Vladimir Fabregas Surigué de Alencar

Reconhecimento Distribuído de Voz Contínua com Amplo Vocabulário para o Português Brasileiro

TESE DE DOUTORADO

DEPARTAMENTO DE ENGENHARIA ELÉTRICA

Programa de Pós-graduação em Engenharia Elétrica

Rio de Janeiro

Agosto de 2009

Pontifícia Universidade Católica
do Rio de Janeiro

Vladimir Fabregas Surigué de Alencar

Reconhecimento Distribuído de Voz Contínua com Amplo Vocabulário para o Português Brasileiro

Tese de Doutorado

Tese de Doutorado apresentada como requisito parcial para obtenção do título de Doutor pelo Programa de Pós-Graduação em Engenharia Elétrica da PUC-Rio.

Orientador: Prof. Abraham Alcaim

Rio de Janeiro

agosto de 2009

PONTIFÍCIA UNIVERSIDADE CATÓLICA
DO RIO DE JANEIRO

Vladimir Fabregas Surigué de Alencar

Reconhecimento Distribuído de Voz Contínua com Amplo Vocabulário para o Português Brasileiro

Tese apresentada como requisito parcial para obtenção do grau de Doutor pelo Programa de Pós-Graduação em Engenharia Elétrica do Departamento de Engenharia Elétrica do Centro Técnico Científico da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Prof. Abraham Alcaim
Orientador
Centro de Estudos em Telecomunicações - PUC-Rio

Dra. Marley Maria Bernardes Rebuzzi Vellasco
Departamento de Engenharia Elétrica - PUC-Rio

Prof. Sergio Lima Netto
COPPE/UFRJ

Prof. Fernando Gil Vianna Resende Jr.
UFRJ

Prof. Fábio Violaro
UNICAMP

Profa. Rosângela Fernandes Coelho
IME

Prof. José Eugenio Leal
Coordenador Setorial do Centro
Técnico Científico - PUC-Rio

Rio de Janeiro, 24 de agosto de 2009

Vladimir Fabregas Surigué de Alencar

Graduou-se em Engenharia de Telecomunicações na UFF (Universidade Federal Fluminense) em 2003. Defendeu sua Dissertação de Mestrado em Março de 2005 pelo Departamento de Engenharia Elétrica da PUC-Rio.

Ficha Catalográfica

Alencar, Vladimir Fabregas Surigué de

Reconhecimento Distribuído de Voz Contínua com Amplo Vocabulário para o Português Brasileiro / Vladimir Fabregas Surigué de Alencar; orientador: Abraham Alcaim. – 2009.

131 f.: il. ; 30 cm

Tese (Doutorado em Engenharia Elétrica) – Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2009.

Incluí referências bibliográficas.

1. Engenharia elétrica - Teses. 2. Reconhecimento de voz distribuído 3. LSF 4. LPC 5. ISF 6. HMM 7. Redes IP 8. Redes Móveis Celulares 9. ITU-T G.723.1 10. AMR-NB 11. AMR-WB. 12. Redes Neurais I. Alcaim, Abraham. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Elétrica. III. Título.

CDD:621.3

Para meus pais Juarez e Laura, minha irmã Tatiana
e minha namorada Daniele
pelo carinho, apoio e confiança.

Agradecimentos

Ao meu orientador, Professor Abraham Alcaim, pela oportunidade, apoio e incentivo para a realização deste trabalho.

Ao corpo docente do CETUC, pelo aprendizado proporcionado.

Ao CNPq e à PUC-Rio, pelos auxílios concedidos.

Agradeço de forma especial aos meus pais, que mais uma vez foram fundamentais na minha vida, à minha irmã e à Daniele, que estiveram ao meu lado nesta empreitada.

Aos Professores que participaram da minha comissão examinadora.

A Audioteca Sal & Luz por todo apoio e esforço na construção e gravação da Base de vozes utilizada neste trabalho.

A todos os amigos que fiz no CETUC, que me proporcionaram não apenas momentos de aprendizagem, mas momentos de companheirismo que espero que se perpetuem.

Resumo

Alencar, Vladimir Fabregas Surigué; Alcaim, Abraham. **Reconhecimento Distribuído de Voz Contínua com Amplo Vocabulário para o Português Brasileiro.** Rio de Janeiro, 2009. 131p. Tese de Doutorado - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

Esta Tese visa explorar as oportunidades de melhoria do desempenho dos Sistemas Automáticos de Reconhecimento de voz com amplo vocabulário para o Português Brasileiro quando aplicados em um cenário distribuído (Reconhecimento de Voz Distribuído). Com esta finalidade, foi construída uma base de vozes para reconhecimento de voz contínua para o Português Brasileiro com 100 locutores, cada um falando 1000 frases foneticamente balanceadas. 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, em particular os codecs ITU-T G.723.1, AMR-NB e AMR-WB. Para um bom funcionamento dos Sistemas Automáticos de Reconhecimento de voz é necessário que os atributos de reconhecimento sejam obtidos a uma taxa elevada, porém os codificadores de Voz para Telefonia IP e Móvel Celular normalmente geram seus parâmetros a taxas mais baixas, o que degrada o desempenho do reconhecedor. Usualmente é utilizada a interpolação linear no domínio das LSFs (*Line Spectral Frequencies*) para resolver este problema. Nesta Tese foi proposta a realização da interpolação com a utilização de um Filtro Digital Interpolador que demonstrou ter um desempenho de reconhecimento muito superior ao da interpolação linear. Foi avaliado também o uso das ISFs (*Immittance Spectral Frequencies*) interpoladas como atributo de reconhecimento, as quais se mostraram inadequadas para esta finalidade, assim como as LSFs. Outro aspecto de fundamental importância para os reconhecedores de voz distribuídos é a recuperação de perda de pacotes, que tem impacto direto no desempenho de reconhecimento. Normalmente os codificadores inserem zeros nos pacotes perdidos ou interpolam linearmente os pacotes recebidos visando restaurar estes pacotes. Foi proposta nesta tese uma nova técnica baseada em Redes Neurais que se mostrou mais eficiente na restauração destes pacotes com a finalidade da realização do reconhecimento.

Palavras-chave

Reconhecimento de Voz Distribuído; LSF; LPC; ISF; HMM; Redes IP; Redes Móveis Celulares; ITU-T G.723.1; AMR-NB; AMR-WB; Redes Neurais

Abstract

Alencar, Vladimir Fabregas Surigué; Alcaim, Abraham (Advisor). **Distributed Recognition for Continuous Speech in Large Vocabulary Brazilian Portuguese.** Rio de Janeiro, 2009. 131p. DSc. Thesis - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

This Thesis aims at exploring several approaches for performance improvement of the Automatic Speech Recognition System with large vocabulary for the Brazilian Portuguese when applied in a distributed scenario (Distributed Speech Recognition). With this purpose, a speech database for continuous speech recognition for the Brazilian Portuguese with 100 speakers was constructed, each one uttering 1000 phonetic balanced sentences. The recording was carried out in a studio (environment without noise) with a specification of recording that would be able to allow the input of several speech codecs in Cellular Mobile Telephony and IP Networks, in particular the ITU-T G.723.1, AMR-NB and AMR-WB. In order to work properly, Automatic Speech Recognition Systems require that the recognition features be extracted at a high rate. However, the Speech codecs for Cellular Mobile Telephony and IP Networks normally generate its parameters at lower rates, which degrades the performance of the recognition system. Usually the linear interpolation in the LSF (*Line Spectral Frequencies*) domain is used to solve this problem. In this Thesis the accomplishment of the interpolation with the use of a Digital Filter Interpolator was proposed and demonstrated to have a higher performance than the linear interpolation in recognition systems. The use of the interpolated ISFs (*Immittance Spectral Frequencies*) was also evaluated as recognition feature, which had shown to be inadequate for this purpose, as well as the LSFs. Another very important aspect for the distributed speech recognizers is the recovery of lost packets, that has direct impact in the recognition performance. Normally the coders insert zeros in the lost packets or interpolate linearly the received packets aiming to restore them. A new technique based on Neural Networks was proposed in this thesis that showed to be more efficient in the restoration of these lost packets with the purpose of speech recognition.

Keywords

Distributed Speech Recognition; LSF; LPC; ISF; HMM; IP Networks; Cellular Mobile Networks; ITU-T G.723.1; AMR-NB; AMR-WB; Neural Networks

Sumário

Lista de figuras

Lista de tabelas

Lista de acrônimos

ACELP	*Algebraic Code-Excited Linear Prediction*
AMFCC	*Autocorrelation Mel-Frequency Cepstral Coefficients*
AMR-NB	*Adaptive Multi-Rate Narrowband*
AMR-WB	*Adaptive Multi-Rate Wideband*
ATK	*Application Toolkit for HTK*
CDMA	*Code Division Multiple Access*
CELP	*Code-Excited Linear Predictive*
CMR	*Comprimento Médio de Rajada*
CNPq	Conselho Nacional de Pesquisa e Desenvolvimento
CS-ACELP	*Conjugate Structure- Algebraic Code-Excited Linear Prediction*
CSR	*Continuous Speech Recognition*
DC	*Direct Current*
DP	*Dynamic Programming*
DSR	*Distributed Speech Recognition*
EVRC	*Enhanced Variable Rate Coder*
FFT	*Fast Fourier Transform*
FIR	*Finite Impulse Response*
GSM	*Global System for Mobile Communication*
GSM-EFR	*GSM-Enhanced Full Rate*
GSM-FR	*GSM-Full Rate*
GSM-HR	*GSM-Half Rate*
HMM	*Hidden Markov Model*
HTK	*HMM Toolkit*
IFFT	*Inverse Fast Fourier Transform*
IMT-2000	*International Mobile Telecommunications-2000*
IP	*Internet Protocol*
ISF	*Immittance Spectral Frequencies*
ITU-T	*International Telecommunication Union - Telecommunication Standardization Sector*
LM	*Language Mode*

LP	*Linear Prediction*
LPC	*Linear Predictive Coding*
LPCC	*LPC Cepstrum*
LSF	*Line Spectral Frequencies*
LVDCSR	*Large Vocabulary Distributed Continuous Speech Recognition*
LVR	*Large Vocabulary Recognition*
MFCC	*Mel-Frequency Cepstral Coefficients*
MLPCC	*Mel LPC Cepstral Coefficients*
MPCC	*Mel Pseudo-Cepstral Coefficients*
MPCEP	*Mel Pseudo-Cepstrum*
MP-MLQ	*Multi-pulse Maximum Likelihood Quantization*
ONG	Organização Não Governamental
PCC	*Pseudo-Cepstral Coefficients*
PCEP	*Pseudo-Cepstrum*
PCM	*Pulse-Code Modulation*
PCP	*Probabilidade Condicional de Perda*
PPI	*Probabilidade de Perda Incondicional*
PSTN	*Public switching Telecommunications Network*
PS-PA	*Pitch-Synchronous Peak Amplitude*
PSVQ	*Predictive Split Vector Quantizer*
PS-ZCPA	*Pitch-Synchronous Zero Crossings with Peak Amplitudes*
QCELP	*Qualcomm Code-Excited Linear Predictive*
RAM	*Random Access Memory*
RAV	Reconhecimento Automático de Voz
RAS	*Relative Autocorrelation Sequence*
RTT	*Round-Trip Time*
SID	*Silence Descriptor*
SMQ	*Split Matrix Quantization*
S-MSVQ	*Split-Multistage Vector Quantization*
SSCH	*Subband Spectral Centroid Histograms*
SVQ	*Split Vector Quantization*
TDMA	*Time Division Multiple Access*
TIA	*Telecommunications Industry Association*

TPP	*Taxa de Perda de Pacote*
USA	*United States of America*
VAD	*Voice Activity Detection*
VSELP	*Vector Sum Excited Linear Predictive*
WCDMA	*Wideband Code Division Multiple Access*
$\overline{WRR}$	*Average Word Recognition Rates*
ZCPA	*Zero Crossings with Peak Amplitudes*
3GPP	*3rd Generation Partnership Project*

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.

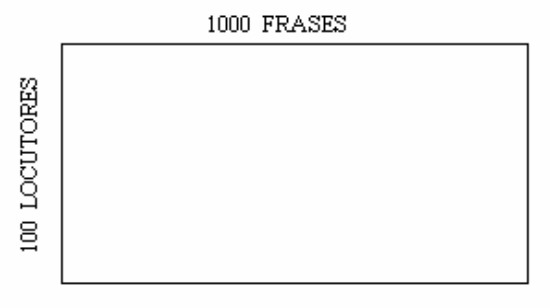

Figura 1.1 – Representação gráfica da base construída

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.

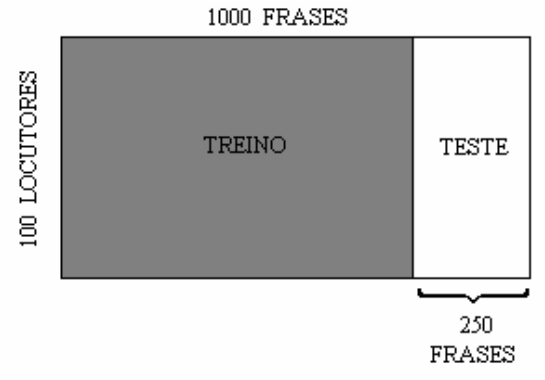

Figura 1.2 – Representação gráfica do cenário 1 (dependente dos 100 locutores)

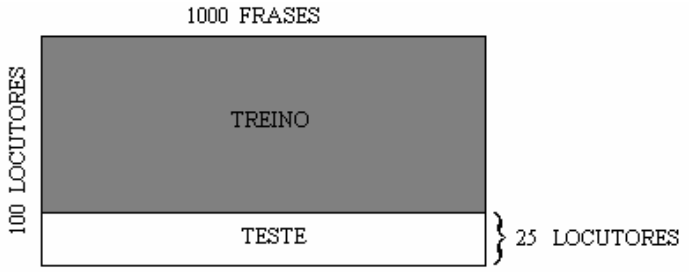

Figura 1.3 – Representação gráfica do cenário 2 (independente do locutor com todas as frases usadas para teste e treino do sistema)

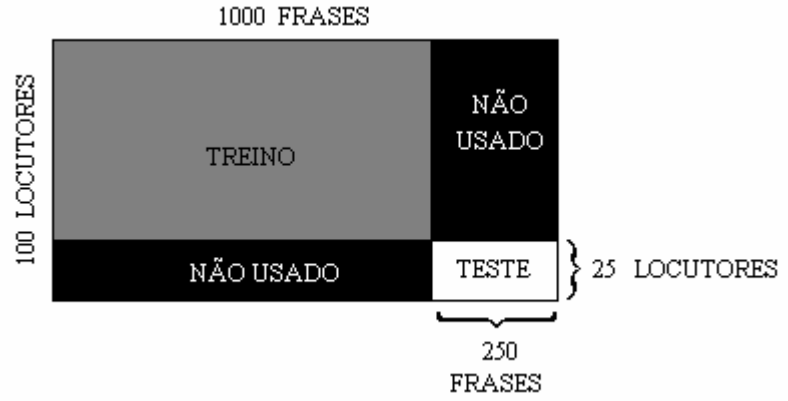

Figura 1.4 – Representação gráfica do cenário 3 (independente do locutor e do texto)

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.

2
Reconhecimento Voz Contínua em Sistemas Distribuídos

O conceito de reconhecimento de voz distribuído (*DSR – Distributed Speech Recognition*) foi desenvolvido como uma forma eficiente de transladar a tecnologia de Reconhecimento Automático de Voz (RAV) para o ambiente móvel e redes IP.

A idéia do *DSR* consiste em usar um *front-end* local, de onde os parâmetros de voz são obtidos e transmitidos através de um canal de dados, até um *back-end* onde se localiza o reconhecedor de voz. Esta idéia pode ser observada na Fig. 2.1.

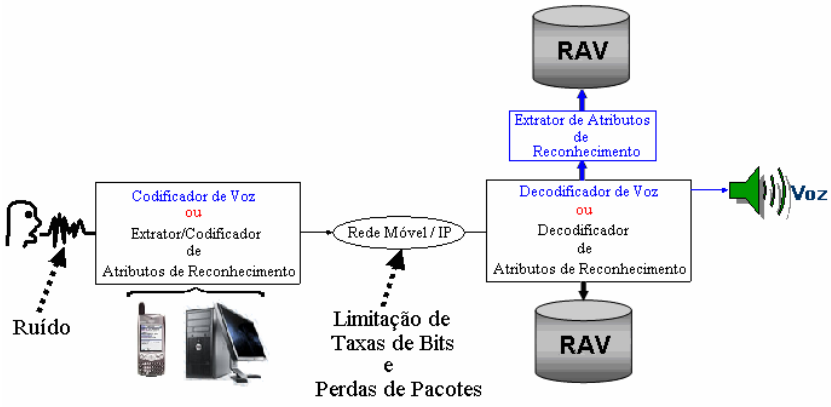

Figura 2.1 Sistemas de Reconhecimento Distribuído – Diagrama Básico

Outra característica importante desta abordagem está no fato de que uma análise relativamente simples de voz é realizada no *front-end* local, enquanto que a maior parte do processamento é colocada no servidor de reconhecimento que pode ser facilmente atualizado para novas tecnologias e serviços, sem custo adicional para o usuário [3].

2.1.
Extrator de Atributos em Sistemas Distribuídos

A partir das características básicas de um *DSR*, é importante analisar as abordagens de reconhecimento com os respectivos atributos utilizados, de forma a definir um bom sistema a ser implementado e o que nele pode ser melhorado.

Cabe ressaltar que a dedução matemática detalhada dos atributos a serem utilizados no reconhecimento será feita no Capítulo 4. Aqui serão apresentadas três abordagens diferentes e os atributos de reconhecimento mais utilizados em cada uma delas [4], [5] e [6].

