Metodologias para Projeto e Aquisição de uma Base de Dados Lingüísticos Visando ao Treinamento e à Avaliação de Sistemas de Reconhecimento de Fala

Edmilson Morais[1], Jussara M. Viera[2], Pablo Arantes[2], Ana Cristina F. Matte[3]

[1] FEEC - Faculdade de Engenharia Elétrica e Computação, UNICAMP
[2] IEL - Instituto de Estudos da Linguagem, UNICAMP
[3] FALE - POSLIN - Estrutura Sonora da Linguagem, UFMG

{emorais}@decom.fee.unicamp.br

Abstract. *The aim of this work is to describe a methodology for designing and recording linguistic databases for training and evaluation of speech recognition systems. All the methods presented on this paper were specifically developed for Hidden Markov Model based speech recognition systems. Moreover, the techniques and recommendations for database design and recording presented here are specific for speech recognition applications such as embedded systems for mobile phones, Palm-Top, Toys, audio and video equipments and information kiosks.*

Resumo. *O objetivo deste trabalho é descrever uma metodologia para projeto e aquisição de bases de dados lingüísticos, voltadas ao treinamento e avaliação de sistemas de reconhecimento automático de fala. Todas as técnicas para projeto de bases de fala descritas neste artigo serão voltadas para sistemas de reconhecimento de fala baseados na tecnologia de Modelos Ocultos de Markov e para tarefas específicas de reconhecimento de fala, tais como: sistemas embarcados para telefonia móvel, Palm-Top, brinquedos, produtos eletroeletrônicos (áudio e vídeo), portais de voz e quiosques para informações.*

1. Introdução

Todos os sistemas modernos de reconhecimento automático de fala são baseados em métodos estatísticos tais como HMM (Hidden Markov Models) e ANN (Artificial Neural Networks). Todos estes métodos demandam, em geral, uma grande massa de dados lingüísticos para que sejam treinados e avaliados adequadamente. Além disso, muitos dos algoritmos utilizados nos sistemas de reconhecimento de fala são dependentes de aspectos lingüísticos e, portanto, requerem pesquisas e desenvolvimentos especificamente direcionados para a língua abordada. Conscientes deste fato, alguns projetos e associações foram criados na Europa e nos Estados Unidos com o objetivo de projetar, coletar e distribuir bases de dados lingüísticos para várias das línguas faladas no mundo [1, 8, 9].

Apesar de algumas iniciativas Européias e Norte Americanas para construção de bases de dados lingüísticos datarem do início da década de 90, até o presente momento, os autores deste projeto desconhecem a existência de uma base de dados lingüísticos

sobre o português brasileiro, doravante PB, que tenha sido especificamente projetada para motivar pesquisas e desenvolvimentos na área de reconhecimento automático de fala no Brasil e que seja de domínio público. Em outras palavras, os autores deste projeto desconhecem a existência de uma base de dados voltada para o reconhecimento de fala do PB, que seja de larga extensão, que tenha sido devidamente projetada, adquirida e rotulada, e que esteja disponível gratuitamente para Universidades e empresas de base tecnológica.

A solução até então adotada por muitos grupos de pesquisa no Brasil tem sido a construção de bases de dados locais e de uso particular. Alunos de Mestrado e Doutorado que trabalham com reconhecimento de fala têm despendido um tempo enorme no desenvolvimento de bases que, em geral, não são construídas de maneira apropriada e que, além disso, não possuem a extensão suficiente para validar os novos métodos, técnicas ou algoritmos propostos[1]. Outro ponto extremamente importante associado à ausência de uma base de dados lingüísticos, comum a vários grupos de pesquisa, é a impossibilidade de uma comparação fidedigna dos resultados experimentais obtidos entre os grupos.

Nos últimos dez anos, inúmeros grupos de pesquisa e empresas de base tecnológica têm sido criados na Europa, Estados Unidos e Japão [3] visando ao desenvolvimento de sistema para reconhecimento automático de fala. As seis áreas mais focadas para possíveis aplicações são: (1) Telefonia móvel, (2) Sistemas de informação – Portais de voz ou quiosques de informação, (3) Dispositivos de áudio e vídeo, (4) Dispositivos automotivos, (5) Brinquedos e (6) "Palm-Top". A existência de bases de dados lingüísticos para o Inglês, para o Japonês e para várias outras línguas Européias, têm sido de fundamental importância para o sucesso de tais grupos e empresas.

