Bancos de Fala para o Português Brasileiro

Vanessa Marquiafavel Serrani
Universidade Estadual Paulista (IBILCE)
marquiafavel@gmail.com

Luis Felipe Uebel
ASR Labs
luis.uebel@asrlabs.com

Resumo

Reconhecimento e síntese de fala demandam grandes quantidades de dados para aplicações comerciais. Em inglês americano, existem grandes quantidades de bancos de fala para a produção de modelos acústicos. Isto não é realidade para muitas línguas, incluindo o Português Brasileiro. Este trabalho apresenta o desenvolvimento de dois bancos de fala para reconhecimento com 248 locutores (224 sentenças) e outro com 1.226 locutores (550 já gravados e 665 sentenças), e um banco para síntese com 1.220 sentenças. Também é mostrado um sistema para selecionar as sentenças gravadas, dicionário fonético para suportar esta seleção, e modos de gravar e validar um grande banco de fala.

1. Introdução

O reconhecimento de fala permite que um computador possa traduzir um comando vocálico produzido por um humano em comando para abrir uma página na Internet, comandar um robô, ditar uma carta, ou transcrever uma conversa em vídeo ou via telefone. O processo inverso da pronúncia de um texto é chamado de síntese de fala.

O reconhecimento e a síntese de fala necessitam de amostras de vozes humanas para poderem trabalhar com eficiência. No reconhecimento de fala, a métrica mais importante é o nível de reconhecimento. Caso o sistema não traduza corretamente para um texto o que o usuário pronunciou, de nada vale o mesmo. Outros fatores como o consumo de memória ou de processamento deixam de ser relevantes caso o fator mais importante, o nível de reconhecimento, não for elevado. Com o objetivo de alcançar altos níveis de reconhecimento, são necessárias amostras de todos os fonemas usados na aplicação, sendo pronunciados por uma quantidade muito elevada de locutores com todos os sotaques presentes naquela língua. Isto equivale a dizer que uma grande quantidade de locutores necessita gravar uma quantidade elevada de sentenças.

Na síntese de fala, é necessária uma quantidade elevada de sentenças sendo pronunciadas por um único locutor. Existem técnicas de produção de vozes sintetizadas que necessitam de pouca quantidade de sentenças, ou a voz pode ser produzida por uma quantidade grande de locutores usando poucas sentenças cada. Para alcançar o nível de naturalidade e inteligibilidade a nível comercial é necessário que um único locutor grave centenas de sentenças com a mesma entonação, volume, velocidade e expressão, o que usualmente é chamado de "*persona*" da voz (Cohen 2004), ou seja, associar voz personalizada a uma marca ou empresa.

De um lado existe a necessidade de uma grande quantidade de fala para a formação do banco (reconhecimento e síntese) e de outro lado existem aspectos financeiros e de tempo para a construção de um banco desta magnitude.

Dependendo da aplicação, o banco pode ser constituído simplesmente dos comandos usados nesta aplicação. Comandos como "siga", "pare", "para direita", "para a esquerda", podem ser gravados quando a aplicação for simplesmente comandar algumas funções de um robô. Outro aspecto interessante é determinar qual o público alvo da aplicação. Caso sejam jovens entre 18 e 20 anos, bastar encontrar uma grande quantidade de jovens nesta faixa etária dispostos a gravar esses quatro comandos.

Quando a aplicação é mais complexa, como ditar uma carta ou reconhecimento de fala espontânea presentes em vídeos postados na Internet, a quantidade de pessoas necessárias e a quantidade de sentenças aumentam. Neste caso é preciso achar um conjunto de sentenças que possuam todos os fonemas presentes no Português Brasileiro.

Quanto maior o conjunto de sentenças a serem gravadas, maiores são os custos e o tempo das gravações, e maiores são os custos da validação dos dados.

O trabalho descreve o desenvolvimento de três bancos de fala para o Português Brasileiro (dois para reconhecimento e um para a síntese de fala), dicionário fonético das palavras coletadas com múltiplas transcrições regionais e sentenças gravadas.