A. Extrator de atributos usando os parâmetros do codificador de voz

Um esquema deste tipo é ilustrado na Fig. 2.2, onde pode ser observada a sua adequação às situações em que se queira realizar o reconhecimento e a recuperação da voz do locutor.

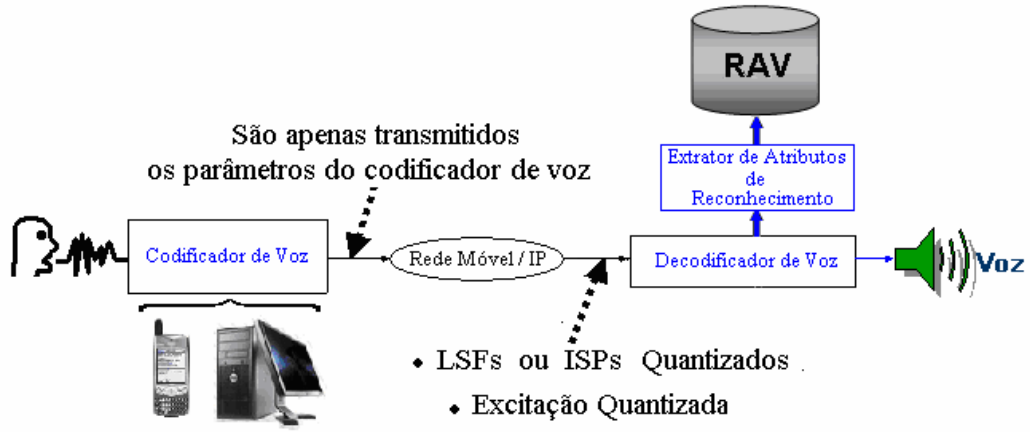

Figura 2.2 – Sistema de reconhecimento de voz distribuído baseado nos parâmetros de voz do codificador

Neste sistema existe uma ampla variedade de atributos de reconhecimento que podem ser obtidos a partir dos parâmetros do decodificador no processo de recuperação da voz, o que simplifica bastante o extrator de atributos.

O codificador de voz transmite as LSFs ou os ISFs e os parâmetros da excitação quantizados, como apresentado na Fig. 2.2. Estes parâmetros da voz trafegam pela rede e chegam ao receptor, onde deseja-se realizar o reconhecimento.

Diretamente dos parâmetros LSFs quantizados pode-se obter atributos de reconhecimento, sendo eles em número de quatro: PCC (*Pseudo-Cepstral Coefficients*) [1], PCEP (*Pseudo-Cepstrum*) [4], MPCC (*Mel Pseudo-Cepstral Coefficients*) [1] e MPCEP (*Mel Pseudo-Cepstrum*) [4].

Dos parâmetros ISFs quantizados tem que se investigar o desempenho quando utilizados diretamente no reconhecimento, bem como possíveis transformações para atributos de reconhecimento.

Dos parâmetros LSF quantizados ou dos ISF quantizados o decodificador de voz do receptor obtém os parâmetros LPC de onde se podem extrair dois atributos de reconhecimento: LPCC (*LPC Cepstrum*) [5] e MLPCC (*Mel LPC Cepstral Coefficients*) [5].

Seria interessante também neste cenário ser analisada a combinação da informação da excitação com os atributos de reconhecimento obtidos de LSF, ISF ou LPC de forma a obter um novo vetor de atributos que tenha melhor desempenho que os originais.

B. Extrator de atributos a partir de voz decodificada

Uma ilustração deste sistema é apresentada na Fig. 2.3, onde se pode observar que o mesmo tem que recuperar a voz do locutor, para efetuar o reconhecimento, o que tem demonstrado desempenho inferior ao das demais abordagens [6].

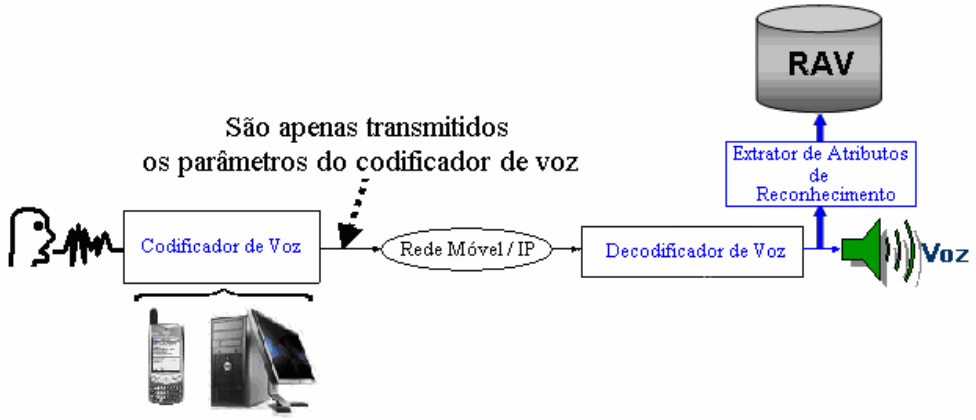

Figura 2.3 – Sistema de reconhecimento de voz distribuído baseado em voz decodificada

Da voz reconstruída podem ser obtidos vários atributos de reconhecimento, dentre os quais: MFCC (*Mel-Frequency Cepstral Coefficients*) [7], AMFCC (*Autocorrelation Mel-Frequency Cepstral Coefficients*) [8], RAS (*Relative Autocorrelation Sequence*) [9], ZCPA (*Zero Crossings with Peak Amplitudes*) [10], PS-ZCPA (*Pitch-Synchronous Zero Crossings with Peak Amplitudes*) [11],

PS-PA (*Pitch-Synchronous Peak Amplitude*) [12], SSCH (*Subband Spectral Centroid Histograms*) [13].

C. Extrator de atributos para reconhecimento no terminal do usuário

Uma ilustração deste sistema é apresentada na Fig. 2.4, onde se pode observar que o mesmo é bastante adequado para situações onde se deseja realizar apenas o reconhecimento, não sendo um requerimento a recuperação de voz.

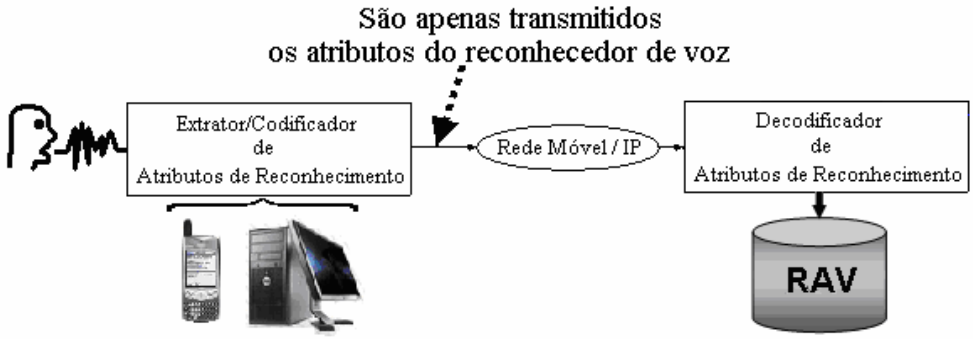

Figura 2.4 – Sistema de reconhecimento de voz distribuído com codificação dos atributos de reconhecimento no *front-end* local

Este sistema pode ser combinado com um sistema de codificação de voz, porém, isto implicará em maior quantidade de informação a ser transmitida no canal e maior processamento do *front-end* local.

Os atributos de reconhecimento gerados pelo codificador de atributos do *front-end* local são os mesmos obtidos de voz reconstruída apresentados no reconhecedor de voz, a partir de voz decodificada. No entanto, os mesmos são obtidos de voz original. Estes sistemas apresentam desempenho igual aos sistemas de reconhecimento automático de voz onde todo o processamento é feito em um único local [6].

2.2. Reconhecimento de Voz Contínua

Sistemas atuais de reconhecimento de voz contínua (*CSR – Continuous Speech Recognition*) com amplo vocabulário (*LVR – Large Vocabulary Recognition*) são estritamente baseados nos princípios de reconhecimento

estatístico de padrões. Os métodos básicos em que se aplicam estes princípios têm ainda forte influência de sistemas pioneiros da década de 70 [14], [15]. A arquitetura representada na Fig. 2.5 é praticamente um consenso na área e é composta pelos seguintes componentes: modelos acústicos, léxico de palavras (opcional), modelo linguístico e, principalmente, o decodificador. Estes blocos serão explorados no restante desta seção.

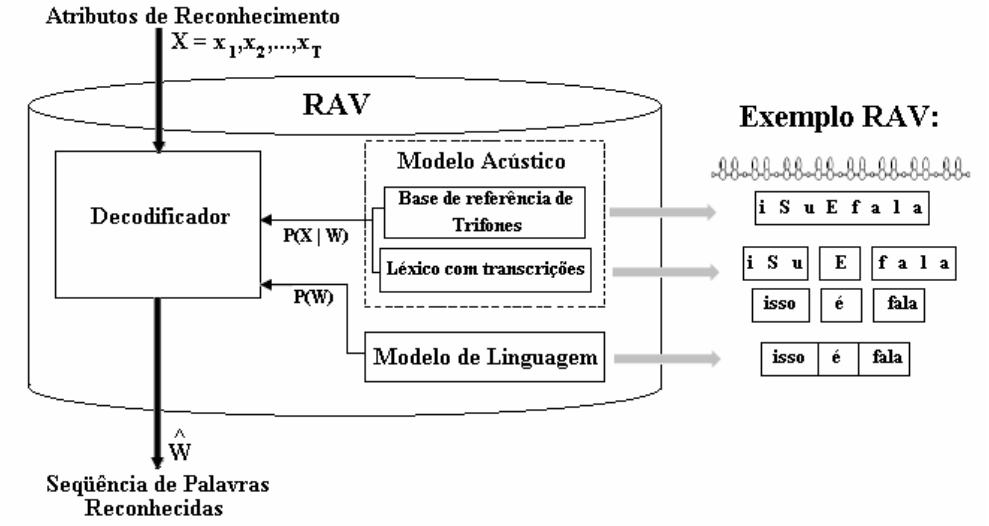

Figura 2.5 – Diagrama em blocos de um sistema de reconhecimento automático de voz baseado em modelos estatísticos de subunidades de palavras [16]

Um sistema de reconhecimento de voz converte o sinal acústico observado em sua representação ortográfica correspondente. Houve um grande avanço na resolução deste problema através da utilização de um modelo estatístico de distribuição conjunta, $P(W,X)$, entre a sequência de palavras pronunciadas W e a sequência de informações acústicas observadas X, numa abordagem conhecida como modelo fonte canal [17]. Nesta abordagem, o sistema de reconhecimento procura uma estimativa $\hat{W}$, da sequência de palavras pronunciadas, a partir da evidência acústica observada X, usando a distribuição de probabilidades a posteriori, $P(W \mid X)$. Para minimizar a taxa de erro, o sistema escolhe a sequência de palavras que maximiza essa distribuição

$$\hat{W} = \arg\max_{w}\left[P(W \mid X)\right] = \arg\max_{w}\left[\frac{P(W)P(X \mid W)}{P(X)}\right] \qquad (2.1)$$

Em (2.1), após a aplicação do Teorema de Bayes, a distribuição a posteriori é decomposta em $P(W)$, a probabilidade a priori da sequência de palavras W, e $P(X|W)$ que é a probabilidade de observar a evidência acústica X quando a sequência W é pronunciada. A distribuição $P(W)$ refere-se às palavras que poderiam ter sido pronunciadas (a fonte) e está associada a um modelo de linguagem. O modelo de probabilidade de uma observação $P(X|W)$ (o canal) é chamado de modelo acústico.

A. Modelo Acústico

O propósito da modelagem acústica é prover um método que calcule a verossimilhança de qualquer sequência de vetores X, dada uma sequência de palavras W. A princípio, a distribuição de probabilidade requerida, $P(X|W)$, pode ser modelada através de inúmeras palavras e o cálculo estatístico de sequências de vetores correspondentes. No entanto, esse método é impraticável para sistemas com amplo vocabulário e, ao invés disso, as palavras modeladas pelo sistema são usualmente decompostas em seus respectivos fones.

Nesta tese, seguindo algumas das implementações com melhor desempenho no cenário de reconhecimento distribuído de voz contínua para amplo vocabulário [18], cada fone foi representado por um HMM (*Hidden Markov Model*) de primeira ordem que contém três estados emissores e uma topologia simples do tipo esquerda-direita (*left-right*). Estados de entrada e saída, não emissores, foram acrescidos à modelagem para facilitar a união entre modelos, ficando o modelo conforme ilustrado na Fig 2.6. O estado de saída (não emissor – não possui função densidade de probabilidade atrelada a este estado) do modelo de um fone pode ser unido com o estado de entrada (não emissor) de outro para criar um HMM composto. Isto permite que modelos de fone sejam unidos para formar palavras e estas unidas para formar frases completas. No entanto, o modelo de pausa, por ser estacionário, foi representado por uma topologia mais simples, constituída apenas de um estado emissor de saída.

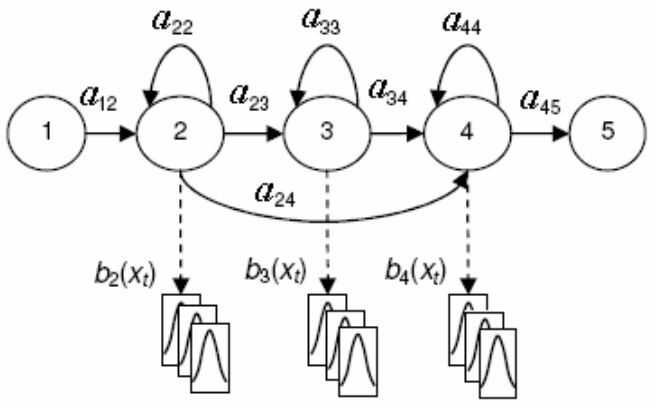

Figura 2.6 – Modelo de fone baseado em HMM

HMM é uma máquina de estados finita que modifica seu estado a cada unidade de tempo e, a cada tempo t e estado j, um vetor acústico de fala x_t é gerado com função densidade de probabilidade $b_j(x_t)$, composta nesta tese pela mistura de 20 (M) gaussianas contínuas multivariadas com covariância diagonal [18], pois em [16] formaram simuladas com outras quantidades de gaussianas (1, 4, 8, 12, 16) e a com 20 gaussianas foi a que apresentou melhor desempenho:

$$b_j(x_t) = \sum_{m=1}^{M} c_{jm}\, \eta(x_t, \mu_{jm}, \Sigma_{jm}) \tag{2.2}$$

onde c_{jm} é o peso do componente m da mistura, no estado j, e $\eta(x, \mu, \Sigma)$ denota uma Gaussiana multivariada de vetor média μ e matriz covariância Σ,

$$\eta(x, \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{-\frac{1}{2}(x-\mu)' \Sigma^{-1} (x-\mu)} \tag{2.3}$$

onde $(\cdot)'$ indica transposta e $|\cdot|$ indica determinante.

Além disso, a transição de um estado i para um estado j também é probabilística e governada por uma probabilidade discreta a_{ij}.

A função densidade de probabilidade conjunta condicionada de uma sequência de vetores X e uma sequência de estados S, dado um modelo M, é calculada pelo produto das probabilidades de transição de estados com a função

densidade de probabilidade de emissão de saída. Assim, a função densidade de probabilidade conjunta condicionada de uma sequência de vetores acústicos X e uma sequência de estados $S = s(1), s(2), s(3), ..., s(T)$ é

$$p(X, S \mid M) = a_{s(0)s(1)} \prod_{t=1}^{T} b_{s(t)}(x_t) a_{s(t)s(t+1)} \tag{2.4}$$

onde o estado inicial $s(0)$ é restringido ao estado de entrada e $s(T+1)$ é restringido ao estado de saída. Na prática, apenas a sequência observada X é conhecida, enquanto a suposta sequência S fica escondida pelo modelo. Por isto a técnica é denominada Modelos Escondidos de Markov. A função densidade de probabilidade condicionada $p(X \mid M)$, necessária para resolver (2.1), é facilmente encontrada somando a Equação (2.4) sobre todas as possibilidades de sequências de estado

$$p(X \mid M) = \sum_{s} p(X, S \mid M) \tag{2.5}$$

Um método recursivo e eficiente para realizar este cálculo é o algoritmo avanço-retorno (*Forward-Backward*). Uma característica crucial desse algoritmo é que ele também permite o cálculo da probabilidade de se estar num estado específico do modelo em um instante específico de tempo. Isso leva a um procedimento simples e eficiente conhecido como algoritmo Baum-Welch [19], o qual procura estimativas de máxima verossimilhança para o conjunto de parâmetros de cada HMM.

No exposto até aqui, assumiu-se de forma implícita ser necessário apenas um HMM por fone, porém na prática, efeitos contextuais causam uma larga variação na maneira como sons diferentes são produzidos, chamando-se de alofones os fones existentes em diferentes contextos. Logo, para se conseguir uma boa discriminação fonética, distintos HMMs devem ser utilizados para representar os fones em cada um dos diferentes contextos. Nesta tese foram utilizados trifones para resolver esse problema, onde cada um dos fones possui modelos distintos de HMM para os pares formados por fones situados à direita e a esquerda. Os contextos de trifones podem abranger ou não as fronteiras entre palavras. Quando

são consideradas as fronteiras os trifones são chamados de trifones entre-palavras (*cross-word triphones*) o que resulta em uma modelagem mais precisa e com melhor desempenho, o que será utilizado nesta tese. O contrário ocorre quando as fronteiras são desconsideradas, onde temos os chamados trifones intrapalavra (*word-internal triphones*), o que resulta em um sistema com menor habilidade em se modelar efeitos contextuais nas fronteiras de palavras e na modelagem de fala fluente, resultando em um desempenho um pouco pior no reconhecimento.