Motivados pela enorme importância que uma base de dados lingüísticos, de alta qualidade e de domínio público terá no desenvolvimento da área de reconhecimento de fala no Brasil, os autores deste trabalho vêm por meio deste propor uma metodologia para a construção de tal base de dados.

A Seção 2 deste artigo descreve em detalhes a metodologia proposta para projeto e aquisição da base de dados. Nesta seção são apresentados detalhes sobre a escolha das aplicações-alvo, seleção dos locutores, projeto do *corpus*, gravação e etiquetagem das sentenças, análise das gravações, licença de uso, documentação e disponibilização do material. A Seção 3 conclui este trabalho apresentando algumas considerações finais.

2. Metodologia

Algumas das principais etapas na criação de uma base de dados lingüísticos, voltada ao treinamento e avaliação de sistemas de reconhecimento de fala, são:

- Definição das prováveis aplicações-alvo
- Seleção dos locutores
- Seleção das sentenças a serem gravadas
- Aquisição
- Segmentação e etiquetagem

[1] Infelizmente, muitas defesas de Tese na área de Reconhecimento Automático de Fala sobre o português brasileiro, defendidas no Brasil, terminam com a velha retórica: "Não havia dados suficientes para validar o algoritmo proposto".

- Análise da qualidade acústica das gravações
- Avaliação da qualidade vocal dos locutores
- Licença de uso
- Documentação e disponibilização da base de dados

2.1. Definição das Prováveis Aplicações-Alvo

Treinar e avaliar sistemas de reconhecimento de fala capazes de operar com alto desempenho em qualquer tarefa e/ou ambiente acústico é um desafio que o estado-da-arte da tecnologia de fala ainda não é capaz de atingir. A maneira mais simples e comumente utilizada para contornar este problema é desenvolver sistemas específicos para determinadas aplicações e/ou condições acústicas. O desenvolvimento de tais sistemas dependentes de tarefa demanda tecnologias específicas e bases de dados lingüísticos especialmente projetadas para tais fins. Partindo-se de tal premissa, toda a metodologia apresentada neste artigo foi desenvolvida visando a aplicações de reconhecimento de fala especificamente voltadas para um número reduzido de aplicações-alvo:

- Sistemas embarcados para telefonia móvel
- "*Palm-Tops*"
- Brinquedos
- Produtos eletroeletrônicos (eletrodomésticos, sistemas de áudio e vídeo...)
- Portais de voz
- Quiosques para informações

A motivação para a escolha de tais aplicações-alvo foi a realização de alguns estudos europeus indicando o potencial econômico de tais áreas [2].

2.2. Seleção dos locutores

Recomenda-se que todos os candidatos a locutores sejam submetidos a um protocolo de entrevista e avaliação. Este protocolo deve ser aplicado por especialistas em Lingüística e Fonoaudiologia e deve possuir os seguintes itens:

- *Identificação.* Aqui serão registradas informações, tais como: nome, idade, data de nascimento, sexo, cidade em que viveu na maior parte da infância e da adolescência e a naturalidade dos pais. Outro item importante a ser registrado é o uso ou não de algum tipo de aparelho ortodôntico ou prótese dentária.
- *Caracterização da saúde vocal*: Aqui será realizada uma caracterização de hábitos como fumo e ingestão de bebida alcoólica. Uso de medicamentos (o que inclui, por exemplo, anticoncepcional ou qualquer outro hormônio para mulheres). Estado das vias aéreas superiores (laringe, faringe, nariz, especialmente). Queixas vocais e auditivas. Possíveis alterações vocais (em mulheres) associadas ao ciclo menstrual.
- *Avaliação de aspectos vocais e de produção da fala.* Aqui serão avaliados aspectos tais como: Ritmo, intensidade, qualidade vocal, níveis de inteligibilidade, ressonância e articulação.

