se destaca de todos os outros como sendo o mais claramente identificável –, todos os outros estados apresentam contrastes relevantes entre si.

Um dos fatores que compromete fortemente a qualidade do discurso é o timbre robótico. Este têm duas causas: a qualidade do banco de dados de dífonos e o método de sintetização do Mbrola. Ambos fogem do escopo deste trabalho. Esta particularidade foi considerada crítica para a realização de testes com grande público, de modo que se pudesse analisar os resultados estatisticamente. A avaliação foi feita, portanto, de forma individual apenas por colegas próximos. Pediu-se que os ouvintes não levassem em consideração o timbre robótico dos exemplos.

Embora um teste formal com grande público não tenha sido realizado, ao mostrar os exemplos da seção anterior (a frase “Sistemas de síntese de voz também tem emoções”) para colegas, algumas conclusões podem ser tiradas com respeito a identificação direta do estado emocional. Em geral, a resposta apresentada foi:

“[a frase] claramente tem emoção, mas não sei dizer exatamente qual. Poderia ser X ou Y”.

Na maioria das vezes, o ouvinte acerta a emoção dentro de seus chutes. Isto vai ao encontro da ideia de que não bastam os caracteres acústicos do som para o julgamento completo da emoção, vindo também à baila expressões e gestos físicos, e, sem sombra de súvida de grande importância, o conteúdo textual.

1.19.1 Significado Textual Posto à Prova

Para comprovar a estreita relação entre a informação acústica e a textual, realizou-se o seguinte teste:

Foi dada à característica de *triste* à frase “eu consegui tirar dez na prova”. A maioria dos ouvintes identificou a frase como estranha e pouco natural. Paralelamente, deu-se o tom de *feliz* à mesma frase, e ela foi apresentada a outras pessoas. A maioria destas identificou a frase como “natural e bem pronunciada”, embora a emoção feliz não estivesse muito evidente. Ambos exemplos encontram-se em anexo, com os nomes *Exemplo2-Feliz.wav* e *Exemplo2-Triste.wav*.

Disto concluímos que certos conteúdos textuais exigem um determinado tratamento acústico para que a frase soe **natural**.

1.19.2 Naturalidade e Variabilidade

Como proposto na seção 1.7, o fator crucial para a naturalidade da fala é a variabilidade. Neste sentido, pode-se considerar o resultado dos estados emocionais menos como expressões de sentimentos arrebatadores do locutor virtual – já que os resultados atingidos não foram o bastante para a identificação imediata do estado emocional – e mais como manifestações da variabilidade do discurso natural.

Foi unânime entre os ouvintes que os exemplos apresentados soam mais naturais que a versão monotônica (f0 constante, fonemas de duração e intensidade constante). Ademais, frases com conteúdo textual tendendo claramente para um estado emocional tiveram seu significado aumentado quando a elas foram dadas características emocionais – vide exemplo anterior.

Todos estes fatores apontam para uma fala menos maquinal e mais natural.

1.19.3 Percepção e Importância do Foco

Foi constatada a importância do conteúdo textual na identificação da emoção, ao lado dos traços prosódicos. Resta ainda um fator a ser considerado na percepção do discurso: a importância que pode ser dada a certas palavras. Ao ressaltar uma determinada palavra, estamos fazendo uma escolha acústica para evidenciar um elemento textual.

Assim, como era de se esperar, ao colocar alguns elementos de foco, as emoções dos exemplos se tornaram mais óbvias ao ouvinte. Se isto se deu pela escolha da palavra a receber o foco, ou pelo tratamento acústico em si, será de difícil avaliação. Abaixo estão algumas frases apresentadas. Os grifos representam os termos em foco.

"Páre de brincar é vá fazer o <u>seu</u> dever de casa!"
(*Exemplo3-Bravo.wav*)

"Eu <u>consegui</u> terminar a tarefa!"
(*Exemplo4-Feliz.wav*)

"Eu já esqueci de <u>tudo</u>"
(*Exemplo5-Triste.wav*)

Disto conclui-se que o tratamento de foco é crucial para a percepção da intenção do locutor.

Considerações Finais

Além da complexidade textual e toda a riqueza inerente ao discurso escrito, o discurso falado traz consigo uma gama imensa de variáveis que o tornam em um primeiro momento quase intangível. No entanto, as pessoas estão em geral aptas a processar o conteúdo textual juntamente com o conteúdo acústico da fala, de maneira a gerar o sentido.