O uso de misturas de Gaussianas nas distribuições de saída permite que as distribuições de cada estado sejam modeladas com muita precisão. No entanto, quando trifones são usados, o sistema resultante requer o treino de muitos parâmetros, logo precisa de uma base muito grande para ser treinado. O problema de muitos parâmetros e poucos dados de treino é absolutamente crucial no projeto de um reconhecedor de fala estatístico.

Uma tentativa de se resolver este problema, e que será utilizado nesta tese, é a união de estados (*state-tying*) [20], [21], nesta técnica são unidos os estados que são acusticamente indistinguíveis (chamados comumente de senones). Isso permite que os dados associados a cada estado individualmente sejam combinados num recurso em comum e, assim, estimados de forma mais robusta. Após as uniões, vários estados passam a compartilhar as mesmas distribuições.

A escolha sobre quais estados devem ser unidos será realizada através de árvores de decisão fonética [22], [23], [24]. Este método envolve a construção de uma árvore binária para cada fone e posição de estado. Cada árvore tem, em cada um dos nós, uma pergunta fonética do tipo sim/não, como “O contexto à esquerda é nasal?”. Inicialmente, todos os estados de um dado fone são posicionados no nó raiz da árvore. De acordo com as respostas, o conjunto de estados é sucessivamente dividido até que os estados tenham alcançado os nós terminais da árvore. Todos os estados depositados em um mesmo nó terminal são então unidos.

As questões em cada nó são escolhidas para maximizar a verossimilhança entre os dados de treino e o conjunto resultante da união de estados. Na prática, as árvores de decisão fonética resultam em grupamentos de estados compactos e de boa qualidade, os quais possuem dados suficientes para estimar de forma robusta as misturas de Gaussianas das funções densidade de probabilidade de emissão de saída. Além disso, elas podem ser usadas para sintetizar um HMM para qualquer

possível contexto, apareça ele nos dados treino, ou não, simplesmente descendo na árvore e usando as distribuições de estado associadas aos nós terminais.

B. Modelo de linguagem

O modelo de linguagem (*LM – Language Model*) pode ser formulado como a probabilidade $P(W)$ definida por

$$P(W) = P(W_1^k) = P(w_1,...,w_k) \tag{2.6}$$

onde reescreve-se $P(W)$ como $P(W_1^k)$ acrescentando os índices 1 e k de forma a representar uma sequência com k palavras. A probabilidade conjunta de ocorrência de palavras da (2.6) pode ser substituída por um produto de suas probabilidades condicionais da seguinte forma

$$P(W_1^k) = P(w_1)\prod_{i=2}^{k} P(w_i \mid w_1,...,w_{i-1}) \tag{2.7}$$

Uma maneira simples, porém efetiva, de se obter estas probabilidades é com a utilização de n-gramas, onde se assume que w_k depende apenas das n-1 palavras anteriores a ela, reduzindo a equação (2.7) para

$$P(W_1^k) \approx P(w_1)\prod_{i=2}^{k} P(w_i \mid W_{i-(n-1)}^{i-1}) \tag{2.8}$$

Na tentativa de capturar a correlação existente entre palavras vizinhas, os n-gramas acabam simultaneamente absorvendo sintaxe, semântica e pragmática existente nas frases observadas. Isso os faz extremamente efetivos em idiomas como inglês ou português onde a ordem das palavras é importante visto que os efeitos contextuais mais fortes normalmente vêm dos vizinhos mais próximos. Além disso, as distribuições de probabilidade dos n-gramas podem ser computadas diretamente de textos prontos e, portanto, não há necessidade de se definir regras explícitas de linguística como uma gramática formal do idioma.

A princípio, os n-gramas podem ser estimados através de simples contadores de frequência e armazenados em uma tabela de memória (*look-up table*). Para o caso de trigramas (n=3), que será adotado em todos os testes desta tese [18]

$$P(w_k \mid w_{k-2}, w_{k-1}) = \frac{t(w_{k-2}, w_{k-1}, w_k)}{b(w_{k-2}, w_{k-1})} \tag{2.9}$$

onde $t(w_{k-2}, w_{k-1}, w_k)$ é o número de vezes que o trigrama w_{k-2}, w_{k-1}, w_k aparece nos dados de treino e $b(w_{k-2}, w_{k-1})$ é o número de vezes que o bigrama w_{k-2}, w_{k-1} aparece. Para lidar convenientemente com a condição de início de frase, associa-se uma nova palavra, w_0, a um símbolo de início de frase, com probabilidade $P(w_0) = 1$. Utilizando (2.7) e (2.8) tem que $P(w_1) = P(w_1 \mid w_0)$, podendo $P(w_1 \mid w_0)$ ser estimado pelo número de vezes que a palavra w_1 começou a frase dividida pelo número de frases da base utilizada para gerar o modelo de linguagem. Em seguida prossegue-se na forma indicada por (2.9) para o cálculo distribuição da probabilidade dos trigramas. Pode-se, claramente, perceber o problema de que, para um vocabulário com V palavras, existem V^3 potenciais trigramas. Para o vocabulário de 60.080 palavras desta tese (aproximadamente 10^{15} possíveis trigramas), este número é extremamente grande. O fato de que muitos dos possíveis trigramas não aparecerão nos dados de treino (240.000 frases da base de treino utilizada nesta tese extraídos de [25]) e muitos outros aparecerão apenas uma ou duas vezes, faz com que a estimativa de (2.9) seja muito pobre. Em resumo, há um problema agudo de escassez de dados.

Uma das soluções de treinamento projetada como solução à escassez de dados é o uso de uma combinação de técnicas de desconto (*discounting*) e retrocesso (*backing-off*) [26], [27]. Desconto significa que as contagens de trigramas que ocorrem com maior frequência são reduzidas e a massa probabilística resultante em excesso é redistribuída entre os trigramas que ocorrem com menos frequência, ou mesmo não ocorram. O retrocesso é aplicado quando há pouquíssimos trigramas para formar qualquer tipo de estimativa (ex.: apenas uma ou duas ocorrências nos dados de treino). Ele envolve a substituição

da probabilidade de um trigrama por uma probabilidade de bigrama escalada, que é

$$P(w_k \mid w_{k-2}, w_{k-1}) = B(w_{k-2}, w_{k-1}) P(w_k \mid w_{k-1}) \tag{2.10}$$

onde $P(w_k \mid w_{k-1})$ é a probabilidade bigrama e $B(w_{k-2}, w_{k-1})$ é um fator de escalamento calculado de forma a garantir que a soma das probabilidades trigrama some um

$$\sum_{w_{k-2}, w_{k-1}} P(w_k \mid w_{k-2}, w_{k-1}) = 1 \tag{2.11}$$

Ainda que uma estimativa robusta das probabilidades de trigramas requeira um cuidado considerável, os problemas decorrentes são solucionáveis e um bom desempenho tem sido obtido. A técnica de n-gramas possui deficiências óbvias que resultam da sua inabilidade de explorar restrições de maior amplitude como a concordância entre sujeito e verbo de uma frase [28]. Como consequência, várias alternativas foram estudadas como modelos baseados em árvore [29], modelos em treliça [30], modelos com gatilhos [31], modelos com histórico [32] e n-gramas variáveis [33]. Entretanto, em geral, todas estas tentativas resultaram em apenas pequenos ganhos de desempenho sobre um considerável custo computacional. Por esse motivo, após quase duas décadas de pesquisa, os modelos de linguagem conhecidos como bigramas e trigramas ainda dominam os sistemas *LVR*. Nestes sistemas, têm se usado, para modelagem da linguagem, bases de dados de 1 milhão a 500 milhões de palavras, correspondendo a vocabulários de 1 mil a 267 mil palavras distintas, para construção dos modelos trigrama [17] (nesta tese utilizou-se uma base de dados com 3.993.906 palavras, correspondendo a um vocabulário de 60.080 palavras distintas – sendo este o tamanho do vocabulário utilizado no decodificador – em 240.000 frases extraídas de [25]).

C. Decodificador

O termo decodificador, no contexto de reconhecimento de voz, foi criado numa analogia à terminologia usada nos métodos que utilizam estados finitos para decodificação [34] no campo da teoria da informação.

A decodificação é um processo de busca no qual uma sequência de vetores correspondentes a características acústicas (atributos) do sinal de voz é comparada com modelos de palavras. De uma maneira geral, o sinal de voz e suas transformações não fornecem uma indicação clara das fronteiras entre palavras nem do número total de palavras em uma dada locução, de modo que a determinação destas é parte do processo de decodificação. Neste processo, todos os modelos das palavras (formadas por seus respectivos modelos de fones) são comparados com uma sequência de vetores acústicos. O número de modelos cresce com o vocabulário, e pode gerar espaços de busca grandes, o que torna o processo de busca oneroso em termos computacionais, e portanto lento. Em geral, esta etapa do reconhecimento, nos sistemas modernos, é responsável por praticamente todo o esforço computacional no reconhecimento de fala contínua e, portanto, é a que determina a velocidade final desses sistemas.

Durante o processo de maximização de (2.1), repetida por conveniência a seguir, o termo $P(X|W)$ é expandido em função do modelo acústico, vinculando os estados dos HMMs à emissão de saídas nos mesmos. Dessa forma, $P(X|W)$ passa a ser calculado como a soma de todas as possibilidades de transições entre as possíveis sequências de estados do modelo sob hipótese

$$\begin{aligned}\hat{W}_1^N &= \arg\max_{W_1^N}\left\{P(W_1^N)P(X_1^T|W_1^N)\right\}\\ &= \arg\max_{W_1^N}\left\{P(W_1^N)\sum_{S_1^T}P(X_1^T,S_1^T|W_1^N)\right\}\\ &\approx \arg\max_{W_1^N}\left\{P(W_1^N)\max_{S_1^T}P(X_1^T,S_1^T|W_1^N)\right\}\end{aligned} \tag{2.12}$$

onde $W_1^N = w_1...w_N$ representa a hipótese de sequência de palavras (compostas por seus respectivos HMMs de subpalavras), $S_1^T = \{s_1...s_T\}$, a hipótese de sequência

de estados dentro do modelo e $X_1^T = \{x_1 ... x_T\}$, os vetores acústicos observados. O somatório nesta equação é então substituído por uma maximização, num processo referido como Aproximação de Viterbi [15]. Ao invés de somar sobre todos os caminhos, considera-se apenas o caminho mais provável.

Neste processo de maximização, o espaço de busca pode ser descrito como uma rede onde se busca o melhor alinhamento temporal entre a sequência de entrada e os estados dos modelos. A busca pode ser realizada em dois níveis: no nível de estados (S_1^T) e no nível de palavras (W_1^N). É possível recombinar eficientemente as hipóteses nos dois níveis usando programação dinâmica (*DP - Dynamic Programming* [35]), limitando a explosão combinatória do número de hipóteses de busca.

Nesta tese será usada a decodificação de Viterbi [16] com feixe de busca que é um algoritmo de programação dinâmica que procura no espaço de estados a mais provável sequência de estados que modele o trecho de fala de entrada. O espaço de estados é construído pelos chamados Palavra-HMMs, que são formados pela concatenação dos HMMs dos trifones que as constituem. Todos os Palavra-HMMs constituídos dessa forma são percorridos pelo algoritmo de busca em paralelo. Como o espaço de estados é grande, mesmo para aplicações com vocabulário de tamanho médio, a heurística da busca em feixe é normalmente aplicada para limitar a busca, através da poda (*pruning*) de estados menos prováveis. A combinação do algoritmo de busca e do método de poda utilizados é referida como busca em feixe de Viterbi (*Viterbi beam search*). A decodificação de Viterbi é uma busca síncrona no tempo que processa a fala, segmento a segmento, atualizando todos os estados associados a um segmento antes de passar para o próximo.

Por ser síncrona, a busca em feixe é possivelmente a técnica mais utilizada em sistemas de reconhecimento de voz, desde as referências [36], [37], [38] até hoje em dia. É um algoritmo de busca ao estilo da busca em largura (*breadth-first*), no qual os nós de uma determinada altura h (distância entre um nó e o nó raiz) são analisados antes de passar para nós em uma altura $h+1$ em relação à raiz. Porém, diferente da busca em largura, a busca em feixe expande, a cada passo, apenas os nós que apresentam uma probabilidade alta de sucederem. Apenas estes nós expandidos permanecem no feixe e o restante é ignorado

(podado) aumentando assim a eficiência da busca. A Fig. 2.7, ilustra os passos de busca em feixe. Os nós em cinza são aqueles que foram mantidos na busca. Os nós pontilhados são aqueles que foram explorados, porém pela sua baixa probabilidade foram eliminados.

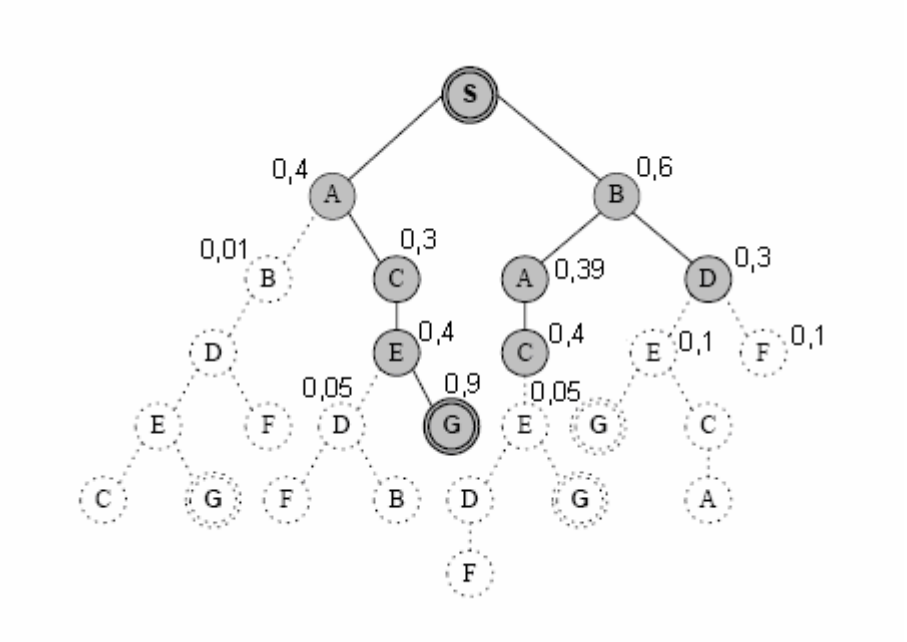

Figura 2.7 – Dinâmica da Busca em Feixe [18]

2.3. Conclusão

Neste capítulo foi feita a apresentação dos sistemas de reconhecimento de voz distribuídos, abordando a extração de atributos e o reconhecimento de voz contínua.

No capítulo seguinte, será feita a apresentação dos codificadores de voz que serão utilizados na montagem dos sistemas de reconhecimento distribuído no ambiente celular e de voz sobre IP.

3
Codificadores de Voz em Telefonia IP e Móvel Celular

Para o desenvolvimento de sistemas de reconhecimento de voz distribuídos voltados a rede IP e rede móvel celular é necessário conhecer os codificadores de voz utilizados nas mesmas.

Os padrões de codificadores para redes IP e ambientes móveis celulares são enumerados abaixo, colocando-se na seguinte ordem as informações – nome do padrão / tipo do *codec* – ano – taxa – tipo de rede em que é utilizado:

- ITU-T G.723.1 / CELP (*Code-Excited Linear Predictive*) – 1995 – 5,3 ou 6,3 kbit/s – IP
- ITU-T G.729 / CS-ACELP (*Conjugate Structure-Algebraic Code-Excited Linear Prediction*) – 1995 – 8 kbit/s – IP
- TIA (*Telecommunications Industry Association*) – IS-54 / VSELP (*Vector Sum Excited Linear Predictive*) - 1992 – 8 kbit/s – Cel USA TDMA (*Time Division Multiple Access*)
- VSELP – Japão – 1993 – 6,7 kbit/s – Cel Japonês TDMA
- TIA – IS-95 / QCELP (*Qualcomm Code-Excited Linear Predictive*) – 1993 – 1 / 2 / 4 ou 8 kbit/s – Cel USA CDMA (*Code Division Multiple Access*)
- TIA – IS-96-A / QCELP – 1995 – 1,2 / 2,4 / 4,8 / 9,6 kbit/s – Cel USA CDMA
- GSM (*Global System for Mobile Communication*) – HR (*Half Rate*) / VSELP – 1995 – 5,6 kbit/s – Cel Europeu GSM
- GSM – EFR (*Enhanced Full Rate*) / ACELP – 1997 – 12,2 kbit/s – Cel Europeu GSM
- GSM – AMR-NB (Adaptive Multi-Rate Narrowband)/ ACELP – 1998 – 4,75 / 5,15 / 5,90 / 6,70 / 7,40 / 7,95 / 10,2 / 12.2 kbit/s – Cel Europeu GSM

- TIA – IS-641 (substitui o IS-54) / ACELP – 1997 – 7,4 kbit/s – Cel USA TDMA
- TIA – IS-733 / QCELP – 1998 – 1,8 / 3,6 / 7,8 / 14,4 kbit/s – Cel USA CDMA
- TIA – IS-127 EVRC (*Enhanced Variable Rate Coder*) / ACELP – 1998 – 1,2 / 4,8 / 9,6 kbit/s – Cel USA CDMA
- AMR-WB (*Adaptive Multi-Rate – Wideband*) / ACELP – 2001 – 6,6 / 8,85 / 12,65 / 14,25 / 15,85 / 18,25 / 19,85 / 23,05 / 23,85 kbit/s – Cel 3G Europa, Japão, USA, Coréia – WCDMA e IP

Destes codificadores apenas 3 serão apresentados neste capítulo, os *codecs* de voz padrão ITU-T G.723.1 [39], AMR-NB [40] e AMR-WB [41], pois são os codificadores mais utilizados pela industria de telecomunicações nos dias atuais nas redes em analise nesta tese. O codificador ITU-T G.723.1 é um codificador de voz padronizado pela ITU (*International Telecommunication Union*) para utilização em redes IP. Já os codificadores AMR-NB e AMR-WB são codificadores de voz padronizados pelo 3GPP (*3rd Generation Partnership Project*) para utilização em redes móveis celulares, sendo que o AMR-WB foi também padronizado pelo ITU para uso em redes IP sobre a norma ITU-T G722.2 [42].