2. Seleção das Frases compostas do Banco de Fala

O primeiro banco gravado, denominado ASR-DB1, é constituído de 200 sentenças definidas no projeto NURC (Alcaim 1992), sendo o mesmo utilizado em outros sistemas de reconhecimento de fala (Ynoguti 1999). Com o objetivo de aumentar a abrangência do conjunto de sentenças, 24 outras foram acrescentadas. Este conjunto adicional cobre números cardinais e ordinais, direções, comandos, meses do ano e nomes do zodíaco. Alguns números estavam contidos nas duzentas frases do projeto NURC, mas foram repetidos para melhorar o treinamento dos modelos acústicos. As outras palavras representam aplicações usuais de reconhecimento de fala. As mais de 700 palavras únicas contidas no conjunto de frases são também as mais usualmente verificadas nos *links* de páginas da Internet brasileira, objetivo principal do banco.

Com o objetivo de aumentar a cobertura fonética, foi desenvolvido um algoritmo de seleção das sentenças que constituem o banco, de modo a ter a maior cobertura possível com a menor quantidade de sentenças. Em (Rebollo et al 2005) foi desenvolvido um sistema de seleção de sentenças que divide o corpus CETENFolha (Corpus de Extractos de Textos Electrônicos NILC/Folha de São Paulo) em 400 conjuntos de 1000 sentenças cada e o conjunto selecionado é o que possui a maior quantidade de fonemas distintos. O algoritmo desenvolvido neste trabalho seleciona a sentença do corpus que possui a maior quantidade de fonemas distintos que ainda não foram encontrados. Portanto, este conjunto de fonemas que ainda não foram selecionados varia cada vez que uma nova frase é selecionada. Aí reside a complexidade do algoritmo. O processo de seleção das sentenças consiste em:

- Coleta de um conjunto de sentenças corretamente transcritas;
- Separação dos textos em sentenças;
- Produção de um dicionário fonético com todas as palavras contidas no
- conjunto de sentenças usadas no sistema;
- Expansão fonética das frases em "*cross-word*" para determinar o
- conjunto de fonemas presentes no conjunto de sentenças;
- Seleção das sentenças até que o sistema cubra todos os fonemas.

Bons linguistas podem demorar até um ano para obter um conjunto de sentenças que cubra os fonemas de uma língua, mas não têm como determinar qual o conjunto de fonemas mais usualmente pronunciados nesta língua. Com o objetivo de determinar primeiramente qual é este conjunto, um banco com sete milhões de sentenças foram coletadas. Cada palavra dessas sentenças foi checada frente a um dicionário ortográfico e 3,9 milhões de sentenças, corretamente transcritas, foram selecionadas. O sistema leva 55 minutos para processar este conjunto de sentenças e selecionar qual o menor conjunto de sentenças que descreve todos os fonemas com um determinado número mínimo de ocorrências no conjunto de sentenças processadas. O número mínimo de fonemas presente no conjunto de sentenças é determinado pelo usuário.

O algoritmo foi adaptado para selecionar sentenças a serem usadas em aplicações de adaptação ao locutor usando *lattice based MLLR* e *discriminative linear transforms* (Uebel 2001, 2002). As duas técnicas aumentam o nível de reconhecimento em condições difíceis (ruído, descasamento entre áudio do treinamento e teste). Neste caso, os fonemas são agrupados e contados como se fossem um único.

2.1 Algoritmo de Seleção de Sentenças

Primeiramente, o algoritmo procura os fonemas com o mínimo de ocorrência definido pelo usuário. O segundo passo é descobrir qual sentença possui a maior quantidade de fonemas procurados. Os fonemas encontrados são retirados da lista de fonemas a serem procurados e uma nova sentença que contém a maior quantidade de fonemas procurados é selecionada. A pesquisa é concluída quando não existem mais fonemas a serem procurados. O algoritmo acaba selecionando fonemas não procurados, uma vez que uma sentença irá ser constituída de fonemas procurados e não procurados.

Com a velocidade que o algoritmo é executado, o linguista pode redefinir uma

seleção de sentenças baseadas na quantidade de fonemas a serem cobertos pela quantidade de sentenças, e avaliar o tempo de gravação e os custos envolvidos.

