1
Introdução

Esta tese tem como principal objetivo analisar e propor esquemas eficientes de reconhecimento de voz distribuído em redes IP (*Internet Protocol*) e redes de telefonia móvel celular, inovando na proposta de novas técnicas para melhorar o desempenho de reconhecimento nestas redes.

Neste capítulo, seção 1.1, será apresentada uma breve introdução da problemática envolvida no reconhecimento de voz em redes IP e redes de telefonia móvel celular. Na seção 1.2, será apresentada a Base de Voz construída e utilizada neste trabalho. Finalmente, a seção 1.3 descreve a estrutura dos capítulos e um breve resumo do conteúdo desta tese.

1.1.
Sistemas de Reconhecimento de Voz em Ambiente Celular/Redes IP

O desenvolvimento tecnológico do mundo atual tem estimulado demanda cada vez maior por máquinas inteligentes. Dentro desse panorama, a área de reconhecimento automático de voz (RAV) é uma das que têm despertado maior interesse, apesar da grande complexidade envolvida em termos de projeto e de operação. Esse interesse crescente tem sido evidente tanto no âmbito das indústrias como dos centros de pesquisa no mundo inteiro.

Tendo em vista o crescimento gigantesco da Internet e dos sistemas de comunicações móveis celulares, as aplicações de processamento de voz nesses meios têm despertado interesse cada vez maior. Em particular, um problema importante nessa área diz respeito ao reconhecimento de voz em um sistema servidor, a partir de parâmetros acústicos calculados e quantizados no terminal do usuário. O servidor reconhece a voz de acordo com uma aplicação específica e envia de volta, ao usuário, informações relativas à ação tomada a partir do reconhecimento de voz.

Os parâmetros acústicos podem ser os especificados por um *codec* de voz – caso o serviço de voz seja também utilizado – ou podem ser os vetores de

atributos que serão efetivamente empregados pelo reconhecedor de voz. Em qualquer situação, esses parâmetros serão digitalizados, através de um esquema de codificação de baixas taxas, e transmitidos ao servidor em canais de comunicações que usualmente apresentam limitação de faixa, como a Internet e os sistemas de telefonia móvel. É exatamente por essa limitação que os parâmetros devem utilizar esquemas de compressão que sejam eficientes. No nosso contexto essa eficiência é medida no sentido de servir bem ao propósito de reconhecimento e não de qualidade de voz. Esse seria um cenário típico de aplicação da tecnologia de reconhecimento de voz à Internet e aos canais de telefonia móvel celular.

Devido à alta complexidade computacional e à grande quantidade de memória requerida em sistemas de RAV, se torna muito atraente a opção por sistemas de reconhecimento de voz distribuídos. Em sistemas desse tipo, o processamento é distribuído entre o terminal do usuário (telefone celular, computador pessoal) e o terminal de recepção em uma rede de comunicações (estação base em redes de telefonia móvel, servidor central em redes IP).

Os problemas relacionados ao projeto de reconhecedores de voz distribuídos para operação na Internet e em redes de telefonia móvel são acentuados pelas altas taxas de erro de bits e perdas de pacotes, fora outros problemas usuais na concepção de sistemas de RAV, como o ruído ambiente.

Além disso, os esquemas de codificação de voz usados operam a baixas taxas de bits e utilizam, em geral, codificação preditiva linear (LPC – *Linear Predictive Coding*), com base em um modelo de produção da fala. Nesse modelo, um sinal de excitação é aplicado a um filtro só de pólos (caracterizado por parâmetros LPC), que representa a informação da envoltória espectral do sinal de voz. Usualmente os parâmetros LPC são transformados para LSF (*Line Spectral Frequencies*) ou ISF (*Immittance Spectral Frequencies*), devido às propriedades atraentes destes últimos para os processos de quantização e interpolação. No caso de sistemas de RAV distribuídos é preferível utilizar diretamente os parâmetros do *codec* do que extraí-los a partir do sinal decodificado (voz reconstruída no decodificador) [1]. A realização desse processamento envolve um grande número de aspectos e estratégias para concepção de reconhecedores de voz eficientes.