Na seção 3.1 deste capítulo será feita a apresentação do *codec* ITU-T G.723.1. Na seção 3.2 será descrito o *codec* AMR-NB e na seção 3.3 o *codec* AMR-WB. Finalmente, a seção 3.4 contém uma breve conclusão.

3.1. ITU-T G.723.1

O codificador ITU-T G.723.1 é um dos mais usados padrões para redes IP. O mesmo está presente em diversos produtos de grandes fabricantes e operadoras de telecomunicações. O *codec* ITU-T G.723.1 permite a codificação de voz a taxas de 6,3 kbit/s ou 5,3 kbit/s [39]. A taxa mais elevada fornece uma voz de melhor qualidade, porém a taxa mais baixa também fornece uma boa qualidade de voz. A diferença entre essas taxas resulta do tipo de excitação a ser utilizada e transmitida para o decodificador. Na taxa de 6,3 kbit/s, o codificador utiliza para a

excitação o MP-MLQ (*Multi-pulse Maximum Likelihood Quantization*), enquanto que na taxa de 5,3 kbit/s é empregado o ACELP (*Algebraic Code-Excited Linear Prediction*).

Este codificador é projetado para operar com um sinal digital, obtido primeiramente filtrando o sinal analógico de entrada com um filtro para telefonia (Recomendação ITU-T G.712 [43]), seguido de amostragem à taxa de 8 kHz e conversão para um PCM (*Pulse-Code Modulation*) de 16 bits, o qual será a entrada do codificador. A saída do decodificador deve ser convertida novamente para analógico, de forma similar.

O codificador opera sobre quadros de 240 amostras cada, o que equivale a 30 ms a uma taxa de amostragem de 8 kHz. O sinal sofre uma filtragem passa-altas a fim de remover a componente DC (*Direct Current*) e, em seguida é dividido em 4 sub-quadros de 60 amostras cada. Para todo sub-quadro é realizada uma análise LPC de ordem 10. Os parâmetros LPC do último sub-quadro são quantizados usando um quantizador PSVQ (*Predictive Split Vector Quantizer*), fazendo com que as LSFs sejam codificadas e transmitidas a cada 30 ms. Os demais parâmetros LPC dos outros sub-quadros em conjunto com o LPC do último sub-quadro, serão utilizados apenas para direcionar a busca da excitação de forma a considerar propriedades psicoacústicas [39].

O diagrama esquemático do codificador é apresentado na Fig. 3.1, onde se podem observar seus blocos básicos, bem como sua complexidade estrutural, a qual implica também em um grande consumo de recursos do terminal do usuário.

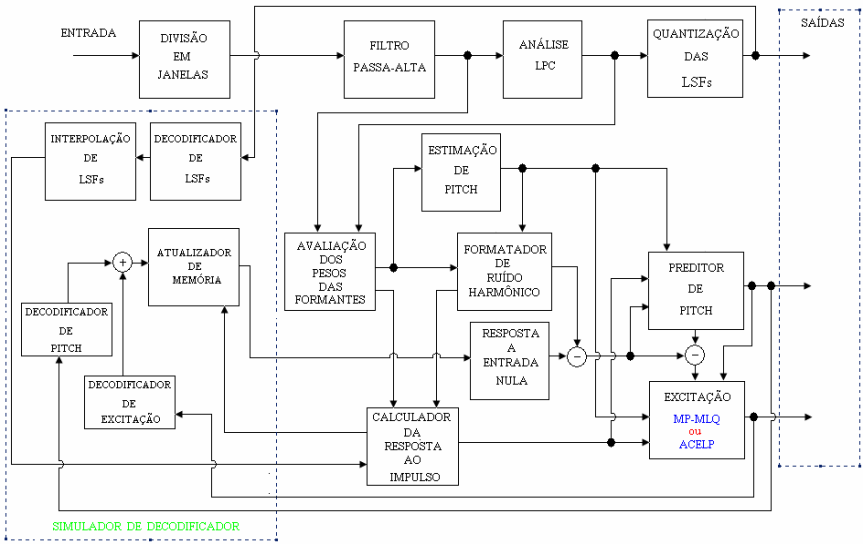

Figura 3.1 – Diagrama de blocos do codificador de voz do ITU-T G.723.1

Outra informação bastante importante sobre o codificador são as máscaras de alocação de bits utilizadas por ambas as taxas (define como serão utilizados os bits disponíveis em cada taxa de codificação) [39], apresentadas na Tab. 3.1, Essa informação dará subsídio a algumas afirmações sobre o sistema de reconhecimento distribuído a ser analisado quando se utiliza o codificador ITU-T G.723.1.

Modo	Parâmetro	1 - subquadro	2 - subquadro	3 - subquadro	4 -subquadro	total por quadro
6,3 kbit/s	LPC					24
	Dicionário Adaptativo	7	2	7	2	18
	Ganhos	12	12	12	12	48
	Posição dos Pulsos	20	18	20	18	73
	Sinais dos Pulsos	6	5	6	5	22
	Índice do Grid	1	1	1	1	4
	Total					189
5,3 kbit/s	LPC					24
	Dicionário Adaptativo	7	2	7	2	18
	Ganhos	12	12	12	12	48
	Posição dos Pulsos	12	12	12	12	48
	Sinais dos Pulsos	4	4	4	4	16
	Índice do Grid	1	1	1	1	4
	Total					158

Tabela 3.1 – Tabela de alocação de bits para o codificador ITU-T G.723.1

Vale ressaltar que como a taxa de transmissão das LSF e a precisão com que as mesmas são transmitidas (número de bits por quadro) são comuns tanto ao codificador funcionando a 5,3 ou 6,3 kbit/s, os resultados para os atributos de reconhecimento que dependam apenas das LSFs quantizadas ou dos parâmetros LPC recuperados das LSFs quantizadas não terão seu resultado de reconhecimento afetado pela variação da taxa de operação do codificador. Já no caso da voz reconstruída, deverá haver uma maior degradação do desempenho dos atributos obtidos através da mesma, para a taxa de 5,3 kbit/s.

A estrutura do decodificador apresentada pela norma ITU-T G.723.1 é aqui ilustrada na Fig. 3.2. A operação de decodificação é realizada quadro a quadro, primeiramente os índices dos parâmetros LSF quantizados são decodificados, sendo consecutivamente interpolados e usados para construir o filtro de síntese. Para cada sub-quadro, a excitação e o pitch são decodificados, sendo os mesmos combinados para gerar a entrada do filtro de síntese, que terá como saída a voz reconstruída a ser ainda filtrada pelo filtro de formantes, que tem como objetivo reduzir o ruído de quantização enfatizando as frequências das formantes e reduzindo os vales presentes no espectro do sinal, sendo sua forma apresentada em [39].

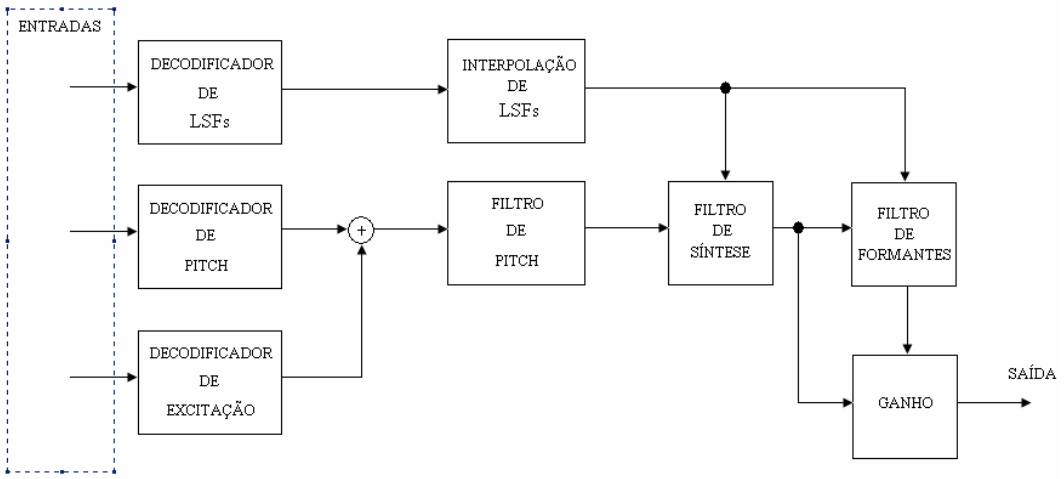

Figura 3.2 – Diagrama de blocos do decodificador de voz do ITU-T G.723.1

O *codec* utiliza também um detector de atividade da voz (VAD - *Voice Activity Detection*). Este detector de atividade decide se cada quadro é composto por voz ou por silêncio com base na energia do sinal amostrado. Os trechos de silêncio no discurso do emissor são codificados a uma taxa denominada SID (*Silence Descriptor*) que reproduz as características do silêncio produzindo o chamado "ruído de conforto" [44].

Nesta tese foi utilizado o código de referência para a aritmética de ponto flutuante definido em [45] para implementar o sistema de reconhecimento de voz distribuído usando o codificador ITU-T G.723.1.

3.2. Adaptive Multi-Rate Narrowband (AMR-NB)

O *codec* AMR-NB [40] é um codificador da família CELP (*Code-Excited Linear Predictive*) que é muito utilizado em sistemas celulares. Como o próprio nome sugere, ele pode operar em diversas taxas, que variam desde 4,75 até 12,2 kbit/s. Uma versão mais recente deste codec, o AMR-WB [41], permite trocas de taxas desde 6,6 até 23,85 kbit/s, o que será apresentado com mais detalhes na próxima seção deste capítulo.

Inicialmente o GSM (*Global System for Mobile Communications*) foi utilizado com os codificadores GSM-FR (*Full-Rate*) [46] e GSM-HR [47]. Com o intuito de oferecer qualidade equivalente às redes PSTN (*Public Switching*

Telecommunications Network), foi desenvolvido o GSM-EFR [48] e, em 1998, – a referência pode chamar atenção como sendo de 2008, porém pelo fato destas normas para codificadores estarem em constante revisão, foi considerada a data da revisão que gerou a versão utilizada como base para esta tese – o AMR-NB foi adotado na fase 2+ do GSM (referente à *Release* 98 – Rel'98 – de software dos sistemas que adotam o padrão GSM). Atualmente o AMR-WB (*Adaptative Multi-Rate Wideband*) é recomendado pelo 3GPP como padrão a ser utilizado em aplicações e serviços de terceira geração definidos pelo IMT-2000 (*International Mobile Telecommunications - 2000*). Até a adoção do AMR-NB, a maioria dos padrões de codificação de voz utilizava taxas fixas.

O AMR-NB fornece 8 taxas de codificação conforme mostrado na Tab. 3.2. Ele também implementa um detector de atividade da voz (VAD)[49]. Este detector de atividade decide se cada quadro é composto por voz ou por silêncio com base na energia do sinal amostrado. Os trechos de silêncio no discurso do emissor são codificados a uma taxa denominada SID (*Silence Descriptor*) que reproduz as características do silêncio, produzindo o chamado "ruído de conforto" [50]. Além disso, o AMR-NB possui mecanismo de substituição e silenciamento de quadros perdidos que diminui os efeitos da perda de pacotes na rede [51]. A finalidade da substituição é atenuar e ocultar o efeito dos quadros perdidos. A finalidade de silenciar a saída, no caso de muitos quadros perdidos, é indicar a interrupção do canal ao usuário e evitar a geração de possíveis sons inoportunos como um resultado do procedimento de substituição de quadro. Para aumentar a qualidade subjetiva, quadros de fala perdidos são substituídos tanto por repetição como por extrapolação de bom(ns) quadro(s) de fala anterior(es). Esta substituição é feita de forma que o nível de saída diminua gradualmente, resultando em silêncio a partir de um determinado número de quadros perdidos [51].

Modo	Taxa (kbps)	Classes de bits			Total
		A	B	C	
AMR-NB_4.75	4.75	42	53	0	95
AMR-NB_5.15	5.15	49	54	0	103
AMR-NB_5.90	5.90	55	63	0	118
AMR-NB_6.70	6.70	58	76	0	134
AMR-NB_7.40	7.40	61	87	0	148
AMR-NB_7.95	7.95	75	84	0	159
AMR-NB_10.2	10.2	65	99	40	204
AMR-NB_12.2	12.2	81	103	60	244
SID	1.80	39	0	0	39

Tabela 3.2 – Taxa de codificação do AMR-NB e alocação de bits nas classes

Ao receber o sinal de fala, o codificador AMR-NB faz a amostragem do sinal a uma taxa de 8 kHz para gerar quadros de 20 ms (correspondendo a 160 amostras). Cada quadro de 20 ms de voz produz, 95, 103, 118, 134, 148, 159, 204 ou 244 bits de informação dependendo da taxa de codificação utilizada. Após a codificação da voz, os bits são separados em três categorias (A, B e C) conforme a sua importância. Durante a codificação do canal, tais bits são protegidos de acordo com a importância que lhe foi atribuída (codificação de canal mais poderosa ou menos poderosa). No entanto, modernamente a maioria dos equipamentos decodificadores continua a decodificação dos bits menos significantes (Classes B e C), mesmo que tenham sido detectados, previamente, erros severos nos bits de Classe A. Parte-se do princípio de que é mais conveniente tentar-se recuperar a inteligibilidade da voz a partir de quadros com erro do que a partir de quadros que tenham sido suprimidos [52].

O fluxo completo de sinal no codificador é mostrado na Fig. 3.3. A análise LPC é executada duas vezes por quadro para o modo AMR-NB 12.20 e uma vez para os demais modos. Para o modo AMR-NB 12.20, os dois conjuntos de parâmetros LPC são convertidos para LSFs (*Line Spectral Frequencies*) e são conjuntamente quantizados usando-se SMQ (*Split Matrix Quantization*) com 38 bits. Para os outros modos, o único conjunto de parâmetros LPC é convertido para LSFs e é quantizado vetorialmente usando-se SVQ (*Split Vector Quantization*). Os LSFs são uma representação dos parâmetros LPC no domínio da frequência. Detalhes da conversão dos parâmetros LPC para a representação LSF podem ser encontrados em [53]. O quadro de fala é dividido em 4 sub-quadros de 5 ms cada (40 amostras). Os parâmetros do dicionário adaptativo e fixo são transmitidos na cadência de sub-quadro. Os parâmetros LSFs quantizados e não quantizados ou suas versões interpoladas são usados dependendo do sub-quadro. Um período de pitch em malha aberta é estimado em todos os sub-quadros (exceto para os modos AMR 5.15 e AMR 4.75, nos quais isto é feito uma vez por quadro), baseado no sinal de fala ponderado perceptualmente. Esta estimativa é realizada para simplificar a análise do pitch e para realizar a busca em malha fechada utilizando valores próximos ao que foi estimado em malha aberta [54].

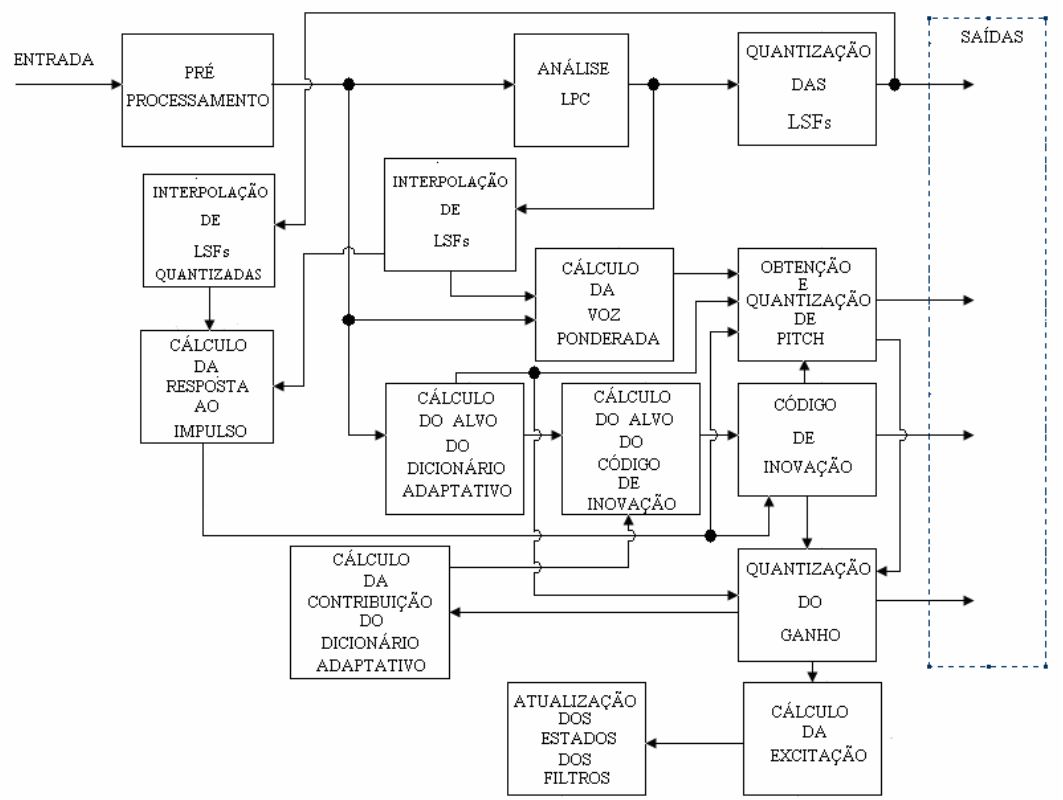

Figura 3.3 – Diagrama de blocos do codificador de voz do AMR-NB

Outra informação bastante importante sobre o codificador são as máscaras de alocação de bits utilizadas nas diversas taxas [54], apresentadas na Tab. 3.3. Essa informação dará subsídio a algumas afirmações sobre o sistema de reconhecimento distribuído a ser analisado quando se utiliza o codificador AMR-NB.