2.2. Sentenças usadas na Seleção

Foram 3.946.744 sentenças extraídas de jornais, revistas, notícias e livros encontrados na Internet. O conjunto de sentenças é constituído de 44.518.978 palavras e 213.654 palavras únicas (sem repetição). As palavras mais freqüentes no banco são: "de" (4,24%), "o" (3,27%), "a" (3,25%), "e" (2,54%), "do" (2,15%), "que" (2,10%), "da" (1,81%), "em" (1,16%), "para" (1,12%) e "um" (0,93%).

Da expansão fonética do tipo "*word internal*", resultaram em 211.927.983 fonemas (trifones, bifones e monofones) e 23.107 fonemas únicos, e na expansão do tipo "*cross word*" foram 36.373 fonemas únicos. Considerando-se que podem ocorrer erros na geração do dicionário fonético, um corte no número mínimo de ocorrências de um fonema é necessário. Outro parâmetro importante é o número de fonemas por sentença. Sentenças muito longas são difíceis de serem lidas e resultam em muitos erros de leitura e, consequentemente, aumento no tempo e custos de validação.

O número mínimo de ocorrências de um fonema influencia no número de sentenças selecionadas, já que uma quantidade maior de fonemas deverá ser coberta nas sentenças selecionadas. A figura abaixo apresenta a cobertura fonética em função da quantidade máxima de fonemas em uma sentença selecionada dado pelo número de corte (CORTE). Na figura são apresentados cortes entre 50 fonemas e 150 fonemas por sentença no máximo e 28 fonemas no mínimo. Foram 781 simulações de aproximadamente 55 minutos cada (716 horas no total).

2.3. Considerações sobre as Sentenças Selecionadas

O processo de seleção de um determinado conjunto de sentenças para gravar um banco de fala depende, não somente de características técnicas de uma determinada língua (fonemas mais usualmente pronunciados), mas também de aspectos mais subjetivos como a facilidade em pronunciar uma determinada sentença, menor quantidade de palavras pronunciadas, e menor quantidade de palavras de origem estrangeira. O processo de seleção do conjunto de sentenças levou em consideração o seguinte:

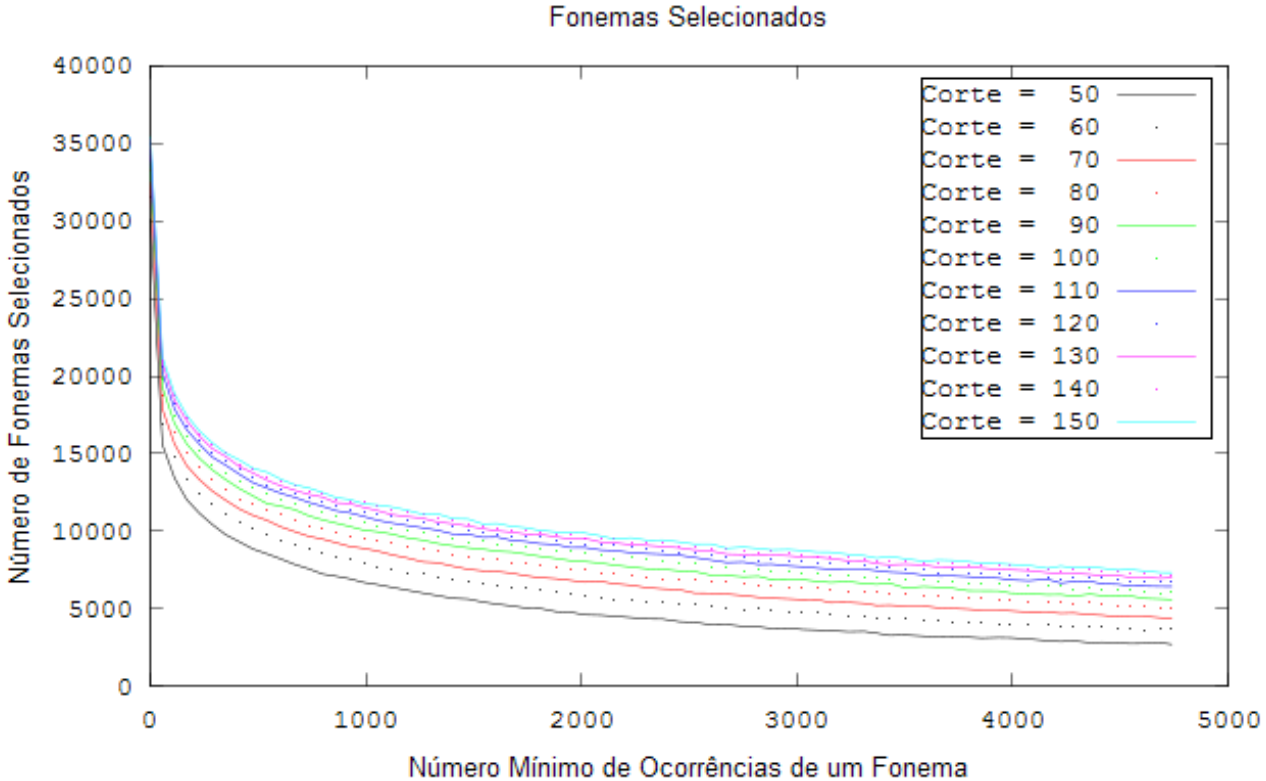

Figura 1. Número de Fonemas nas Sentenças Selecionadas.

1. Cobertura fonética de no mínimo 93,5 % de fonemas presentes na língua Portuguesa, sendo que esta cobertura foi verificada em função da distribuição dos fonemas presentes no banco de fala (3.946.744 de sentenças, 211.927.983 fonemas e 36.373 distintos);
2. O conjunto de sentenças selecionadas possui, no mínimo, os sete mil fonemas mais frequentemente encontrados na língua Portuguesa. O conjunto selecionado possui 353 sentenças, 4.175 palavras, e 7.216 fonemas distintos, que representam 93,8 % do total de fonemas presentes no banco de dados.

Portanto, em função da cobertura fonética, o conjunto de sentenças pode ser utilizado para o treinamento de modelos acústicos de excelente qualidade, com uma cobertura fonética excepcional e que podem ser utilizados no reconhecimento de fala contínua e espontânea, uma vez que as sentenças foram selecionadas utilizando-se uma expansão fonética do tipo "*cross word*".

Outras sentenças e palavras avulsas foram acrescentadas ao conjunto de sentenças para melhorar o nível de reconhecimento em situações difíceis, como é o caso dos números, comandos usados para comandar o Navegador Internet e outras tarefas de comando de um modo geral. Os números acima mencionados não levam em consideração estas sentenças e palavras avulsas.

As palavras que constam das sentenças gravadas no banco de fala representam 65,48% de todas as palavras do banco de dados, ou seja, as palavras que foram gravadas representam mais de 2/3 de todas as palavras que constam do banco de dados. Este é mais um dado que mostra quão acertada foi a seleção do conjunto de sentenças.

2.4. Sentenças do ASR-DB2

As primeiras cinco sentenças do conjunto de 353 encontram-se na Tabela 1. Além das 353 sentenças, foram acrescentadas 23 sentenças curtas para aumentar a cobertura fonética, 6 sentenças para adaptação ao locutor (aproximadamente 30 segundos de fala), 19 sentenças de domínio específico (similares as usadas no ASR-DB1), e 49 comandos de controle de um Navegador Internet.

Tabela 1. Lista de Sentenças Foneticamente Balanceadas

	Texto
1	O Corinthians passou para a segunda fase ao desclassificar o Vitória da Bahia.
2	Quem comprou telefones em noventa e três acabou o ano com uma boa surpresa.
3	Esse consumidor está em busca de um produto cada vez melhor e mais barato.
4	Os freios a disco nas quatro rodas podem ter ABS opcional, não travam.
5	Depoimento garante retorno ao gabinete civil, da sucursal de Brasília.