Não serão aceitos locutores que apresentarem as seguintes características:

- Omissões, substituições, adições e transposições articulatórias, mesmo que se constituam por razões sociolingüísticas

- Movimentos de mandíbula que confiram uma articulação naturalmente travada ou exagerada
- Excesso de pressão aérea na produção de fonemas plosivos (que possam levar a variações demasiadamente bruscas de amplitude)
- Uso profissional da voz, por exemplo, locutores de rádio (por apresentarem uma fala significativamente diferente da fala de um locutor padrão)

As seguintes características serão toleradas nos locutores:

- Protrusão de língua nos fonemas /t/, /d/, /n/, /s/ e /z/
- Regionalismos quanto aos fonemas /r/, /t/, /d/, /s/ /ʃ/
- Graus leves de ressonância vocal nasal, rouquidão e soprosidade
- Uso de aparelhos ortodônticos e próteses dentárias desde que a produção de fala e voz satisfaça o nível de qualidade desejado

O treinamento de sistemas de reconhecimento estatístico de fala demanda bases de dados lingüísticos ricas em variabilidades acústicas. A forma mais usual de se obter tais variabilidades é por meio de um número elevado de locutores, com características dialetais diversas e com idades variadas. As Tabelas 1 e 2 a seguir apresentam sugestões quanto ao número total de locutores e suas respectivas faixas etárias e distribuições geográficas e dialetais. A Tabela 2 leva em consideração tanto a diversidade de dialetos quanto a importância econômica da região.

Tabela 1: Faixa etária dos locutores

Faixa etária	N° de locutores	N° de homens	N° de mulheres
De 18 a 30 anos	550	225	225
De 31 a 45 anos	300	150	150
De 46 a 60 anos	150	75	75
Total	1000	500	500

Tabela 2: Distribuição dialetal dos locutores

Região/Estado	Dialetos	População nacional	Número de locutores
Sul	Paranaense, Catarinense, Gaúcho	15%	200
São Paulo	Região metropolitana, Litonarâneo, Centro paulista, Oeste paulista	23%	225
Sudeste	Carioca, Mineiro, Capixaba	21%	225
Nordeste	Baiano, Pernambucano, Cearense	25%	275
Norte e Centro Oeste	Centro Oeste, Amazonense	16%	75
Total		100%	1000

2.3. Seleção das sentenças a serem gravadas

As sentenças a serem gravadas devem ser definidas em função das aplicações-alvo. Aspectos importantes a serem considerados durante o processo de construção ou seleção destas sentenças são:

- *As sentenças devem ser lidas ou pronunciadas espontaneamente?* Sentenças lidas são adequadas ao treinamento de sistemas de reconhecimento de fala bem articulada e pronunciada sem hesitações. Sistemas visando ao reconhecimento de fala espontânea devem ser treinados com o uso de bases de fala espontânea.
- *Quais as variabilidades fonético-acústicas que realmente são importantes no treinamento e avaliação dos sistemas?* As variabilidades espectrais limitadas aos

segmentos fonéticos da fala (fones, difones e trifones nos mais variados contextos) são as características mais relevantes para o bom treinamento de sistemas de reconhecimento baseados em HMMs. Aspectos prosódicos ou supra-segmentais não são, em geral, bem explorados pelos sistemas baseados em HMMs e, portanto, sua presença na base de dados não é muito relevante.

- *É ou não importante incluir sentenças ou palavras específicas para as aplicações-alvo?* Como os sistemas baseados em HMMs empregram métodos estatísticos que aprendem a partir dos exemplos de treinamento, é de se esperar que treinar o sistema com palavras que apresentem uma alta probabilidade de ocorrência durante o uso do sistema, irá provavelmente aumentar o desempenho do mesmo. Entretanto, esta inclusão de palavras específicas deve ser realizada com cuidado para evitar grandes alterações no balanceamento fonético-acústico do *corpus*.
- *Como construir ou selecionar um conjunto ótimo de sentenças?* Uma das técnicas mais usuais é a de selecionar, a partir de um grande *corpus* (por exemplo, sentenças extraídas do Jornal Folha de São Paulo), um subconjunto de sentenças que satisfaça as especificações fonético-acústicas consideradas mais relevantes. Além das sentenças selecionadas, é prática comum, como citado acima, a inclusão de sentenças e/ou palavras específicas para as aplicações-alvo que se deseja contemplar.

A Tabela 3 apresenta uma sugestão para as sentenças, palavras e comandos a serem gravados. É importante enfatizar mais uma vez que os itens da Tabela 3 foram definidos em função das aplicações-alvo citadas na Seção 2.1.