Para isso, conforme mencionado na caracterização do problema, diversos aspectos e estratégias deverão ser considerados. Primeiramente, os parâmetros LSF do *codec* não são necessariamente as melhores opções de atributos a serem

usadas no RAV. Portanto, transformações desses parâmetros são estratégias importantes a serem consideradas. Uma outra estratégia visada no projeto de reconhecedores de voz no ambiente celular e de redes IP consiste em incorporar outros parâmetros já disponíveis no decodificador, de modo a melhorar o desempenho do reconhecedor de voz. É de interesse, também, investigar atributos que sejam mais robustos em presença de ruído. Quando o serviço previsto é apenas o de reconhecimento, é importante examinar não só os novos conjuntos de atributos que sejam mais robustos, assim como novos métodos de codificação específicos para os atributos a serem empregados. Problemas relacionados ao comportamento do sistema em presença de erros no canal e de perdas de quadros, além da escolha do domínio de interpolação dos quadros, também são itens que devem ser examinados.

Dentre as diversas técnicas e problemas aqui apresentados, esta tese apresenta uma nova técnica de interpolação das LSFs que permite a obtenção dos atributos de reconhecimento a uma taxa adequada ao reconhecedor, bem como inova na técnica proposta para recuperação de pacotes perdidos baseada em redes neurais.

O estudo aqui desenvolvido representa uma contribuição original importante e útil às aplicações que necessitam de Reconhecimento de Voz Contínua Distribuído com Amplo Vocabulário. Diversos resultados inéditos de reconhecimento de voz foram obtidos e serão apresentados ao longo desta tese.

1.2. Base de Voz

A Base de vozes utilizada nesta tese foi construída especificamente para a mesma, e visa o treinamento e teste de sistemas de reconhecimento de voz contínua para o Português Brasileiro com amplo vocabulário e independentes do locutor (100 locutores – 50 locutores do sexo masculino e 50 locutores do sexo feminino, cada um falando todas as 1000 frases foneticamente balanceadas escolhidas para compor a base [2]). Estas frases são compostas de 9 à 12 palavras. A gravação só foi possível devido ao apoio do CNPq (através de projeto aprovado em edital Universal) que permitiu a contratação da Audioteca Sal e Luz (ONG – Organização Não Governamental – que visa a inclusão de deficientes visuais

através da gravação de livros falados) que gravou a base em seus estúdios, obteve os locutores e forneceu os arquivos para avaliação e inserção na base. A verificação de todos os arquivos de áudio para garantir que o padrão em qualidade, nomenclatura, frase lida, taxa de bits, etc, havia sido respeitado demandou um grande esforço de paciência e tempo, devido ao tamanho da base e para garantir a qualidade dos resultados a serem obtidos. A construção desta base pode também ser considerada uma contribuição relevante para o desenvolvimento de sistemas de reconhecimento de voz contínua para o Português Brasileiro.

A gravação foi realizada em estúdio, ambiente sem ruído, com uma especificação de gravação que pudesse abranger a entrada dos diversos codificadores de voz utilizados em Telefonia Móvel Celular e IP (taxa de amostragem 16 kHz e 16 bits por amostra com banda de sinal de 50 – 7000 Hz.). A base de voz produzida nesta tese está sendo disponibilizada para a utilização pública. A Fig. 1.1 é uma representação gráfica desta base e será utilizada para explicar alguns dos experimentos a serem realizados nesta tese.

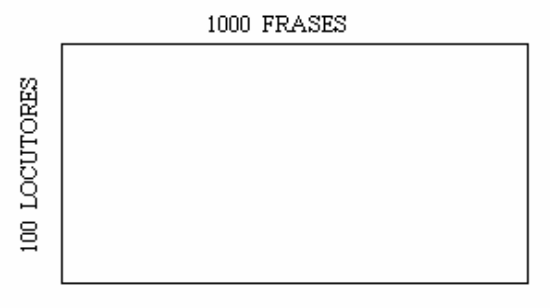

Figura 1.1 – Representação gráfica da base construída

A base foi dividida em três diferentes formas para produzir três diferentes cenários de experimentos a serem utilizados nesta tese. A primeira divisão está representada na Fig. 1.2, que pode ser considerada um cenário de reconhecimento dependente dos 100 locutores. A segunda divisão, representada pela Fig. 1.3, pode ser considerado um cenário independente do locutor com todas as frases usadas para teste e treino do sistema. A terceira divisão, representada Fig. 1.4, é um cenário de independência do texto e do locutor, que é o cenário que mais se aproxima do uso prático de reconhecimento de voz contínua distribuído para amplo vocabulário.