Como algumas aplicações de reconhecimento de fala demandam o reconhecimento de palavras isoladas, foram acrescentadas 45 palavras avulsas relacionadas a números, 28 de letras do alfabeto, 46 de direção e comando, 18 de partes do corpo humano, 36 de utensílios domésticos, e 42 de sensores, cores e estados emocionais. O total de sentenças é de 665, que agrupam a maioria das aplicações em reconhecimento de fala (fala espontânea, ditado, comandos, algarismos e robótica).

2.5. Sentenças para a Síntese de Fala

Diversos mecanismos podem ser utilizados para a elaboração da síntese de fala, como, por exemplo, a concatenação de sentenças ou palavras. O conjunto de sentenças deve possuir mais de uma mostra do mesmo fonema para melhor modelamento acústico.

Desta forma os fonemas possuem, no mínimo, duas ocorrências no conjunto de sentenças selecionadas e, pelo menos, 1.266 ocorrências no banco. O resultado são 632 sentenças foneticamente balanceadas para o Português Brasileiro, 8.046 fonemas (94,6% dos fonemas existentes no banco) e 7.529 palavras. Foram acrescentadas as sentenças para o reconhecimento de fala, 30 nomes próprios mais comumente encontrados no Brasil e outras palavras avulsas. O total de sentenças avulsas é de 558, sendo 47 sentenças para adaptação ao locutor (transformação da voz), e o total de sentenças para o banco de fala em síntese de fala (TTS-DB1) é de 1.220. Comparando a outros bancos, o CMU Artic (Kominek 2004), conjunto de sentenças para o

inglês americano, possui 1.132 sentenças, 10.003 palavras, sendo 2.970 palavras únicas, o que mostra uma grande repetição de palavras nas sentenças. O SynthFemale01 possui 4,392 sentenças em coreano e o projeto NEMLAR possui 2.032 sentenças (42 mil palavras) em árabe.

Comparado a outros bancos citados, o conjunto aqui selecionado possui menos sentenças, facilitando a leitura, maior cobertura fonética e menor repetição de palavras, resultando em 16 vozes sintetizadas usando o Flite+HTS (Black 2007).

3. Dicionário Fonético

O dicionário fonético é constituído pela transcrição ao nível fonético das palavras contidas nos bancos de fala e outras palavras que se queira reconhecer. Os fonemas são a menor unidade de som de uma língua e permitem que seja possível reconhecer palavras não contidas nestes bancos. De um modo simplificado, o "ão" de "tubarão" pode ser utilizado no treinamento de modelos que irão reconhecer "coração" ou "leão". O mesmo processo pode ser usado na síntese de fala.

3.1. Corpus e Transcrição Fonética

Todas as palavras que constituem as sentenças dos bancos de fala foram transcritas foneticamente segundo as variantes padrão (norma oficial) e não-padrão (em decorrência das diferentes regiões geográficas, classes sociais, faixas etárias, etc.) do Português Brasileiro, perfazendo um total de 13 grandes variações dialetais descritas: DF, GO, ES, AM, RJ, SP, interior de SP, MT, MG, PA, BA, Sul e Nordeste em geral.

Segundo pesquisas realizadas (Callou 2003), não podemos tomar como modelo apenas a pronúncia de uma pessoa, de uma classe social ou de uma região. Por isso, apresentamos em nosso dicionário múltiplas transcrições fonéticas das variantes e de vícios de linguagem (alterações sonoras como assimilação, epêntese, etc.), uma vez que a língua não é usada da mesma forma pelas pessoas em todos os momentos. Foram transcritas todas as palavras contidas nos três bancos de fala (ASR-DB1, ASR-DB2 e TTS-DB1) e o sistema de notação adotado para tal foi o SAMPA, alfabeto fonético computável, constituído pelo mapeamento dos símbolos do sistema IPA (Alfabeto Fonético Internacional) para códigos ASCII.

A correta escolha do conjunto de sentenças utilizado para a gravação de um banco de fala é de vital importância para o projeto. Tal importância não se restringe unicamente à cobertura fonética representativa dos sons da língua, mas também à facilidade com que os locutores as leem. Outro item de suma importância deste projeto é a construção de um dicionário fonético de qualidade, pilar para que se obtenham bons resultados nas tarefas relacionadas ao reconhecimento e síntese de fala.