Tabela 3: Itens a serem gravados

Num.	Itens	Quantidade
1	Sentenças lidas	85
2	Sentenças foneticamente compactas	5
3	Palavras foneticamente ricas	5
4	Palavras/frases específicas p/ as aplicações-alvo	150
5	Dígitos conectados	5
6	Dígitos contínuos	5
7	Números telefônicos	5
8	Horas do dia	5
9	E-mail e endereços html	5
10	Dinheiro	5
11	Nomes de cidades	5
12	Nomes próprios	5
13	Dias da semana, mês, ano e datas importantes	5
14	Caracteres especiais de computador	5
15	Palavras soletradas	5
Total		**300**

A seguir são traçadas algumas considerações sobre os itens da Tabela 3

2.3.1 Sentenças lidas

Sugere-se a gravação de 85 sentenças foneticamente ricas para cada um dos 1000 locutores. O objetivo é obter uma cobertura de todos os fones, bem como uma boa cobertura dos difones e trifones mais freqüentes do PB. O termo "sentenças foneticamente ricas" não será utilizado no sentido de uma distribuição de fonemas similar à distribuição "típica" do PB. O termo "sentenças foneticamente ricas" será utilizado neste artigo para expressar:

- Exemplos de treinamento suficientes para todos os fones, incluindo os fones mais raros.
- Boa cobertura dos difones e trifones mais freqüentes. É importante ressaltar a necessidade de se respeitar um bom balanceamento dialetal.
- Número mínimo de exemplos de um determinado fone, para toda a base de dados, igual a 1000. Esta imposição somente deve ser relaxada para o caso de fones considerados muito raros. Apenas 5% do total dos fones podem ser considerados muito raros.

Para que seja alcançada uma boa diversidade acústica no conjunto de sentenças lidas, recomenda-se:

- Não deve existir mais do que 5 exemplares idênticos de cada sentença em todo o *corpus*.
- Cada fonema deve ser pronunciado por pelo menos 95% dos locutores.

Um bom método para a seleção das sentenças pode ser encontrado no site, http://gps-tsc.upc.es/veu/personal/sesma/index.html

Com o objetivo de obter uma boa variabilidade de pronúncias e de contornos prosódicos recomenda-se:

- Sentenças de tamanho variados, entre 8 e 12 palavras
- 90% de frases declarativas, 5% de frases exclamativas, 5% de frases interrogativas

Todas as sentenças devem ser individualmente conferidas para verificar se não há nada semanticamente ofensivo ou inapropriado.

2.3.2 Sentenças foneticamente compactas

Devem ser gravadas 5 sentenças foneticamente compactas. Estas sentenças devem apresentar as seguintes propriedades:

- Larga variabilidade fonético-acústica
- Ser de fácil leitura, isto é, devem minimizar possíveis hesitações ou dificuldades de leitura por parte dos locutores

Estas sentenças foneticamente compactas devem ser comuns a todos os 1000 locutores. Estas sentenças devem ser segmentadas manualmente e utilizadas para o treinamento inicial do sistema.

2.3.3 Palavras foneticamente ricas

Devem ser gravadas 5 palavras com contextos fonéticos relativamente pouco freqüentes, "raros", na língua Portuguesa. Estas palavras devem ser utilizadas para tentar satisfazer a condição de 1000 exemplares de cada fone em todo o *corpus*.

2.3.4 Palavras e frases específicas para as aplicações-alvo

150 palavras/comandos e frases específicas para as aplicações-alvo. A Tabela 4 apresenta alguns possíveis exemplos para palavras/sentenças específicas:

Tabela 4: Exemplo de alguns comandos específicos para as aplicações-alvo

Classes de comandos	Exemplos de comandos específicos para aplicação
Comandos para ativar e desativar sistemas	Ligar, desligar, cancelar, senha, ok, sair...
Dispositivos	CD, DVD, PDA, MP3, microfone, vídeo cassete...
Conectividade	*Bluetooth*, rede, servidor, cliente, sincronizar...

Navegação em diretórios	Menu, diretório, lista, opções, detalhes...
Edição de texto	Copiar, colar, corrigir, ditar, adicionar, inserir...
Dispositivos de vídeo	Maximizar, limpar, zoom, brilho, contraste, cor...
Dispositivos de áudio	Volume, aumentar volume, grave, agudo...
Navegação na Internet	Internet, hyperlink, conectar, responder, enviar, urgente...
Funções para agendas eletrônicas	Calendário, agenda, apontamentos, contatos...
Lazer e diversão	Cinema, teatro, arte, cultura, moda, comédia...