No estudo de como implementar tal variabilidade na fala computacional, de forma a aproximá-la da fala humana e torná-la menos maquinal, foi levantada uma série de qualidades acústicas do som, as quais carregam consigo significado, chamados traços prosódicos. No entanto, estes traços não bastam para o estudo sistemático do complexo funcionamento prosódico das línguas. Para isso valeu-se da teoria da Fonologia Prosódica de Nespor e Vogel, com seus ditos constituintes prosódicos.

Diferindo ligeiramente dos modelos encontrados na literatura, o modelo emocional desenvolvido não leverá em consideração apenas variáveis acústicas. Estas ficarão reduzidas a algumas poucas. Tal fator será compensado pelos parâmetros prosódicos, como a sonoridade dos acentos e a curva entoacional, entre outros, já que não se trata de uma ressíntese da voz neutra, e portanto a informação prosódica não será perdida.

Desta maneira, em um primeiro momento, este trabalho preocupou-se em dar ao texto características prosódicas como acento, entoação e ritmo, de maneira a diminuir o “sotaque” do computador. Um modelo computacional foi desenvolvido abstraindo e simplificando a teoria da Fonologia Prosódica, possibilitando alterar as características acústicas do som, de acordo com as informações prosódicas. Para tanto, parte-se de um texto já escrito em simbologia fonética, no qual os constituintes prosódicos já estão definidos.

A seguir, baseado no modelo tri-dimensional de emoções, foram desenvolvidos conjuntos de parâmetros para caracterizar quatro estados emocionais: neutro, feliz, triste e bravo.

1.20 Sucesso do Projeto

Pode-se dizer que, dada a complexidade do tema tratado, os resultados obtidos foram mais do que satisfatórios. À parte alguns casos pontuais, a maioria dos exemplos alcançou um alto grau de naturalidade, o que demonstra que o modelo prosódico em muito contribuiu para a dicção do computador.

Além disso, foram dadas a diversas frases características emocionais, o que as tornou ainda mais compreensíveis e naturais. Não obstante, a riqueza gerada pelos diferentes estados emocionais contribui para para a variabilidade da enunciação, o que torna também o discurso mais natural.

Para frases emocionalmente ambíguas, a maioria dos ouvintes foi capaz de identificar que a frase fora dita de maneira afetada – com emoções –, mas em muitos casos não pode dizer de qual emoção se tratava, com exceção do estado bravo, o qual ficou bastante claro. Isto, pois as alterações nos parâmetros prosódicos possibilitaram apenas mudanças sutis no estado emocional, as quais, sem ajuda do conteúdo textual ou gestual, não deixam perfeitamente clara a emoção em questão.

Não consideramos grave, no entanto, que os estados emocionais alcançados sejam sutis, considerando muito mais importante o fato de que, quando postos lado a lado, o contraste entre as diferentes pronúncias de uma mesma frase seja evidente.

Além disso, foi criado também um tratamento de foco, o qual permite destacar certas expressões do conteúdo textual, evidenciando assim a intenção do locutor. Tal tratamento foi de importância fundamental na percepção da emoção.

1.21 Ressalvas ao Modelo Proposto

Embora os resultados obtidos sejam, em certa medida, bastante satisfatórios quando desprezado o timbre metálico e robótico da fala, algumas ressalvas devem ser feitas à qualidade do modelo proposto para a caracterização das emoções.

Diferindo drasticamente de grande parte da literatura encontrada, o tratamento emocional aqui está intrinsecamente ligado ao tratamento prosódico. No entanto, é possível que parte considerável do conteúdo emocional da frase esteja dissociado dos níveis de constituintes prosódicos. Ou seja, ao utilizar o sistema de fonologia prosódica, vários parâmetros acústicos foram deixados de lado.

Tal escolha, no entanto, se baseou em dois fatos: o primeiro é de que seria necessário, inevitavelmente, um tratamento prosódico para que a frase monotônica ganhasse sentido, graças a acentuações e curva entocional. O segundo é de que alterar os parâmetros do modelo prosódico para alcançar nuances emocionais se apresenta como uma ótima oportunidade para traçar paralelos entre a prosódia e os estados emocionais, ao invés de lidar com os traços prosódicos apenas em termos acústicos.