3.2. Coleta do Banco de Fala

A coleta do banco de fala com vários locutores lendo as mesmas sentenças previamente escolhidas em função de sua contribuição para a cobertura fonética e para possíveis aplicações possui vantagens como: facilidade na coleta dos dados, uma vez que o conjunto de sentenças não é modificado; rapidez na validação do banco de fala, pois pode ser usada uma validação vertical, ou seja, validar cada sentença para todos os locutores ao invés de todas as sentenças de um locutor, o que possibilita ganhos de produtividade em torno de cinco vezes; e menor taxa de erros por parte do locutor e do validador. Como descrito anteriormente, este tipo de banco de dados permite definir, no menor conjunto de sentenças, todos os trifones mais encontrados na língua.

O perfil dos locutores escolhidos varia entre 13 e 59 anos, devido às aplicações vislumbradas pelos bancos de fala. As aplicações são de reconhecimento de fala espontânea, ditado, comandos de controle de robôs móveis e controle de objetos.

As gravações do banco ASR-DB1 foram realizadas no estado de São Paulo, e o ASR-DB2 no estado de São Paulo e outros estados da Federação. Muitas pessoas gravadas no estado de São Paulo são oriundas de outras regiões, possibilitando uma cobertura de sotaques das 13 regiões dialéticas do Brasil. O tempo médio das gravações do ASR-DB1 variava entre 20 minutos e 2 horas (média de 25 minutos) e

entre 1:10h e 1:30h para o ASR-DB2. Algumas pessoas podem levar até 3 horas, ou menos, como 55 minutos. O tempo varia conforma a velocidade de leitura do locutor, facilidade de leitura, nervosismo, timidez, e problemas de coordenação motora ao utilizar os equipamentos (computador e *mouse*). Algumas pessoas ficam nervosas por acharem que estão sendo testadas. Neste caso o técnico deve orientar o locutor e acalmá-lo para que a gravação tenha prosseguimento. Ruídos causados pelos lábios ao serem abertos no início das gravações e o ruído causado pela respiração resultam em picos nos respectivos espectrogramas e problemas na segmentação dos dados no treinamento acústico. Neste caso o técnico pedia ao locutor para regravar a sentença novamente. O locutor deixava um silêncio antes e depois de pronunciar cada sentença para que a pronúncia da mesma não fosse cortada.

As frases mais problemáticas de serem lidas são as que possuem palavras estrangeiras e excertos literários, por possuírem palavras desconhecidas pelo locutor, este não entende o sentido da frase e acaba cometendo erros de pronúncia e entonação.

3.3. Validação dos Dados

O processo de coleta dos dados é seguido de sua validação, ou seja, verificar se o locutor realmente leu o que deveria ler. O processo é iniciado na coleta dos dados. O técnico fica atento ao que o locutor está pronunciando e, caso não esteja correto, pede que o mesmo repita a leitura da sentença de uma forma correta. Este procedimento tem por finalidade diminuir o tempo e custos do processo de validação do banco de fala.

A validação consiste em escutar todos os arquivos de áudio e transcrever fielmente o que o locutor realmente pronunciou. A transcrição fiel dos arquivos sonoros deve ser a mais verossímil possível. Caso o locutor tenha pronunciado "desta" em vez de "dessa" ou pronunciou "pobema" em vez de "problema", estes fatos devem constar nos arquivos de texto validados. Um dos aspectos que facilitaram e aceleraram o processo de validação dos dados foi utilizar uma validação vertical dos dados, ou seja, frase por frase, em oposição ao método tradicional que valida locutor por locutor. O treinamento do técnico responsável pelas gravações reduziu significativamente o tempo de validação dos arquivos e o processo de validação vertical diminuir os erros na validação uma vez que o validador vai escutar centenas de vezes a mesma sentença e detectar de uma forma mais rápida e precisa pequenos erros como a falta da pronúncia de um simples "s" em uma palavra ou os erros descritos acima.