2.3.5 Sequência de dígitos

- 5 dígitos isolados: Dígitos devem ser pronunciados com uma pausa entre eles. Por exemplo - Dois, três, nove, sete, um, zero...
- 5 dígitos conectados: Os dígitos devem ser pronunciados de forma contínua, sem pausas entre eles. Por exemplo - Cinco - quatro - dez - um - dois - seis...
- 5 dígitos contínuos: Por exemplo - Dois mil quinhentos e cinqüenta e dois

2.3.6 Números telefônicos

5 números de telefones. Escolher números que representem discagens locais, estaduais e internacionais.

2.3.7 Horas do dia

5 expressões de horas do dia. Por exemplo: Cinco horas da tarde. Dezessete horas...

2.3.8 E-mail e endereços html

5 descrições de *e-mails* e endereços *html*.

2.3.9 Expressões descrevendo quantidade de dinheiro

5 sentenças descrevendo dinheiro. Por exemplo: oito mil trezentos e quarenta reais

2.3.10 Nomes de cidades

5 nomes de ruas. Por exemplo: Rua Treze de Maio, Avenida Brasil... Contemplar nomes freqüentes.

2.3.11 Nomes próprios de pessoas

5 nomes próprios (incluindo nomes e sobrenomes). Por exemplo: João Pedro da Silva. Deve-se contemplar nomes próprios freqüentes no Brasil.

2.3.12 Dias da semana, meses, datas importantes e feriados

5 expressões de datas. Por exemplo: Segunda-feira, Março, 21 de Abril, Natal...

2.3.13 Caracteres especiais do teclado do computador

5 caracteres especiais de teclado de computador. Por exemplo: Arroba, Cifrão...

2.3.14 Palavras soletradas

Soletrar palavras não é uma prática muito usual no PB (por se tratar de uma língua quase fonética) Entretanto, nos casos de alguns sobrenomes e nomes de cidades, o soletrado pode às vezes ser importante. São recomendadas 5 palavras soletradas: 2 nomes de pessoas, 2 nomes de cidades e 1 seqüência aleatória de letras.

2.4. Aquisição

2.4.1 Software para aquisição

Recomenda-se que as sentenças e/ou palavras/comandos sejam lidas da tela de um computador. O ideal seria a utilização um software para aquisição com as seguintes funcionalidades:

- Cadastro das informações do protocolo de entrevista e avaliação dos locutores, ver Seção 3.2.
- Condução do processo de gravação, indicando aos locutores o que deve ser pronunciado.
- Visualização gráfica do sinal de voz gravado e aviso sobre possíveis problemas de saturação ou nível muito baixo de sinal.

2.4.2 Cenário de gravação

Sugere-se uma gravação em ambiente silencioso, sujeito apenas a ruídos semelhantes ao de um escritório. A relação sinal/ruído (RSR) deve ser controlada na faixa entre 30 e 60dB, aproximadamente. Os autores deste artigo estão conscientes que algumas das aplicações-alvo sugeridas na Seção 3.1 estarão, muito provavelmente, sujeitas a RSR acima de 60dB. Portanto, talvez fosse mais adequado a realizações de gravações em diferentes cenários, sujeitos a RSR na faixa entre 30 e 90dB. Entretanto, a aquisição de uma base de dados em diferentes cenários e sujeita a elevados níveis de ruído é uma tarefa deveras complexa. Outra tarefa não menos complexa é a segmentação fonético-acústica de uma base de dados que tenha sido adquirida em ambientes com elevado nível ruído. O que os autores deste artigo sugerem é uma aquisição em ambiente de escritório e uma posterior mistura, aditiva ou convolutiva, do sinal gravado com ruídos diversos [6].

Com o objetivo de se obter diferentes relações sinal-ruído, para o sinal gravado, sugere-se a gravação, simultânea, de três canais:

- Gravação a curta distância: entre 3 a 5 cm dos lábios do locutor.
- Gravação a média distância: entre 30 e 40 cm dos lábios do locutor.
- Gravação a longa distância: entre 100 e 110 cm dos lábios do locutor.

2.4.3 Equipamentos de gravação

"*LapTops*" providos de placa de som digital externa de alta qualidade. Microfones do tipo "*headset*" de alta qualidade para realização das gravações a curta distância. Microfone de mesa localizado entre 30 e 40 cm do locutor para realização das gravações a média distância. Microfone de mesa localizado entre 100 e 110 cm do locutor para realização das gravações a longa distância.