1.22 Perspectivas

Ainda que o modelo aqui proposto esteja limitado pela abordagem da Fonologia Prosódica, muito ainda pode ser feito para alcançar resultados mais expressivos. O mais evidente é o estudo de outras formas de curva entoacional, tanto para outras naturezas de sintagma, tais como apostos e interrogações, como para as diferentes emoções. Isto aumentaria exponencialmente a quantidade de diferenças entre os estados emocionais e curvas melódicas do discurso.

Além disso, no modelo proposto, o tratamento de acentuação é homogêneo para todo um sintagma. Variações intra-sintagma também podem ser propostas em futuras expansões do modelo prosódico.

Uma vez obtido um modelo completo o bastante, com parâmetros tais que emoções bastante efusivas possam ser exprimidas, as atenções devem ser voltadas para o modelo emocional.

Neste trabalho, foi proposto um número finito de estados emocionais, o que naturalmente não corresponde à realidade humana. Idealmente, o sistema computacional deveria ser capaz de percorrer todo o espaço tri-dimensional criado pelos eixos de excitação, satisfação e dominação, gerando para cada ponto um conjunto de parâmetros prosódicos correspodente. Com isso, uma gama infinita de emoções complexas poderiam ser expressas.

Assim, numa visão futurística, o computador seria capaz de determinar, por mecanismos internos de inteligência artificial, em qual estado este se encontra, e expressar-se assim com a emoção correspondente. Menos futurística seria a capacidade do computador de identificar, por meio de análise semântica, qual o estado emocional está atrelado ao texto, e escolher emoção correspondente para declamá-lo. Enfim, aplicações estão limitadas apenas pelas fronteiras da criatividade.

Muito pode ainda ser feito no campo de síntese de voz com emoções. Esperamos sinceramente que este trabalho de formatura sirva de inspiração para futuras pesquisas neste campo, e que muito ainda seja feito para a língua portuguesa.

Referências Bibliográficas

ANDRADE, J. N. (1841). *Grammatica elementar da lingua portugueza por systema.* Lisboa: A.S.Coelho.

Associação de Informação Terminológica. (n.d.). Retrieved Junho 02, 2010, from Dicionário de Termos Linguísticos: http://www.ait.pt/recursos/dic_term_ling/index2.htm

BARBOSA, J. S. (1821). *Grammatica philosophica da lingua portugueza ou principios da grammatica geral applicados á nossa linguagem.* Lisboa: Typographia da Real Academia das Sciências. 1871.

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

BURKHARDT, F. (2006). Emotional Speech Synthesis: Applications, History and Possible Future. *ESSV* .

COELHO DE CARVALHO, J. J. (1910). *Prosodia e ortografia.* Lisboa: Imprensa Nacional.

CRYSTAL, D. (1994). *A Dictionary of Linguistics and Phonetics 3rd ed.* Cambridge: Mass.: Blackwell.

DUTOIT, T., & LEICH, H. (1993). MBR-PSOLA: Text to Speech synthesis based on a MBE re-synthesis of the segments database. *Speech Communication* .

FRAGOSO, K. (2009, Dezembro). Sobre a universalidade do grupo clítico como domínio de regras fonológicas. *Letronica v.2* , pp. 101-113.

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* (pp. 533-555). Braga: APL.

HENRIQUES, I. (2009). A importância da sílaba: uma reflexão fonológica. *Volume 1, Número 1* (pp. 37-59). Centro de Linguística da Universidade do Porto.

LEMMETTY, S. (1999). *Review of Speech Synthesis Technology (Master Thesis).* Helsinki University of Technology. Espoo.

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.

MATEUS, M. H., & RODRIGUES, C. (2003). A vibrante em coda no Português Europeu. *Teoria Lingüística. Fonologia e outros temas* (pp. 181-199). Instituto de Linguística Teórica e Computacional.

NESPOR, M. &. (1986). *Prosodic Phonology.* Dordrecht: Foris.

PEPPERKAMP, S. (1996). On the prosodic representation of clitics. *Studia Gramatica - Interfaces in Phonology.* Kleinhenz, Ursula: Akademie Verlag.

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

TAO. (2006, Julho). Prosody Conversion From Neutral Speech. *IEEE Transactions on audio, speech, and languague processing, vol. 14, no. 4* .