Modo	Parâmetro	1 - subquadro	2 - subquadro	3 - subquadro	4 -subquadro	total por quadro
12.2 kbit/s (GSM EFR)	2 conj. LSF					38
	Retardo de Pitch	9	6	9	6	30
	Ganho de Pitch	4	4	4	4	16
	Código Algébrico	35	35	35	35	140
	Ganho de Código	5	5	5	5	20
	Total					244
10.2 kbit/s	conj. LSF					26
	Retardo de Pitch	8	5	8	5	26
	Código Algébrico	31	31	31	31	124
	Ganhos	7	7	7	7	28
	Total					204
7.95 kbit/s	conj. LSF					27
	Retardo de Pitch	8	6	8	6	28
	Ganho de Pitch	4	4	4	4	16
	Código Algébrico	17	17	17	17	68
	Ganho de Código	5	5	5	5	20
	Total					159
7.40 kbit/s (TDMA EFR)	conj. LSF					26
	Retardo de Pitch	8	5	8	5	26
	Código Algébrico	17	17	17	17	68
	Ganhos	7	7	7	7	28
	Total					148
6.70 kbit/s (PDC EFR)	conj. LSF					26
	Retardo de Pitch	8	4	8	4	24
	Código Algébrico	14	14	14	14	56
	Ganhos	7	7	7	7	28
	Total					134
5.90 kbit/s	conj. LSF					26
	Retardo de Pitch	8	4	8	4	24
	Código Algébrico	11	11	11	11	44
	Ganhos	6	6	6	6	24
	Total					118
5.15 kbit/s	conj. LSF					23
	Retardo de Pitch	8	4	4	4	20
	Código Algébrico	9	9	9	9	36
	Ganhos	6	6	6	6	24
	Total					103
4.75 kbit/s	conj. LSF					23
	Retardo de Pitch	8	4	4	4	20
	Código Algébrico	9	9	9	9	36
	Ganhos	8		8		16
	Total					95

Tabela 3.3 – Tabela de alocação de bits para o codificador AMR-NB

O fluxo do sinal no decodificador é mostrado na Fig. 3.4. Os índices transmitidos são extraídos do fluxo de bits recebido e, então, decodificados para obtenção dos parâmetros do codificador em cada quadro transmitido. Estes parâmetros são os vetores LSF, os períodos de pitch fracionários, os vetores de código de inovação, e os respectivos ganhos de pitch e de inovação. Os vetores LSF são convertidos para os coeficientes LPC do filtro e interpolados para obter filtros LPC em cada sub-quadro. Então, a cada 40 amostras de sub-quadro:

- a excitação é construída adicionando-se os vetores de código adaptativo e de inovação escalados pelos seus respectivos ganhos;
- a fala é reconstruída filtrando-se a excitação através do filtro LPC de síntese. Finalmente, o sinal de fala reconstruído passa por um filtro de formantes que tem como objetivo reduzir o ruído de quantização enfatizando as frequências das formantes e reduzindo os vales presentes no espectro do sinal, sendo sua forma apresentado em [40].

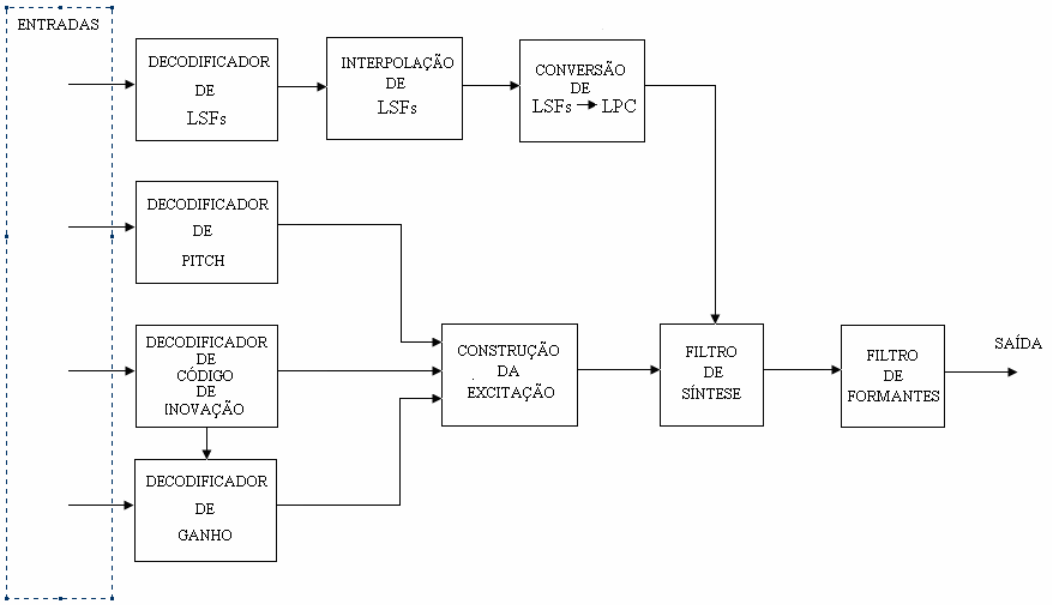

Figura 3.4 – Diagrama de blocos do decodificador de voz do AMR-NB

Nesta tese foi utilizado o código de referência para a aritmética de ponto flutuante definido em [55] para implementar o sistema de reconhecimento de voz distribuído usando o codificador AMR-NB. Porém, é possível também implementar o sistema utilizando o código do mesmo codificador utilizando a aritmética de ponto fixo, conforme definido em [56] quando o processador utilizado, ou o sistema operacional, não for compatível com a aritmética de ponto flutuante.

3.3. Adaptive Multi-Rate Wideband (AMR-WB)

O *codec* AMR-WB [41] é um codificador da família CELP (*Code-Excited Linear Predictive*) que já é utilizado em sistemas celulares (dependerá apenas do

modelo do aparelho em uso pelo usuário – já existem modelos no mercado com esta funcionalidade e da cobertura da operadora – já existem locais onde este codificador já é suportado), porém também é utilizado para redes IP [42], principalmente para a codificação de áudio para vídeo conferências. Como o próprio nome sugere, ele pode operar em diversas taxas, que variam desde 6,6 até 23,85 kbit/s.

O AMR-WB fornece 9 taxas de codificação conforme mostrado na Tab. 3.4. Ele também utiliza um detector de atividade da voz (VAD) [57]. Este detector de atividade decide se cada quadro é composto por voz ou por silêncio com base na energia do sinal amostrado. Os trechos de silêncio no discurso do emissor são codificados a uma taxa denominada SID (*Silence Descriptor*) que reproduz as características do silêncio produzindo o chamado "ruído de conforto" [58]. Além disso, o AMR possui mecanismo de substituição e silenciamento de quadros perdidos que diminui os efeitos da perda de pacotes na rede [59].

Modo	Taxa (kbps)	Classes de bits		Total
		A	B	
AMR-WB_6.60	6.60	54	78	132
AMR-WB_8.85	8.85	64	113	177
AMR-WB_12.65	12.65	72	181	253
AMR-WB_14.25	14.25	72	213	285
AMR-WB_15.85	15.85	72	245	317
AMR-WB_18.25	18.25	72	293	365
AMR-WB_19.85	19.85	72	325	397
AMR-WB_23.05	23.05	72	389	461
AMR-WB_23.85	23.85	72	405	477
SID	1.75	40	0	40

Tabela 3.4 – Taxa de codificação do AMR-WB e alocação de bits nas classes

Após a geração da fala, o codificador AMR-WB faz a amostragem do sinal a uma taxa de 16 kHz para gerar quadros de 20 ms (correspondendo a 320 amostras). Cada quadro de 20 ms de voz produz 132, 177, 253, 285, 317, 365, 397, 461, 477 bits de informação dependendo da taxa de codificação utilizada. Após a codificação da voz, os bits são separados em duas categorias (A, B) conforme a sua importância. Podemos observar, então, que diferentemente do AMR-NB, o AMR-WB não possui bits categorizados como de classe C. Durante a codificação do canal, tais bits são protegidos de acordo com a importância que lhe foi atribuída (codificação de canal mais poderosa ou menos poderosa). No entanto, modernamente a maioria dos equipamentos decodificadores continua a

decodificação dos bits menos significantes (Classe B), mesmo que tenham sido detectados, previamente, erros severos nos bits de Classe A. Parte-se do princípio de que é mais conveniente tentar-se recuperar a inteligibilidade da voz a partir de quadros com erro do que a partir de quadros que tenham sido suprimidos [52].

O fluxo completo de sinal no codificador é mostrado na Fig. 3.5. Depois da sub-amostragem (decimação), filtragem passa-altas e pré-ênfase, a análise LPC é realizada uma vez por quadro. O conjunto de parâmetros LPC é convertido para os ISFs (*Immittance Spectrum Frequencies*) e quantizados vetorialmente usando-se S-MSVQ (*Split-Multistage Vector Quantization*). Detalhes da conversão dos parâmetros LPC para a representação ISF podem ser encontrados em [60]. No capítulo 4 será explicado porque se optou neste codificador pela conversão dos LPC para ISF e não para LSF, apresentando as manipulações matemáticas para esta conversão LPC para ISF. O quadro de fala é dividido em 4 sub-quadros de 5 ms cada (64 amostras com taxa de amostragem de 12,8 kHz). Os parâmetros do dicionário adaptativo e fixo são transmitidos na cadência de sub-quadro. Os parâmetros LPC quantizados e não quantizados ou suas versões interpoladas são usados dependendo do sub-quadro. Um período de pitch em malha aberta é estimado em todos os sub-quadros ou uma vez por quadro baseado no sinal de fala ponderado perceptualmente. Esta estimativa é realizada para simplificar a análise do pitch e para realizar a busca em malha fechada utilizando valores próximos ao que foi estimado em malha aberta [61]. Em adição a estes parâmetros, os índices de ganho da banda alta são computados para o modo de 23.85 kbit/s.

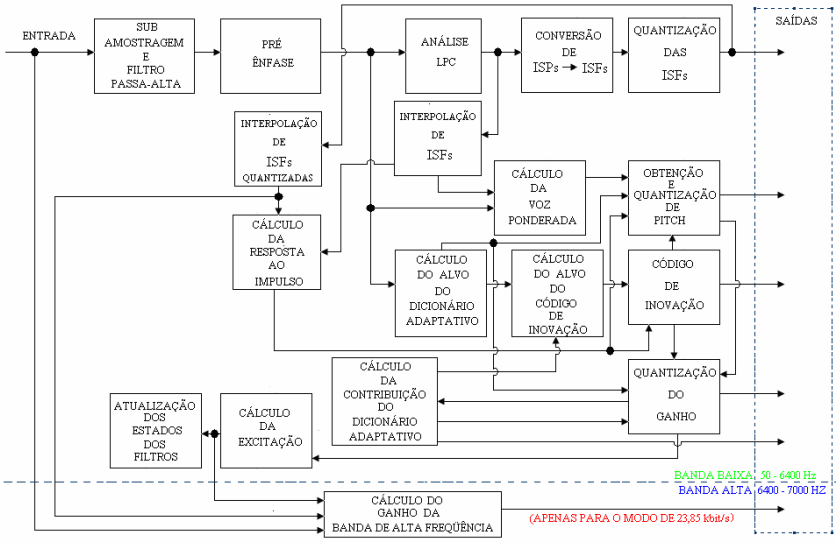

Figura 3.5 – Diagrama de blocos do codificador de voz do AMR-WB

Outra informação bastante importante sobre o codificador são as máscaras de alocação de bits utilizadas nas diversas taxas [61], apresentados na Tab. 3.5. Essa informação dará subsídio a algumas afirmações sobre o sistema de reconhecimento distribuído a ser analisado quando se utiliza o codificador AMR-WB.

Modo	Parâmetro	1 - subquadro	2 - subquadro	3 - subquadro	4 -subquadro	total por quadro
23.85 kbit/s	VAD-flag					1
	ISF					46
	Filtro LTP	1	1	1	1	4
	Retardo de Pitch	9	6	9	6	30
	Código Algébrico	88	88	88	88	352
	Ganho de Código	7	7	7	7	28
	Energia HB	4	4	4	4	16
	Total					477
23.05 kbit/s	VAD-flag					1
	ISF					46
	Filtro LTP	1	1	1	1	4
	Retardo de Pitch	9	6	9	6	30
	Código Algébrico	88	88	88	88	352
	Ganhos	7	7	7	7	28
	Total					461
19.85 kbit/s	VAD-flag					1
	ISF					46
	Filtro LTP	1	1	1	1	4
	Retardo de Pitch	9	6	9	6	30
	Código Algébrico	72	72	72	72	288
	Ganho de Código	7	7	7	7	28
	Total					397
18.25 kbit/s	VAD-flag					1
	ISF					46
	Filtro LTP	1	1	1	1	4
	Retardo de Pitch	9	6	9	6	30
	Código Algébrico	64	64	64	64	256
	Ganhos	7	7	7	7	28
	Total					365
15.85 kbit/s	VAD-flag					1
	ISF					46
	Filtro LTP	1	1	1	1	4
	Retardo de Pitch	9	6	9	6	30
	Código Algébrico	52	52	52	52	208
	Ganhos	7	7	7	7	28
	Total					317
14.25 kbit/s	VAD-flag					1
	ISF					46
	Filtro LTP	1	1	1	1	4
	Retardo de Pitch	9	6	9	6	30
	Código Algébrico	44	44	44	44	176
	Ganhos	7	7	7	7	28
	Total					285
12.65 kbit/s	VAD-flag					1
	ISF					46
	Filtro LTP	1	1	1	1	4
	Retardo de Pitch	9	6	9	6	30
	Código Algébrico	36	36	36	36	144
	Ganhos	7	7	7	7	28
	Total					253
8.85 kbit/s	VAD-flag					1
	ISF					46
	Retardo de Pitch	8	5	8	5	26
	Código Algébrico	20	20	20	20	80
	Ganhos	6	6	6	6	24
	Total					177
6.60 kbit/s	VAD-flag					1
	ISF					36
	Retardo de Pitch	8	5	5	5	23
	Código Algébrico	12	12	12	12	48
	Ganhos	6	6	6	6	24
	Total					132

Tabela 3.5 – Tabela de alocação de bits para o codificador AMR-WB

O fluxo do sinal no decodificador é mostrado na Fig. 3.6. Os índices transmitidos são extraídos do fluxo de bits recebido e, então, decodificados para obtenção dos parâmetros do codificador em cada quadro transmitido. Estes parâmetros são os vetores ISF, os períodos de pitch fracionários, os vetores de código de inovação, os respectivos ganhos de pitch e de inovação e os parâmetros do preditor de período longo (os quais são realizados para a banda baixa de 50 – 6400 Hz). No modo 23.85 kbit/s também são decodificados os índices da banda alta (6400 – 7000 Hz). Os vetores ISF são convertidos para os coeficientes LPC do filtro e interpolados para obter filtros LPC em cada sub-quadro. Então, a cada 64 amostras de sub-quadro:

- a excitação é construída adicionando-se os vetores de código adaptativo e de inovação escalados pelos seus respectivos ganhos;
- a fala é reconstruída filtrando-se a excitação através do filtro LPC de síntese;
- A voz reconstruída passa por um filtro de de-ênfase.

Finalmente, o sinal de fala da banda baixa (50 – 6400 Hz) reconstruído é sobre-amostrado para 16 kHz e o sinal de voz da banda alta é filtrado em (6400 – 7000 Hz) na taxa de 16kHz e adicionado à banda baixa.

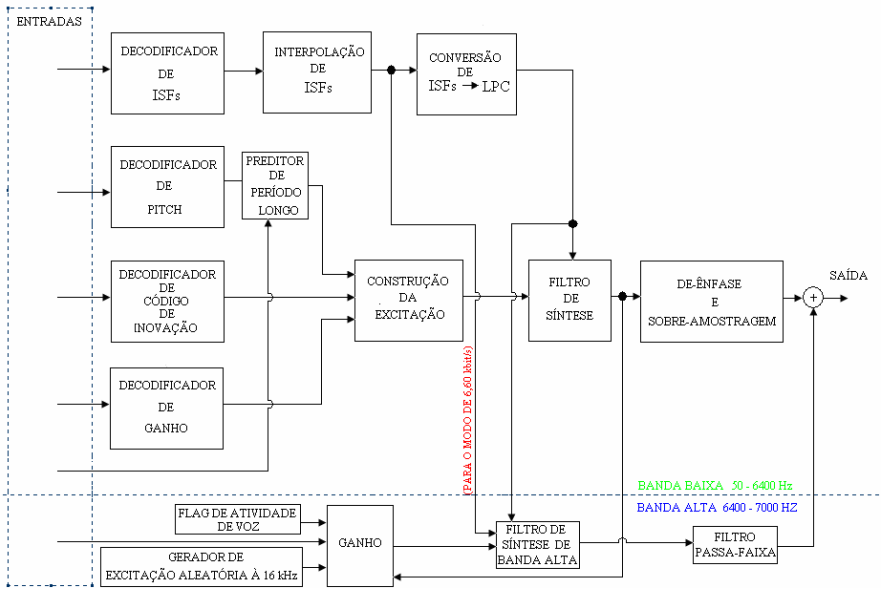

Figura 3.6 – Diagrama de blocos do decodificador de voz do AMR-WB

O código de referência para o codificador AMR-WB na aritmética de ponto flutuante foi definido em [62], porém a norma também implementa o codificador utilizando a aritmética de ponto fixo, conforme definido em [63] para quando o processador utilizado, ou o sistema operacional, não for compatível com a aritmética de ponto flutuante.