2.4.4 Condições de gravação

Os três canais devem ser gravados, simultaneamente, com taxa de amostragem de 22kHz e quantizados com 16 bits.

2.5. Segmentação e etiquetagem

Depois de gravadas, todas as sentenças devem ser transcritas ortograficamente. Esta transcrição consiste na verificação do que realmente foi falado pelo locutor. Se houver alguma diferença entre as sentenças originais e o que foi falado pelo locutor então as devidas correções devem ser efetuadas.

Deve ser realizada a transcrição ortográfico-fonética de todas as palavras presentes no *corpus*. Para que isto seja feito, torna-se necessário a definição de um alfabeto fonético e também a construção de um transcritor ortográfico-fonético. A Tabela 5 apresenta uma proposta para o alfabeto fonético a ser utilizado.

Tabela 5: Proposta para o alfabeto fonético a ser utilizado: SAMPA-PB

SampaPB	IPA	Ex.
Vogais plenas		
a	a	s**a**co
E	E	s**e**co(verb)
O	ɔ	s**o**co(verb)
e	e	s**e**co(adj)
o	o	s**o**co
i	i	s**i**co
u	u	s**u**vo
Vogais reduzidas		
A	ɐ	sac**a**
I	I	saqu**e**
U	Y	sac**o**
&	ɐ̟	pêss**e**go
@	ɒ̟	cômodo
Consoante de ataque		
p	p	**p**ata
t	t	**t**ata
k	k	**c**ata
b	b	**b**ata

SampaPB	IPA	Ex.
d	d	**d**ata
g	g	**g**ata
f	f	**f**aca
s	s	**s**aca
S	ʃ	**ch**ata
v	o	**v**aca
z	z	**z**aca
Z	ʒ	**j**aca
r	r, χ, x	car**r**o
m	m	**m**ata
n	n	**n**ata
J	ŋ	**nh**oque
R	ɾ	**p**rato
l	l	ga**l**o
L	ʟ	ga**lh**o
Fones em coda		
N	n, ~	san**to**
5	s	casca
4	ɬ,x, ɾ	car**t**a

SampaPB	IPA	Ex.
j	j	cai**x**a
w	w	ma**u**
6	ɐ̯	pár**ia**
Ditongos nasais		
aNw	ɐ̃w̃	m**ão**
ANw	ɐ̃w̃	bênç**ão**
aNj	ɐ̃j̃	m**ãe**
eNj	ẽj̃	d**en**te
oNj	õj̃	p**õe**
Ditongos orais		
aj	aj	p**ai**
Ej	Ej	papé**is**
ej	ej	**ei**xo
oj	oj	p**ois**
Oj	ɔj	d**ói**
uj	uj	f**ui**
aw	aw	m**au**
Ew	ɛw	c**éu**
ew	ew	m**eu**

SampaPB	IPA	Ex.
iw	iw	r**iu**
Ow	ɔw	s**ol**
ow	ow	s**ou**
uw	uw	az**ul**
Iw	Iw	diár**io**
I@	Iɒ̟	Dár**io**
Uw	ʊw	cônsu**l**
U@	ʊɒ̟	côngr**uo**
Ij	Ij	sér**ie**
I&	Iɐ̟	cár**ie**
Uj	ʊj	tên**ue**
U&	ʊɐ̟	tên**ue**
I6	Iɐ̯	pár**ia**
U6	ʊɐ̯	contín**ua**
Alofonias importantes		
T	tʃ	**t**ime
D	dʒ	**d**ica

Deve ser realizada uma transcrição fonética larga (não muito detalhada) de todas as sentenças. Esta transcrição pode ser obtida através do uso de regras automáticas para a inserção de possíveis efeitos de co-articulação entre palavras. [5]

O último passo é a segmentação fonético-acústica das sentenças. A segmentação manual de todas as sentenças a serem gravadas seria uma tarefa extremamente tediosa e economicamente inviável. Por esta razão, a solução proposta pelos autores deste projeto é a construção de segmentador fonético-acústico semi-automático para o PB, baseado em HMM [7].

2.6. Análises

Avaliação acústica das gravações: Realização de algumas análises acústicas para verificar a qualidade sonora das gravações (Relação sinal ruído, nível de amplitude das gravações...).

Avaliação da qualidade vocal dos locutores: Realização de algumas análises para caracterizar a qualidade vocal dos locutores (Qualidade acústica das vogais, *Jitter*, *Shimmer,* taxa de locução de cada locutor...).