A validação é peça fundamental no treinamento de modelos acústicos no reconhecimento e síntese de fala, principalmente quando técnicas mais avançadas de modelamento acústico são empregadas, como treinamento discriminativo. Uma validação automática foi empregada para detectar arquivos corretamente pronunciados, facilitando (acelerando) o processo de validação e diminuindo seus custos.

4. Conclusões

O trabalho apresenta o desenvolvimento de três bancos de fala para o Português Brasileiro, dois especificamente para reconhecimento e um para síntese. O ASR-DB1 possui 248 locutores, 224 sentenças, microfone de 100 a 10 kHz, 16 *bits* e 48 kHz de taxa de amostragem. O ASR-DB2 irá gravar 1.226 locutores (550 já gravados), 665 sentenças, microfone de alta fidelidade do tipo *headset* (banda passante de 80 Hz a 15 kHz) e placa de som com 24 *bits* e 96 kHz de taxa de amostragem. O TTS-DB1 possui 1.220 sentenças, microfone de 20 Hz a 20 kHz, 24 *bits* e 96 kHz de amostragem. O trabalho também apresentou um algoritmo de seleção de sentenças que permite determinar com exatidão os fonemas mais usuais do Português Brasileiro usando um conjunto de 3,9 milhões de sentenças ortograficamente corrigidas.

5. Agradecimentos

Agradecemos o apoio da FAPESP (PIPE – 2005/59953-0), Finep (Subvenção Econômica à Inovação – 4717/06) e CNPq (RHAE Pesquisador na Empresa – 558052/2008-8) pelos recursos destinados ao projeto.

Agradecemos o pesquisador Mauro Miazaki pela implementação da primeira versão do sistema de seleção de sentenças, Henrique Ferraz, Vinicius Mittitier, Franclin Barros e

Roseli Oliveira Silva pelas gravações do banco ASR-DB2.

6. Referências

Alcaim, Abraham, Solemicz, José Alberto e Morais, João Antônio de. 1992. Frequência de Ocorrência dos Fones e Listas de Frases Foneticamente Balanceadas para o Português Falado no Rio de Janeiro. *Revista da Sociedade Brasileira de Telecomunicações*, Vol. 7.

Black, Alan, Zen, Heiga e Tokuda, Keiichi. 2007. Statistical Parametric Speech Synthesis. *ICASSP 2007*, Honolulu, Havaí, EUA: 1229-1232.

Callou, Dinah & Leite, Ione. 2003. *Iniciação à Fonética e à Fonologia*, 5 ed., Rio de Janeiro: Zahar.

Rebollo Couto, Leticia; Moraes, J.A. de, Resende Jr, F.G.V.; Cirigliano, R.J. R.; Barbosa, F.L.F.; Vianna, C.M. 2005. Um Conjunto de 1000 Frases Foneticamente Balanceadas para o Português Brasileiro Obtido Utilizando a Abordagem de Algoritmos Genéticos, *Anais do XXII Simposio Brasileiro de Telecomunicacoes (SbrT 2005)*, Campinas, Brazil, pp. 544-549.

Cohen, Michael H.; Giangola, James. P.; Balogh, Jennifer. 2004. *Voice User Interface Design.* Addison-Wesley Professional, 368 p.

Kominek, John & Black, Alan. W. 2004. The CMU Artic Speech Databases. Proceedings of *5th ISCA Speech Synthesis Workshop*, Pittsburgh, EUA, pp. 223-224.

Uebel, Luis Felipe & Woodland, P. C. 2001. Improvements in Linear Transform based Speaker Adaptation. *In: IEEE – International Conference on Acoustics Speech and Signal Processing*, 7-11 May 2011, Salt Lake City, Utah, EUA, Vol. 1, pp. 49-52.

Uebel, Luis Felipe. 2002. *Speaker Normalisation and Adaptation in Large Vocabulary Speech Recognition*. Tese de Doutorado. University of Cambridge. England.

Ynoguti, Carlos Alberto. 1999. Reconhecimento de Fala Contínuo usando Modelos Ocultos de Markov. Tese de Doutorado, UNICAMP, Campinas-SP.