3.4. Conclusão

Neste capítulo foi feita a apresentação dos codificadores de voz que serão utilizados na montagem dos sistemas de reconhecimento distribuído no ambiente celular/voz sobre IP.

No capítulo seguinte, será feita a apresentação teórica dos atributos de reconhecimento que serão utilizados para a implementação do sistema de reconhecimento de voz distribuído para o Português Brasileiro com amplo vocabulário.

4
Atributos para Reconhecimento de Voz Distribuído

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 [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 [1]. Portanto, transformações desses parâmetros são estratégias importantes a serem consideradas.

Neste capítulo serão apresentadas as deduções matemáticas dos atributos de reconhecimento de voz distribuído a serem utilizados nesta tese.

Na seção 4.1 deste capítulo será feita a apresentação dos atributos obtidos a partir dos parâmetros LPC do decodificador e na seção 4.2 serão apresentados os atributos obtidos dos parâmetros LSF. A seção 4.3 descreve os parâmetros ISF. A seção 4.4 é dedicada ao atributo mais amplamente usado, obtido da voz reconstruída pelos codificadores (MFCC - *Mel-Frequency Cepstral Coefficients*) . Finalmente, a seção 4.5 contém uma breve conclusão.

4.1.
Atributos Extraídos de LPCs

Nesta seção são analisados os parâmetros de reconhecimento que podem ser extraídos diretamente dos parâmetros LPC (*Linear Predictive Coefficients*), sem a necessidade de reconstrução do sinal de voz para obtenção dos atributos. Esta abordagem se deve ao fato de que, dentro dos decodificadores de voz utilizados para telefonia celular e voz sobre IP, já serem produzidos naturalmente, no seu processo de recuperação de voz, os parâmetros LPC, em um estágio anterior à reconstrução da voz. Sendo assim, parâmetros de reconhecimento de voz, obtidos neste estágio, são menos complexos computacionalmente do que os obtidos de voz reconstruída, pois evitam a necessidade de recuperação da mesma. Além disso, a MFCC de voz reconstruída apresenta resultado pior [6].

Primeiramente, será feita, na seção 4.1.1, uma apresentação matemática do método de obtenção dos parâmetros LPC que serão a base dos atributos de reconhecimento apresentados nas seções 4.1.2 e 4.1.3.

4.1.1.
Linear Predictive Coding (LPC)

A idéia básica da análise LPC consiste em uma amostra do sinal de fala ser modelada por uma combinação linear de suas p amostras passadas, dada por

$$s(n) \approx a_1 s(n-1) + a_2 s(n-2) + \ldots + a_p s(n-p) \tag{4.1}$$

onde os coeficientes $a_1, a_2, \ldots, a_p$ são recalculados para cada janela do sinal, pois, em pequenos trechos, o sinal pode ser assumido como sendo estacionário.

A equação anterior pode ser convertida em uma igualdade, incluindo um termo de excitação do sinal, $Gu(n)$, onde $u(n)$ é a excitação normalizada e G é o seu ganho.

$$s(n) = Gu(n) + \sum_{i=1}^{p} a_i s(n-i) \tag{4.2}$$

Isso nos leva a uma função de transferência do trato vocal

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{i=1}^{p} a_i z^{-i}} \tag{4.3}$$

Os coeficientes $a_1, a_2, \ldots, a_p$ são os parâmetros LPC do sinal. Eles são calculados considerando-se a predição linear dada por

$$\tilde{s}(n) = \sum_{i=1}^{p} a_i s(n-i) \tag{4.4}$$

e o seu erro de predição é

$$e(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{i=1}^{p} a_i s(n-i) \tag{4.5}$$

Os coeficientes são escolhidos a fim de minimizar uma função do erro de predição. Para isso, dentro de uma janela de sinal de tamanho N, o erro médio quadrático definido por

$$E_l = \sum_{n=0}^{N-1} (e(n))^2 = \sum_{n=0}^{N-1} \left(s(n) - \sum_{k=1}^{p} a_k s(n-k) \right)^2 \tag{4.6}$$

deve ser derivado em função de cada coeficiente a_i e igualado a zero

$$\frac{\partial E_l}{\partial a_i} = 0, \qquad i = 1,2,\ldots,p \tag{4.7}$$

onde l é o índice do segmento considerado.

Obtendo

$$\sum_{n=0}^{N-1} s(n-i)s(n) = \sum_{k=1}^{p} a_k \left(\sum_{n=0}^{N-1} s(n-i)s(n-k) \right), \qquad i = 1,2,\ldots,p \tag{4.8}$$

Pode-se definir os coeficientes de correlação como

$$\varphi_l(i,k) = \sum_{n=0}^{N-1} s(n-i)s(n-k) \tag{4.9}$$

e então

$$\sum_{k=1}^{p} a_k \varphi_l(i,k) = \varphi_l(i,0), \qquad i = 1,2,\ldots,p \tag{4.10}$$

A solução de p equações lineares resulta em p coeficientes LPC que minimizam o erro de predição. Com a_i satisfazendo a (4.10), o erro de predição total em (4.6) assume o seguinte valor

$$E_l = \sum_{n=0}^{N-1} s^2(n) - \sum_{k=1}^{p} a_k \sum_{n=0}^{N-1} s(n)s(n-k) = \varphi(0,0) - \sum_{k=1}^{p} a_k \varphi(0,k) \tag{4.11}$$

Com uma simples substituição de variáveis, (4.9) pode ser reescrita como

$$\varphi_l(i,k) = \sum_{n=-i}^{N-1-i} s(n)s(n+i-k) = \sum_{n=-k}^{N-1-k} s(n)s(n+k-i) \tag{4.12}$$

Como o sinal é processado em janelas de duração finita $(0 \le n \le N-1)$, sendo o sinal zero fora da janela, os limites do somatório podem ser alterados

$$\varphi_l(i,k) = \sum_{n=0}^{N-1-(i-k)} s(n)s(n+i-k) = \sum_{n=0}^{N-1-(k-i)} s(n)s(n+k-i) \equiv r(|i-k|) \tag{4.13}$$

Com a alteração dos limites do somatório em (4.13) temos a autocorrelação do sinal. Neste caso, a equação (4.10) torna-se

$$\sum_{k=1}^{p} a_k r(|i-k|) = r(i), \qquad i = 1,2,\ldots,p \tag{4.14}$$

Este é chamado de método da autocorrelação e é utilizado pelos codificadores aqui utilizados. O sistema de equações pode ser visto na sua forma matricial

$$\begin{bmatrix} r(0) & r(1) & r(2) & \cdots & r(p-1) \\ r(1) & r(0) & r(1) & \cdots & r(p-2) \\ r(2) & r(1) & r(0) & \cdots & r(p-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r(p-1) & r(p-2) & r(p-3) & \cdots & r(0) \end{bmatrix} \begin{bmatrix} a(1) \\ a(2) \\ a(3) \\ \vdots \\ a(p) \end{bmatrix} = \begin{bmatrix} r(1) \\ r(2) \\ r(3) \\ \vdots \\ r(p) \end{bmatrix} \tag{4.15}$$

Como a matriz é do tipo Toeplitz [64-66], o melhor método para resolvê-la é utilizar o algoritmo de Levinson-Durbin [64-66], que também é utilizado pelos codificadores para resolver o sistema de equações, e é dado por

$$\begin{aligned} &\textit{Valores iniciais}: E^{(0)} = r(0), \qquad k_0 = 0 \\ &\textit{Iteração}: \quad 1 \le i \le p \\ &E^{(i)} = \left(1 - k_{i-1}^2\right) E^{(i-1)} \\ &k_i = \left\{ r(i) - \sum_{j=1}^{i-1} \alpha_j^{(i-1)} r(|i-j|) \right\} \Big/ E^{(i)} \\ &\alpha_i^{(i)} = k_i \\ &\alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)}, \qquad 1 \le j < i \\ &\textit{Resultado}: \quad a_i = LPC \quad \textit{coeficients} = \alpha_i^{(p)}, \qquad 1 \le i \le p \end{aligned} \tag{4.16}$$

Os parâmetros de reconhecimento que podem ser obtidos dos parâmetros LPC são os parâmetros LPCC (*LPC Cepstrum*) e MLPCC (*Mel-Frequency* LPCC). Os parâmetros LPCC serão obtidos a partir dos parâmetros LPC por uma fórmula recursiva a ser deduzida na seção 4.1.2., sendo aqui apresentados, pois são a base da obtenção dos atributos MLPCC. Porém não serão utilizados em simulações, pois têm um pior desempenho no reconhecimento do que o MLPCC

[5]. Já os parâmetros MLPCC serão obtidos dos LPCCs através de uma rede de filtros passa-tudo a ser apresentada na seção 4.1.3.

4.1.2. LPC *Cepstrum* (LPCC)

O processo de obtenção dos parâmetros LPCC a partir dos coeficientes LPC será formulado no domínio da Transformada-Z, com o cálculo da resposta ao impulso do logaritmo complexo do sistema LPC, o que é análogo ao cálculo do Cepestro no domínio da Transformada Discreta de Fourier [5].

Primeiramente, se constrói a função de transferência do sistema LPC de ordem p, que é dada por

$$H(z) = \sum_{n=0}^{+\infty} h[n] z^{-n} = \frac{G}{A(Z)} = \frac{G}{1 - \sum_{i=1}^{p} a_i z^{-i}} \tag{4.17}$$

onde a_i é o i-ésimo parâmetro LPC e G é o fator de ganho.

Calculando a derivada do polinômio complexo $\ln(H(z))$, em relação a $\rho = z^{-1}$, obtém-se

$$\frac{\partial}{\partial \rho} \ln(H(\rho)) = \frac{\partial}{\partial \rho} [\ln(G) - \ln(A(\rho))] = \frac{\sum_{l=1}^{p} l a_l \rho^{l-1}}{1 - \sum_{i=1}^{p} a_i \rho^{i}} \tag{4.18}$$

Como $H(z)$ é a função de transferência do sistema LPC obtido no transmissor, onde são utilizados métodos para garantir a estabilidade da função $H(z)$, a mesma deverá ter todos os seus pólos dentro do círculo unitário, então $\ln(H(z))$ é unilateral, o que leva a escrever

$$C(z) = \sum_{i=0}^{+\infty} c_i z^{-i} \tag{4.19}$$

onde c_i é o i-ésimo parâmetro LPCC e $C(z)$ é o logaritmo complexo da função de transferência do sistema LPC.

Derivando $C(z)$ em relação a ρ e igualando a (4.18), obtém-se a equação

$$\sum_{j=1}^{+\infty} jc_j \rho^{j-1} = \frac{\sum_{l=1}^{p} la_l \rho^{l-1}}{1 - \sum_{i=1}^{p} a_i \rho^i} \tag{4.20}$$

que pode ser reescrita na forma

$$\left(\sum_{j=1}^{+\infty} jc_j \rho^{j-1} \right) \left(1 - \sum_{i=1}^{p} a_i \rho^i \right) = \sum_{l=1}^{p} la_l \rho^{l-1} \tag{4.21}$$

Comparando os coeficientes das séries de ρ em ambos os lados, chega-se a uma equação recursiva que permite a obtenção dos parâmetros LPCC, onde o parâmetro c_0 é determinado pelo termo constante da definição original de $H(z)$. Essa equação é dada por

$$c_i = \begin{cases} \ln(G) & i = 0 \\ a_1 & i = 1 \\ a_i + \sum_{j=1}^{i-1} \frac{i-j}{i} c_{i-j} a_j & 1 < i \le p \\ \sum_{j=1}^{p} \frac{i-j}{i} c_{i-j} a_j & i > p \end{cases} \tag{4.22}$$

4.1.3. *Mel-Frequency* LPCC (MLPCC)

O processo de obtenção do parâmetro MLPCC passa pela transformação do eixo de frequência real para o eixo de frequência na escala mel dos parâmetros LPCC [4]. Para ser realizada esta transformação, utiliza-se um banco de n filtros passa-tudo de primeira ordem que permite efetuar a transformação do eixo de frequência real para o eixo de frequência na escala mel - onde n é o número de parâmetros LPCC obtidos através de (4.22) - [67]. Todos os filtros deste banco

terão sua função de transferência $\psi(z)$ passa-tudo de primeira ordem [68] dada pela expressão

$$\psi(z) = \frac{z^{-1} - a^*}{1 - az^{-1}} \tag{4.23}$$

devendo cada coeficiente cepestral c_i passar por um filtro diferente deste banco de filtros, onde a é o coeficiente deste filtro passa-tudo e a^* é o conjugado de a.

Como o objetivo de cada filtro é realizar a aproximação da escala mel de frequências, tem-se que analisar o que a função de transferência em (4.23) está realizando com os eixos das frequências. Para isto, será considerado a real, o que facilitará a implementação do filtro [69].

Para que seja feita a análise do que está sendo feito com os eixos de frequência, deve-se reescrever ψ, em função de $e^{j\Omega}$, como

$$\psi\left(e^{j\Omega}\right) = e^{-j\theta(\Omega)} \tag{4.24}$$

onde Ω é a frequência real e

$$\theta(\Omega) = \arctan\left[\frac{\left(1-a^2\right)\operatorname{sen}\Omega}{\left(1+a^2\right)\cos\Omega - 2a}\right] \tag{4.25}$$

é a frequência na escala mel expressa em função da frequência real Ω.

Ao se ajustar a curva de $\theta(\Omega)$ à curva da escala mel, para a frequência de amostragem de 8 kHz, por meio da variação do termo a real, obtém-se $a = 0{,}3624$ [69] e para a frequência de amostragem de 16 kHz, $a = 0{,}6$ [5].

As saídas do banco de filtros serão os parâmetros MLPCC.

4.2.
Atributos Extraídos de LSFs

As *Line spectral frequencies* (LSFs) são usualmente utilizadas para codificação de voz, devido à sua grande eficiência de codificação e suas propriedades atraentes para interpolação [70]. Porém, as LSFs não apresentam bom desempenho quando utilizadas como atributos para reconhecimento de voz [1].

Primeiramente, será desenvolvido matematicamente, na seção 4.2.1., o método de obtenção dos parâmetros LSF a partir dos parâmetros LPC, sendo este desenvolvimento justificado, pois as LSFs serão a base dos atributos de reconhecimento apresentados nas seções 4.2.2., 4.2.3., 4.2.4. e 4.2.5. e para que, quando da apresentação da obtenção dos parâmetros ISFs dos parâmetros LPC na seção 4.3., fique clara a semelhança entre os parâmetros ISF e LSF.

4.2.1.
Line Spectral Frequencies (LSF)

Os coeficientes LSF constituem uma das várias representações possíveis para os coeficientes de predição a_i do filtro de síntese utilizado na análise LPC.

Este filtro é definido por

$$H(z) = \frac{G}{A(z)} = \frac{G}{1 - \sum_{i=1}^{p} a_i z^{-i}} \tag{4.26}$$

onde $A(z)$ é o filtro inverso de ordem p.

Para o cálculo dos coeficientes LSF é necessário definir dois polinômios auxiliares $P(z)$ e $Q(z)$ obtidos a partir de $A(z)$ da seguinte forma [71]:

$$P(z) = A(z) + z^{-(p+1)} A(z^{-1}) \tag{4.27}$$

$$Q(z) = A(z) - z^{-(p+1)} A(z^{-1}) \tag{4.28}$$

onde, $P(z)$ é um polinômio simétrico e $Q(z)$ é um polinômio antissimétrico.

As raízes de $P(z)$ e $Q(z)$ determinam os coeficientes LSF. Estes polinômios possuem ligação direta com o modelo acústico do trato vocal e com os estágios do filtro preditor com estrutura em treliça.

Se $A(z)$ é de fase mínima, ou seja, se $H(z)$ é estável, então:

1. As raízes de $P(z)$ e $Q(z)$ estão sobre o círculo unitário.
2. As raízes de $P(z)$ estão alternadas com as raízes de $Q(z)$, ou seja, $r_1 < q_1 < r_2 < q_2 < \ldots < q_{p+1}$, onde r_i e q_i representam a posição angular da i-ésima raiz de $P(z)$ e $Q(z)$, respectivamente.
3. O filtro $H(z)$ continuará estável após a quantização das raízes de $P(z)$ e $Q(z)$ desde que (1) e (2) sejam respeitados pelos valores quantizados.
4. Sendo $H(z)$ estável, a mesma permanecerá estável após a interpolação.

Além disso, os LSF possuem uma faixa dinâmica bem comportada, possibilitando uma quantização mais eficiente do que as outras formas de representar os coeficientes LPC.

Finalmente, os coeficientes de predição de $A(z)$ são obtidos a partir dos coeficientes de $P(z)$ e $Q(z)$ através da seguinte igualdade polinomial:

$$A(z) = \frac{P(z) + Q(z)}{2} \tag{4.29}$$

Os coeficientes LSF apresentam a importante propriedade de robustez à distorção. De acordo com ela, qualquer alteração sofrida por um desses coeficientes não terá um efeito global. Apenas será afetada a região do espectro próxima a esta frequência. Esta propriedade pode ser explorada em sistemas de codificação da voz, uma vez que o ouvido humano não é muito sensível a variações em frequências elevadas. Nesses sistemas, é possível representar os coeficientes LSF de elevadas frequências com um menor número de bits (o que é realizado normalmente pelos codificadores de voz), o que possibilita uma diminuição da taxa de bits do sistema.

De acordo com as propriedades acima, pode-se concluir que a utilização de coeficientes LSF apresenta vantagens em relação aos coeficientes LPC em termos de transmissão, quantização e interpolação. No entanto, o cálculo direto dos coeficientes LSF exige uma elevada capacidade computacional. A alternativa é calcular os coeficientes LPC e depois transformá-los em LSF como é feito nos codificadores aqui apresentados.