2.7. Licença de uso

Deve-se preparar um termo de licença de uso a ser assinado pelos locutores, disponibilizando os sinais gravados para os fins devidamente especificados.

2.8. Documentação e disponibilização da base de dados

Documentação descrevendo todos os ítens importantes relacionados com a definição das

aplicações-alvo, seleção dos locutores, seleção das sentenças, aquisição, análise, avaliação da base de dados, licenças e disponibilização do material.

Sugere-se a disponibilização da base de dados por meio do uso de CD, DVD ou de serviços de *ftp* ou *http*.

3. Considerações Finais

Neste trabalho foi apresentada uma metodologia para projeto e aquisição de bases de dados lingüísticos visando aos treinamentos e avaliações de sistemas de reconhecimento de fala. A metodologia foi desenvolvida tendo como ênfase algumas aplicações-alvo consideradas comercialmente interessantes. No que diz respeito à seleção dos locutores a serem gravados, foi proposto um protocolo de entrevista e avaliação, bem como distribuições dialetais e de faixa etária dos locutores. Foram discutidos aspectos fonético-acústicos considerados importantes no projeto das sentenças a serem gravadas. Uma descrição detalhada de um conjunto de itens a serem gravados (sentenças, palavras, comandos, soletrado, dígitos...) foi apresentada. Considerações sobre cenários, equipamentos e condições de gravação foram traçadas. Discussões sobre transcrições ortográficas, conversões ortográfico-fonética e segmentação fonética, semi-automática, foram apresentadas. Um alfabeto fonético foi proposto. Algumas análises acústicas e fonético-acústicas a serem realizadas após a aquisição da base de dados foram mencionadas. Finalmente, foram feitos alguns comentários sobre licença de uso, disponibilização da base de dados e documentação.

Os autores deste trabalho acreditam que uma base de dados lingüísticos para o PB, de alta qualidade e de domínio público, será de extrema importância para o desenvolvimento da área de reconhecimento de fala no Brasil. Esperamos que este trabalho possa, de alguma forma, contribuir para o projeto e aquisição de tal base de dados.

Referências

[1] LCD, Linguistic Data Consortium, http://www.ldc.upenn.edu/ . ELRA, European Language Resources Association, http:/www.elra.info/. ELDA, Evaluations and Language Resources Distribution Agency, http://www.elda.fr/sommaire.php

[2] SpeechDat projects., http://www.speechdat.org/. Speecon project., http://www.speechdat.org/speecon/index.html

[3] Empresas na área de Tecnologia da Fala., http://www.scansoft.com, http://www.nuance.com, http://www.research.att.com/programs/VES.html, http://www.research.att.com/programs/VES.html.

[4] Ynoguti C., A., Barbosa, P., A., and Violaro, F., "A Large Speech Database for Brazilian Portuguese Spoken Language Research", Proceedings of the VI Encontro para o Proc. Comp. da Língua Portuguesa, PROPOR'2003, Junho de 2003, Faro, Portugal. pp. 193-196, ISBN 3-540-40436-8.

[5] Albano, E. and A. Moreira, "Archisegment-based letter-to-phone conversion for concatenative speech synthesis in Portuguese", Proceeding of the ICSLP 96, vol 3, pp. 1708-1711, 1996.

[6] Couvreur, L., et al, "On the use of artificial reverberations for ASR in highly reverberant environments". 2° IEEE Benelux Signal Proc. Symposium, Hilvarenbeek, Holanda, Março, 2000.

[7] Wightman, C., W., Talkin, T. D., "The Aligner: Text-to-Speech Alignment Using Markov Models" , Progress in Speech Synthesis, Jan P. H. van Sante… [et al], editors, chapter 25, pp. 313, Spring-Verlang, New York, USA, 1996.

[8] Listerri, J., et al, "Corpus Orales para el Desarrollo de las Tecnologías Hable en Español" , Oralia Análisis del Discurso Oral 8, 2005 (em prensa). http://liceu.uab.es/~joaquim/publicacions/Oralia_04.pdf

[9] Listerri, J., "Transcripicón, Etiquetado y Codificación de Corpus Orales" , In Gómez Guinovart, J., et al (Eds.) Panorama de la Investigación en lingüística informática. RESLA, Revista Española de Lingüística Aplicada, Volumen Monográfico. p. 53-82.