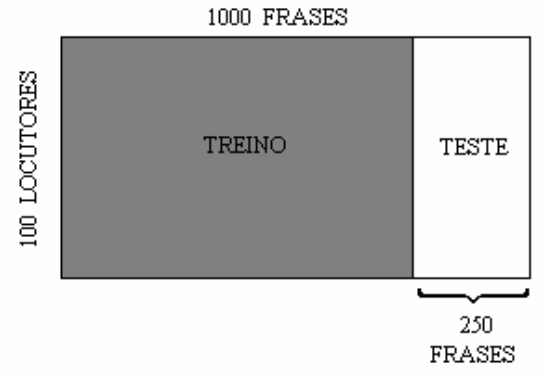

Figura 1.2 – Representação gráfica do cenário 1 (dependente dos 100 locutores)

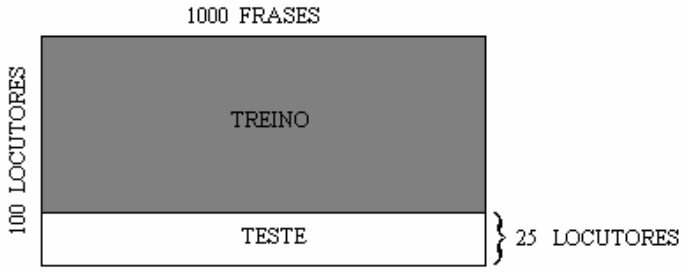

Figura 1.3 – Representação gráfica do cenário 2 (independente do locutor com todas as frases usadas para teste e treino do sistema)

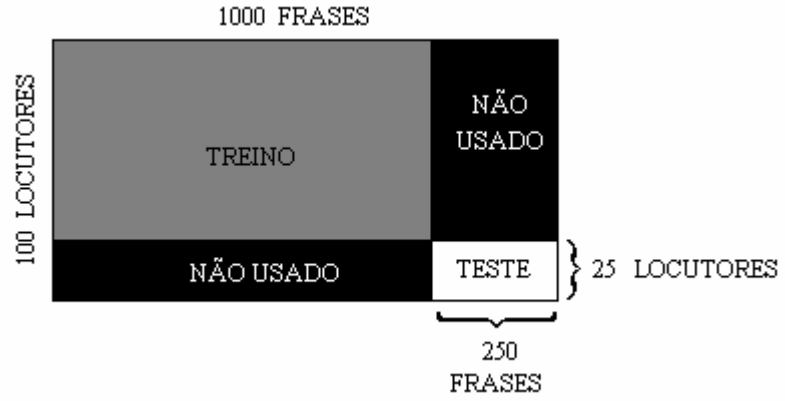

Figura 1.4 – Representação gráfica do cenário 3 (independente do locutor e do texto)

Uma distribuição de 75% e 25% da base foi utilizada respectivamente para treinamento e teste nos dois primeiros cenários de divisão da base. No terceiro cenário foi empregada uma distribuição da base de vozes de 56,25% para treino, 6,25% para teste e 37,5% não foi usada.

1.3. Organização da Tese

Esta Tese está organizada em sete capítulos. Além dessa Introdução seguem-se:

- Capítulo 2 – descreve os sistemas de reconhecimento distribuído, discorrendo sobre a extração de atributos e o reconhecimento de voz contínua.
- Capítulo 3 – é feita uma breve descrição dos codificadores de voz ITU-T (*International Telecommunication Union - Telecommunication Standardization Sector*) G.723.1 para Redes IP, AMR-NB (*Adaptive Multi-Rate Narrowband*) para Telefonia Celular e AMR-WB (*Adaptive Multi-Rate Wideband*) para Telefonia Celular e Redes IP, apresentando as principais características de funcionamento dos mesmos, dando assim subsídios à forma de utilização dos codificadores em sistemas reconhecimento de voz distribuído.
- Capítulo 4 – apresenta a dedução matemática dos parâmetros de reconhecimento de voz utilizados nesta tese.
- Capítulo 5 – analisa os atributos em reconhecimento de voz distribuído, tendo como finalidade apresentar uma nova técnica de interpolação dos atributos de reconhecimento, apresentando os resultados e a conclusão sobre a utilização desta nova técnica com os codificadores ITU-T G.723.1, AMR-NB e AMR-WB.
- Capítulo 6 – aborda o problema de perdas de pacotes em redes IP e redes móveis celulares, apresentando uma nova técnica, baseada em redes neurais, para a reconstrução dos pacotes perdidos, seus resultados e as conclusão sobre a utilização desta nova técnica.
- Capítulo 7 – finaliza o trabalho com algumas conclusões gerais e sugestões para trabalhos futuros.