A obtenção de parâmetros de reconhecimento a partir das LSFs evita a necessidade de utilização de um decodificador de voz, ou da transformação para LPC, no receptor para a realização do reconhecimento. O sistema de reconhecimento de voz distribuído que evita tal utilização se torna mais leve computacionalmente que quaisquer outros baseados em parâmetros que dependam da reconstrução da voz ou dos parâmetros LPC. Os parâmetros de reconhecimento que podem ser obtidos desta forma (diretamente de LSFs) são os parâmetros PCC (*Pseudo-Cepstral Coefficients*), PCEP (*Pseudo-Cepstrum*), MPCC (Mel-Frequency PCC) e MPCEP (Mel-Frequency PCEP). Apenas os atributos MPCC e MPCEP serão utilizados nesta tese para obter os resultados de simulação, pois já foi demonstrado que os mesmos têm melhor desempenho que os atributos PCC, PCEP [72]. Porém a apresentação da dedução matemática dos mesmos se justifica por dela derivar a dedução dos atributos na escala mel.

Cabe ressaltar que estes parâmetros, obtidos diretamente de LSF, são aproximações da obtenção dos parâmetros LPCC e MLPCC, anteriormente apresentados. Estas aproximações têm como finalidade evitar a necessidade de recuperação dos parâmetros LPC, reduzindo a complexidade computacional do sistema e, ao mesmo tempo, buscando não perder o desempenho no reconhecimento.

4.2.2. Pseudo-Cepstral Coefficients (PCC)

O parâmetro PCC é obtido diretamente de LSF, porém a sua dedução passa pela obtenção do parâmetro LPCC a partir de LPC, com manipulações matemáticas e aproximações que permitem obtê-lo diretamente de LSF sem necessitar dos parâmetros LPC. Esses procedimentos serão apresentados em seguida.

Um filtro inverso de ordem p estável, onde todas as raízes se encontram dentro do círculo unitário, pode ser definido por

$$A_p(z) = \sum_{i=0}^{p} a_i z^{-i} \tag{4.30}$$

onde $a_0 = 1$ e a_i é o i-ésimo coeficiente de predição linear (LPCs).

As LSFs de ordem p são definidas como sendo as raízes complexas dos polinômios $P(z)$ e $Q(z)$, as quais são expressas por

$$P(z) = A(z) + z^{-(p+1)} A(z^{-1}) \tag{4.31}$$

$$Q(z) = A(z) - z^{-(p+1)} A(z^{-1}) \tag{4.32}$$

Para obter a relação entre LPCC e LSF é preciso realizar a multiplicação de (4.31) e (4.32), resultando em

$$P(z)Q(z) = A^2(z)\left[1 - R^2(z)\right] = (1 - z^{-2}) \prod_{i=1}^{p} \left(1 - e^{jw_i} z^{-1}\right)\left(1 - e^{-jw_i} z^{-1}\right) \tag{4.33}$$

para p par e maior que 2, onde w_i é o i-ésimo parâmetro LSF e

$$R(z) = \frac{z^{-(p+1)} A(z^{-1})}{A(z)} \tag{4.34}$$

e aplicando o logaritmo nos dois lados de (4.33) chega-se a

$$\begin{aligned} 2\log A_p(z) + \log\left(1 - R^2(z)\right) &= \log\left(1 - z^{-2}\right) \\ &+ \sum_{i=1}^{p} \left(\log\left(1 - e^{jw_i} z^{-1}\right) + \log\left(1 - e^{-jw_i} z^{-1}\right)\right) \end{aligned} \tag{4.35}$$

Fazendo, agora, a expansão em série em ambos os lados de (4.35), obtém-se

$$-2\sum_{n=1}^{\infty} c_n e^{-jwn} + \sum_{n=1}^{\infty} R_n e^{-jwn} = -\sum_{n=1}^{\infty} \frac{1}{n}\left(1+(-1)^n\right)e^{-jwn} - \sum_{n=1}^{\infty} \frac{1}{n}\sum_{i=1}^{p}\left(e^{jnw_i} + e^{-jnw_i}\right)e^{-jwn} \tag{4.36}$$

onde c_n é o n-ésimo parâmetro LPCC que satisfaz a relação

$$\log A_p(e^{jw}) = -\sum_{n=1}^{\infty} c_n e^{-jwn} \tag{4.37}$$

e R_n é a transformada inversa de Fourier de $\log(1-R^2(z))$. Pode-se mostrar que a expansão dada pela equação (4.36) converge [4]. De (4.36) pode-se obter a partir de algumas manipulações matemáticas que

$$c_n = \frac{1}{2n}\left(1+(-1)^n\right) + \frac{1}{n}\sum_{i=1}^{p}\cos nw_i + R_n \tag{4.38}$$

Observando-se a equação (4.38), percebe-se que ainda existe o termo R_n que depende dos parâmetros LPC e que os demais só dependem das LSFs. Sendo assim, será desconsiderado este termo, dando origem à expressão do parâmetro PCC definido por

$$\hat{c}_n = \frac{1}{2n}\left(1+(-1)^n\right) + \frac{1}{n}\sum_{i=1}^{p}\cos nw_i \tag{4.39}$$

É razoável esperar que desprezar o fator $R_n/2$ não venha a prejudicar o desempenho, pois este fator será zero, ou assumirá valores muito pequenos, para a maioria dos casos [1].

4.2.3.
Pseudo-Cepstrum (PCEP)

Com base na dedução matemática dos parâmetros PCC, se torna bastante trivial a obtenção dos parâmetros PCEP. Esses parâmetros são obtidos a partir dos parâmetros PCC, eliminando-se o termo $\frac{1}{2n}\left(1+(-1)^n\right)$ que não depende da voz, ou seja, não depende dos parâmetros LSF. A expressão dos parâmetros PCEP é dada por

$$\hat{d}_n = \frac{1}{n}\sum_{i=1}^{p}\cos nw_i \tag{4.40}$$

Pode-se esperar um bom desempenho espectral dos parâmetros PCEP, pois os mesmos fornecem uma envoltória espectral bastante parecida com a do Cepestro obtido diretamente de voz [1]. O PCEP possui a vantagem de apresentar ainda uma carga computacional mais baixa do que o parâmetro PCC obtido anteriormente.

4.2.4.
Mel-Frequency PCC (MPCC)

Para obter os parâmetros MPCC a partir dos parâmetros PCC basta manipular as LSFs a serem utilizadas em (4.39), onde w_i é substituído por w_i^m, definido pela transformação

$$w_i^m = w_i + 2\tan^{-1}\left(\frac{0{,}45\,sin\,w_i}{1-0{,}45\cos w_i}\right) \tag{4.41}$$

Essa equação consiste em uma forma de se transformar os eixos de frequência de um determinado conjunto de parâmetros nos eixos de frequência da escala mel [73]. Com esta alteração de eixo, obtém-se os parâmetros MPCC, dados pela expressão

$$\hat{c}_n^m = \frac{1}{2n}\left(1+(-1)^n\right)+\frac{1}{n}\sum_{i=1}^{p}\cos nw_i^m \tag{4.42}$$

onde $\hat{c}_n^m$ é o n-ésimo parâmetro MPCC.

4.2.5. Mel-Frequency PCEP (MPCEP)

Para se chegar aos parâmetros MPCEP, basta repetir o procedimento descrito para os parâmetros MPCC, obtendo a seguinte expressão

$$\hat{d}_n^m = \frac{1}{n}\sum_{i=1}^{p}\cos nw_i^m \tag{4.43}$$

onde $\hat{d}_n^m$ é o n-ésimo parâmetro MPCEP.

4.3. Atributos Extraídos de ISFs

As *Immittance spectral frequencies* (ISFs) são utilizadas para codificação de voz no codificador AMR-WB, o codec recomendado para a terceira geração de telefonia celular que está sendo implantado em todo o mundo.

Primeiramente, na seção 4.3.1, será deduzido matematicamente o método de obtenção dos parâmetros ISF a partir dos parâmetros LPC, sendo esta apresentação justificada, pois as ISFs são os parâmetros do codificador utilizado na terceira geração de telefonia celular, sendo assim interessante se pesquisar também atributos que possam ser extraídos diretamente dos mesmos (o que será colocado nesta tese como proposta para futuros trabalhos).

4.3.1. *Immittance Spectral Frequencies* (ISF)

Uma outra representação dos parâmetros LPC são os parâmetros ISF que estão proximamente relacionados aos parâmetros LSF. O modelo ISF [60] é definido usando os polinômios $F(z)$ e $G(z)$ que são definidos como

$$F(z) = A(z) + z^{-p} A(z^{-1}) \tag{4.44}$$

$$G(z) = A(z) - z^{-p} A(z^{-1}) \tag{4.45}$$

onde $F(z)$ é um polinômio simétrico e $G(z)$ é um polinômio antissimétrico.

Note a similaridade entre os polinômios $F(z)$ e $G(z)$ definidos em (4.44) e (4.45) com os polinômios $P(z)$ e $Q(z)$ definidos em (4.27) e (4.28), estando a diferença no fator que multiplica $A(z^{-1})$. Nas LSF tem $z^{-(p+1)}$ e nos ISF tem z^{-p} como este fator. As raízes de $F(z)$ e $G(z)$ determinam os coeficientes ISF [42].

Se $A(z)$ é de fase mínima, ou seja, se $H(z)$ é estável, então:

1. As raízes de $F(z)$ e $G(z)$ estão sobre o círculo unitário.
2. As raízes de $F(z)$ estão alternadas com as raízes de $G(z)$, ou seja, $r_1 < q_1 < r_2 < q_2 < \ldots < q_{p+1}$, onde r_i e q_i representam a posição angular da i-ésima raiz de $F(z)$ e $G(z)$, respectivamente.
3. O filtro H(z) continuará estável após a quantização das raízes de $F(z)$ e $G(z)$ desde que (1) e (2) sejam respeitados pelos valores quantizados.
4. Sendo $H(z)$ estável, a mesma permanecerá estável após a interpolação.

Além disso, os ISF possuem uma faixa dinâmica bem comportada, possibilitando uma quantização mais eficiente, do que as outras formas de representar os coeficientes LPC, o que foi verificado em resultados experimentais em [60].

Finalmente, os coeficientes de predição de $A(z)$ são obtidos a partir dos coeficientes de $F(z)$ e $G(z)$ através da seguinte igualdade polinomial:

$$A(z) = \frac{F(z) + G(z)}{2} \tag{4.46}$$

Os coeficientes ISF apresentam a importante propriedade de robustez à distorção. De acordo com ela, qualquer alteração sofrida por um desses coeficientes não terá um efeito global. Apenas será afetada a região do espectro próxima a esta frequência. Assim como ocorre com as LSFs, esta propriedade pode ser explorada em sistemas de codificação da voz, uma vez que o ouvido humano não é muito sensível a variações em frequências elevadas. Nesses sistemas, é possível representar os coeficientes ISF de elevadas frequências com um menor número de bits (o que é realizado normalmente pelos codificadores de voz), o que possibilita uma diminuição da taxa de bits do sistema. Foi observado em [60] que quando se comparam ISFs às LSFs esta compressão é ainda maior, quando se mantem a qualidade de voz desejada depois da decodificação, ou seja, a ISF tem maior capacidade de armazenamento de informação e proteção da mesma.

De acordo com as propriedades acima, pode-se concluir que a utilização de coeficientes ISF apresenta vantagens em relação aos coeficientes LPC em termos de transmissão, quantização e interpolação. No entanto, o cálculo direto dos coeficientes ISF exige uma elevada capacidade computacional. A alternativa é calcular os coeficientes LPC e depois transformá-los em ISF como é feito no codificador AMR-WB.

Os coeficientes ISFs ainda têm como vantagem sobre as LSFs a redução de carga computacional, pois reduzem em um o número de raízes que precisam ser calculadas no processo de obtenção das ISFs, em detrimento das LSFs [60].

Será apresentado, no capítulo 7 desta tese, que a dedução matemática dos atributos de reconhecimento a partir dos ISFs é uma das propostas de trabalhos futuros.

4.4. Atributo Extraído de Voz Reconstruída (MFCC)

Nesta seção é considerado o atributo (MFCC - *Mel-Frequency Cepstral Coefficients*) que necessita ser obtido a partir de voz. No sistema aqui considerado, o atributo será obtido a partir da voz recuperada no decodificador localizado no receptor do sistema celular ou de voz sobre IP. Por esse motivo, ele foi classificado como atributo extraído de voz reconstruída.

Os coeficientes Mel-cepestrais surgiram devido aos estudos na área de psicoacústica (ciência que estuda a percepção auditiva humana), que mostraram que a percepção humana das frequências de tons puros ou de sinais de voz não segue uma escala linear. Isto estimulou a idéia de serem definidas frequências subjetivas de tons puros, da seguinte forma: para cada tom com frequência *f*, medida em Hz, define-se um tom subjetivo medido em uma escala que se chama escala mel. O mel, então, é uma unidade de medida da frequência percebida de um tom.

Como referência, definiu-se a frequência de 1 kHz, com potência 40 dB acima do limiar mínimo de audição do ouvido humano, como 1000 mels [66]. Os outros valores subjetivos foram obtidos através de experimentos, onde se pedia a ouvintes que ajustassem a frequência física de um tom, até que a frequência percebida fosse igual a duas vezes a frequência de referência; depois, 10 vezes a frequência de referência e assim por diante. Essas frequências teriam os valores de 2000 mels, 10000 mels e assim sucessivamente. O mesmo processo era efetuado na outra direção, ou seja, metade do tom de referência, um décimo do tom de referência, etc. Essas frequências teriam valores de 500 mels, 100 mels, etc. Isto permitiu verificar que o mapeamento entre a escala de frequência real, em Hz, e a escala de frequências percebida, em mel, é aproximadamente linear abaixo de 1000 Hz e, logarítmica, acima.

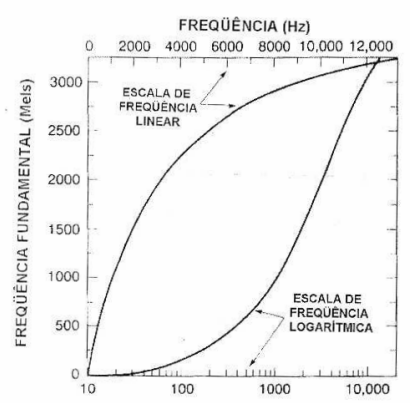

Figura 4.1 – Percepção subjetiva da frequência fundamental de sons sonoros

A Fig. 4.1 apresenta um gráfico da frequência fundamental subjetiva de tons em função da frequência [74]. A curva superior mostra a relação entre aquela e esta em uma escala linear. Pode-se observar que a frequência fundamental subjetiva, em mels, cresce menos e menos rapidamente à medida que há um

aumento linear na frequência. A curva inferior, por outro lado, mostra a frequência fundamental subjetiva em função da frequência em uma escala logarítmica. Pode-se notar na Fig. 4.1, que a frequência fundamental subjetiva é essencialmente linear para frequências inferiores a 1000 Hz.

Um outro importante critério subjetivo de conteúdo de frequência de um sinal é a banda crítica. Alguns experimentos demonstraram que a percepção humana de algumas frequências de sons complexos não pode ser individualmente identificada, dentro de certas bandas. Quando uma componente cai fora da banda, chamada de banda crítica, ela pode ser identificada. Uma explicação apresentada para esse fato foi que a percepção de uma frequência particular pelo sistema auditivo, por exemplo f_0, é influenciada pela energia da banda crítica das frequências em torno de f_0. O valor dessa banda varia nominalmente de 10 a 20 % da frequência central do som, começando em torno de 100 Hz para frequências abaixo de 1 kHz e aumentando em escala logarítmica, acima.

Esses fenômenos (escala mel e banda crítica) sugeriram que seria mais interessante fazer algumas modificações na representação e nas medidas de distâncias espectrais. Tais modificações consistiram, primeiramente, em fazer uma ponderação da escala de frequência para a escala mel e, além disso, incorporar a noção de banda crítica na definição de distorção espectral. Ou seja, ao invés de se usar simplesmente o logaritmo da magnitude das frequências, passou-se a utilizar o logaritmo da energia total das bandas críticas em torno das frequências mel. A aproximação mais utilizada para esse cálculo é a utilização de um banco de filtros triangulares, espaçados uniformemente em uma escala não linear (escala mel).

A técnica de ponderação mel pode ser aplicada a vários tipos de representação espectral. Cabe destaque a representação cepestral, devido à combinação da mesma com a técnica mencionada (mel), ser a mais utilizada e apresentar maior eficácia computacional, sendo chamada de Mel-Cepestral [66].

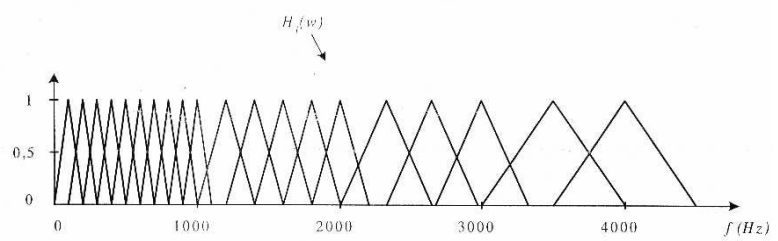

Figura 4.2 – Magnitude do espectro dos filtros de banda crítica

A Fig. 4.2 apresenta a configuração para o cálculo dos coeficientes Mel-Cepestrais. Para a faixa de frequências de interesse da voz humana, utilizam-se 20 filtros centrados nas frequências da escala mel. O espaçamento é de aproximadamente 150 mels e a largura de banda de cada filtro triangular é de 300 mels. Os valores dos centros são apresentados na Tab. 4.1. Como os valores calculados pela Transformada Rápida de Fourier (*Fast Fourier Transform – FFT*) são discretos, a tabela também mostra as aproximações para esses centros quando se utiliza FFT de 1024 pontos e frequência de amostragem de 8 kHz [68].

Filtro i	Centro Desejado (Hz)	Centro Aproximado (Hz)	Banda Crítica (Hz)
1	100	102	100
2	200	203	100
3	300	305	100
4	400	406	100
5	500	500	100
6	600	602	100
7	700	703	100
8	800	805	100
9	900	906	100
10	1000	1000	124
11	1148	1148	160
12	1318	1320	184
13	1514	1516	211
14	1737	1742	242
15	1995	2000	278
16	2291	2297	320
17	2630	2633	367
18	3020	3023	422
19	3467	3469	484
20	4000	4000	556

Tabela 4.1 – Frequências dos centros e banda crítica dos filtros utilizados para cálculo dos coeficientes mel-cepestrais

Inicialmente, divide-se o sinal de voz $s(n)$ em janelas. Para cada janela m estima-se o espectro $S(w,m),$ utilizando-se FFT, cujo espectro de magnitude é dado por

$$|S(w,m)| = (\mathrm{Re}[S(w,m)]^2 + \mathrm{Im}[S(w,m)]^2)^{1/2} \tag{4.47}$$

O espectro modificado $P(i), i = 1,2,\ldots,N_f,$ consistirá na energia de saída de cada filtro, expresso por

$$P(i) = \sum_{k=0}^{N/2} |S(k,m)|^2 H_i\left(k\frac{2\pi}{N}\right) \tag{4.48}$$

onde N é o número de pontos da FFT, N_f é o número de filtros triangulares, $|S(k,m)|$ é o módulo da amplitude na frequência do k-ésimo ponto da m-ésima janela e $H_i(w)$ é a função de transferência do i-ésimo filtro triangular, definido por

$$H_i(w) = \begin{cases} \dfrac{1}{k_i - k_{i-1}}(w - k_{i-1}) & k_{i-1} \le w \le k_i \\ \dfrac{1}{k_i - k_{i+1}}(w - k_{i+1}) & k_i \le w \le k_{i+1} \end{cases} \tag{4.49}$$

onde, k_i é o i-ésimo centro, cujos valores estão mostrados na Tab. 4.1, $k_0 = 0,$ e w é uma escala ajustada de acordo com o número de pontos da FFT, e expressa por

$$w = k\frac{2\pi}{N} \qquad 0 \le k \le N/2 \tag{4.50}$$

Em seguida, define-se o conjunto de pontos $E(k)$ por

$$E(k) = \begin{cases} \log[P(i)] & k = k_i \\ 0 & qq\ outro\ k \in [0, N-1] \end{cases} \tag{4.51}$$

Os coeficientes mel-cepestrais $c_{mel}(n)$ são então obtidos com o uso da Transformada Inversa de Fourier (IFFT), usando-se a seguinte equação:

$$c_{mel}(n) = \frac{1}{N}\sum_{k=0}^{N-1} E(k) e^{j\left(\frac{2\pi}{N}\right)kn} \qquad n = 1,2,\ldots,N_c \tag{4.52}$$

onde N_c é o número de coeficientes desejado.

Como $E(k)$ é simétrico em relação a $N/2$ (ou $\pi/2$) e lembrando que

$$e^{j\left(\frac{2\pi}{N}\right)kn} = \cos\left(\frac{2\pi}{N}kn\right) + j\,\mathrm{sen}\left(\frac{2\pi}{N}kn\right) \tag{4.53}$$

resulta que os termos em seno da (4.52) se cancelam, gerando a equação

$$c_{mel}(n) = \frac{1}{N}\sum_{k=0}^{N-1} E(k)\cos\left(\frac{2\pi}{N}kn\right) \tag{4.54}$$

Ainda usando a simetria e observando que

$$E(0) = E\left(N/2\right) \tag{4.55}$$

obtém-se a expressão

$$c_{mel}(n) = \frac{2}{N}\sum_{k=1}^{\frac{N}{2}-1} E(k)\cos\left(\frac{2\pi}{N}kn\right) \tag{4.56}$$

Sabendo-se que no intervalo $0 \le k \le (N-2)/2$ existirão apenas N_f termos diferentes de zero, que são os correspondentes aos centros dos filtros, e

eliminando-se o fator de escala $2/N$, a equação (4.56) pode ser simplificada, chegando-se à expressão final para os coeficientes MFCC, dado por

$$c_{mel}(n) = \sum_{i=1}^{N_f} E(k_i) \cos\left(\frac{2\pi}{N} k_i n\right) \qquad n = 1,2,\ldots,N_c \qquad (4.57)$$

onde N_c é o número de coeficientes mel-cepestrais desejado, N_f é o número de filtros e k_i é o centro do i-ésimo filtro.

4.5. Conclusão

Neste capítulo foram apresentados a base teórica e os parâmetros/atributos que serão utilizados para a implementação do sistema de reconhecimento de voz distribuído no ambiente celular/voz sobre IP desta tese.

O capítulo seguinte descreve uma nova técnica de interpolação de parâmetros que visa melhorar o desempenho do reconhecedor de voz distribuída quando comparada com a interpolação linear.

5
Métodos de Interpolação dos Atributos

Para um bom funcionamento dos Sistemas Automáticos de Reconhecimento de Voz é necessário que os atributos de reconhecimento sejam obtidos a uma taxa elevada, porém os codificadores de Voz para Telefonia IP e Móvel Celular, usados em cenários distribuídos, normalmente geram seus parâmetros a taxas mais baixas, o que degrada o desempenho do reconhecedor. Usualmente é utilizada a interpolação linear no domínio das LSFs para resolver este problema. Nesta tese foi proposta a realização da interpolação com a utilização de um Filtro Digital Interpolador que demonstrou ter um desempenho de reconhecimento ainda melhor que a interpolação linear.

Na seção 5.1 deste capítulo é apresentada a interpolação linear. Na seção 5.2 é descrita a técnica de interpolação utilizando filtro digital proposta nesta tese. A seção 5.3 é dedicada às simulações utilizando o codec ITU-T G.723.1 usado em Voz para Telefonia IP. Na seção 5.4 são apresentados os resultados de simulação utilizando o codec AMR-NB usado em telefonia celular GSM. Na seção 5.5 são apresentados os resultados de simulação utilizando o codec AMR-WB que está sendo adotado em telefonia celular de terceira geração e redes IP. Finalmente, a seção 5.6 contém uma breve conclusão.

5.1.
Interpolação Linear

A interpolação linear é um dos métodos comumente utilizados para estimar valores entre pares de amplitudes adjacentes de sequências discretas no tempo. Em particular, esta técnica já foi utilizada em reconhecimento de voz distribuído para interpolar as LSFs decodificadas do codificador IS-641 [6] e do codificador ITU-T G.723.1 em [75]. No caso do codificador ITU-T G.723.1 os parâmetros LSF são obtidos a uma taxa de 33 Hz (um conjunto de parâmetros a cada 30 ms) e interpolados para obter uma taxa mais elevada de 100 Hz (um conjunto de parâmetros a cada 10 ms) [6] [75].

A interpolação linear é implementada passando o sinal $x[n]$, que se deseja interpolar linearmente, por um *up-sampler* cuja saída é $x_u[n]$ dado por

$$x_u[n] = \begin{cases} x[n/r], & n = 0, r, 2r, 3r, \ldots \\ 0, & para \quad n \neq 0, r, 2r, 3r, \ldots \end{cases} \tag{5.1}$$

onde $r > 1$ é o fator de sobre-amostragem que se quer utilizar e $r - 1$ é o número de zeros inseridos entre as amostras.

Tendo obtido $x_u[n]$, passa-se o mesmo por um segundo sistema discreto no tempo, que substitui as amostras de valor nulo inseridas pelo *up-sampler* por amostras que estão na linha reta que une o par de entradas $x[n]$ adjacentes às amostras que estão sendo substituídas [68].

O sinal interpolado linearmente é designado por $y[n]$ e pode ser computado para interpolação de fator 2 ($r = 2$ no *up-sampler*) por

$$y[n] = x_u[n] + \frac{1}{2}(x_u[n-1] + x_u[n+1]) \tag{5.2}$$

e para interpolação de fator 3 ($r = 3$ no *up-sampler*) por

$$y[n] = x_u[n] + \frac{1}{3}(x_u[n-1] + x_u[n+2]) + \frac{2}{3}(x_u[n-2] + x_u[n+1]) \tag{5.3}$$

Só foram apresentadas as expressões para o sinal interpolado pelos fatores 2 e 3, pois só serão considerados aumentos de taxa de um determinado parâmetro por estes mesmos fatores para aumento de taxa dos parâmetros dos codificadores.

Na Fig. 5.1 é apresentada a representação gráfica da interpolação Linear de fator 3, onde se pode observar que a mesma não utiliza nenhuma propriedade do sinal que será utilizado no reconhecimento, a não ser preencher as amostras faltantes por valores correspondentes a pontos pertencentes a reta que ligam as amostras conhecidas.

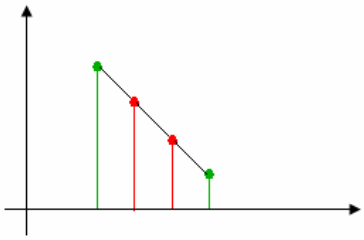

Figura 5.1 – Representação gráfica da interpolação Linear de fator 3

5.2.
Interpolação com Filtro Digital

Nesta tese, está sendo proposta uma nova técnica de interpolação para parâmetros/atributos de reconhecimento de voz distribuída, que foi projetada usando um *up-sampler* e um filtro digital passa-baixa $H(z)$. O *up-sampler* utilizado nesta técnica é o mesmo utilizado na interpolação linear, já apresentado na seção anterior deste capítulo. O filtro passa-baixa $H(z)$ é o responsável pela substituição das amostras zeradas inseridas pelo *up-sampler* pelo valor mais próximo do ideal das mesmas. Para isso, o mesmo elimina a inserção de imagens do espectro original comprimidas pelo fator de sobre-amostragem r que são inseridas pelo *up-sampler* com sua equação geral dada por (5.1). Uma observação importante nesta aplicação de interpolação é que os valores de amostras do sinal de entrada não devem ser alterados na saída (da mesma forma que já ocorre na interpolação linear). Isso implica na utilização de técnicas de projeto de interpoladores ótimos utilizando filtros digitais [68], pois caso contrário, no momento da filtragem por $H(z)$ poder-se-ia ter a substituição das amostras originais, o que levaria, ainda, a uma maior degradação da qualidade do sinal.

Esta técnica proposta tira vantagem das propriedades em frequência do sinal original ao qual se quer aumentar a taxa de amostragem. De forma ilustrativa são representadas nas Fig. 5.2 e 5.3 respectivamente o sinal original e o espectro do sinal original.

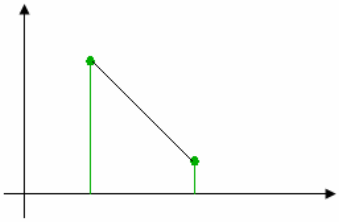

Figura 5.2 – Representação gráfica do Sinal Original

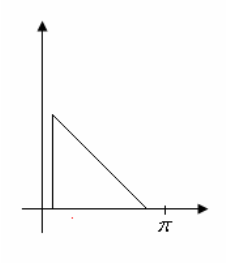

Figura 5.3 – Representação gráfica do Espectro em Freqüência do Sinal Original

Nas Fig. 5.4 e 5.5 são apresentados respectivamente o sinal e seu espectro depois de ser aplicada a sobre-amostragem definida na equação 5.1.

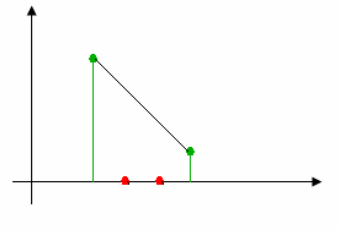

Figura 5.4 – Representação gráfica do Sinal sobre-amostrado de fator 3

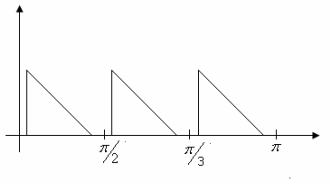

Figura 5.5 – Representação gráfica do Espectro em Freqüência do Sinal sobre-amostrado de fator 3

Nas Fig. 5.6 e 5.7 são apresentados respectivamente o sinal e seu espectro após ser aplicado o filtro passa-baixa proposto sobre o sinal da Fig. 5.4, onde o filtro utilizado preserva o valor das amostras originais.

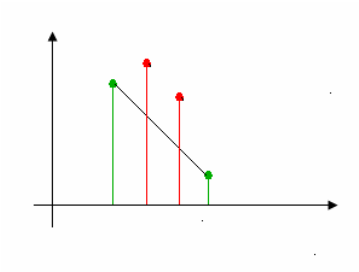

Figura 5.6 – Representação gráfica do Sinal sobre-amostrado de fator 3 depois de filtragem passa-baixa

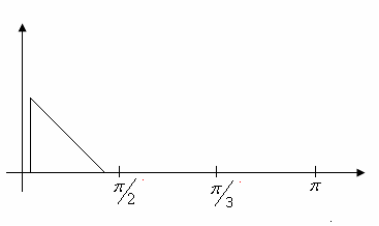

Figura 5.7 – Representação gráfica do Espectro em Freqüência do Sinal sobre-amostrado de fator 3 depois de filtragem passa-baixa

Para que esta nova técnica de interpolação fosse utilizada no cenário de reconhecimento de voz distribuída, assumiu-se que o sinal de entrada possui energia limitada e banda limitada à faixa $0 \le \omega \le \alpha$, onde α deve ser igual ou menor que 0.5 (nesta tese α sempre foi considerado igual a 0.5 em todos os testes). O filtro $H(z)$ é um filtro FIR (*Finite Impulse Response*) simétrico que determina as amostras faltantes, através da minimização do erro médio quadrático, usando o princípio da ortogonalidade [76]. O comprimento do filtro é determinado pela expressão $2rL+1$, onde L é um inteiro que determina o comprimento do filtro (nesta tese foi feito L igual a 4, pois mediante resultados de simulação foi este valor que levou ao comprimento de filtro com melhor desempenho de reconhecimento, tendo sido simulado com L igual 1, 2, 3, 4, 5, 6). Porém, para reduzir a complexidade de calcular diretamente os coeficientes de um filtro de comprimento $2rL+1$ (igual a 25 para o ITU-T G.723.1 e 17 para o AMR-NB), foram utilizadas as técnicas propostas em [76] para desenvolver um interpolador ótimo com menor complexidade e um menor consumo de hardware. O procedimento corresponde em substituir o projeto de um filtro de $2rL+1$ (igual a 25 para o ITU-T G.723.1 e 17 para o AMR-NB) pelo projeto de r (igual a 3 para

o ITU-T G.723.1 e 2 para o AMR-NB) filtros de comprimento $2L+1$ (igual a 9 para o ITU-T G.723.1 e 9 para o AMR-NB).

5.3. Resultados de Simulação para o Codec ITU-T G.723.1

Para o caso do codec ITU-T G.723.1 que opera com uma taxa de codificação de LSFs de 33 Hz, é necessário utilizar um fator de interpolação igual a 3 ($r=3$) para atingir a taxa de geração de atributos requerida para reconhecimento de voz distribuída que é de 100 Hz. Com a determinação do valor de r tem-se que para esta nova técnica de interpolação para reconhecimento de voz distribuído, o filtro $H(z)$ é de comprimento $2rL+1=25$ (tendo assumido $L=4$) e fase 0. Utilizou-se as técnicas propostas em [75] visando reduzir a complexidade através da obtenção de 3 filtros de comprimento $2L+1=9$, o qual foi implementado utilizando a função *interp* do Matlab.

As simulações para o codificador ITU-T G.723.1 foram divididas em dois grupos. No primeiro buscou-se avaliar a interpolação linear nas diversas formas de utilizar a base de vozes conforme apresentado no capítulo 1 seção 1.2, visando avaliar o impacto no desempenho de reconhecimento de acordo com as características da base. Já no segundo grupo de simulações avaliou-se a nova técnica de interpolação utilizando filtros digitais, porém se restringindo ao uso da base de vozes no cenário mais próximo da utilização real de um sistema de reconhecimento de voz distribuído, que é o reconhecimento independente do locutor e do texto usado no treinamento.

Para facilitar o entendimento dos resultados e minimizar que o leitor tenha que se referir aos capítulos anteriores desta tese, é apresentado aqui um breve resumo de forma a facilitar o entendimento sobre os resultados obtidos.

A nova base usada nesta tese foi desenvolvida a partir de um conjunto de 1000 frases foneticamente balanceadas para o Português Brasileiro [2]. Esta base foi composta por 50 locutores masculinos e 50 locutores femininos, cada um repetindo uma vez todas as 1000 frases (3528 palavras). A base foi gravada em estúdio com uma frequência de amostragem de 16 kHz e 16 bits por amostra com largura de banda 50 – 7000 Hz. Esta base foi filtrada e sub-amostrada [68] para alcançar os requerimentos do codificador ITU-T G.723.1 [39].

Nos experimentos, foi considerado um sistema LVDCSR (*Large Vocabulary Distributed Continuous Speech Recognition*) usando o codificador ITU-T G.723.1. O codificador ITU-T G.723.1 é um dos mais usados padrões para redes IP. O mesmo está presente em diversos produtos de grandes fabricantes e operadoras de telecomunicações. O mesmo permite a codificação em 6,3 kbit/s ou 5,3 kbit/s. Em nossos experimentos foram consideradas ambas as taxas de operação. O codificador ITU-T G.723.1 emprega quadros de 30 ms, taxa de amostragem de 8 kHz, e 10 LSFs por quadro. As LSFs são quantizadas em 24 bits por um PSVQ (*Predictive Split Vector Quantizer*) e transmitidas numa taxa de 33 Hz (uma a cada 30 ms). A taxa de 100 Hz para os atributos foi escolhida, pois é o valor empregado usualmente para propiciar um bom desempenho do reconhecedor. Interpolar as LSFs de 1 para cada 30 ms para 1 para cada 10 ms é equivalente a uma interpolação de fator $r = 3$. Baseado nos resultados apresentados em [75], apenas foi considerada a interpolação no domínio das LSFs. Isso significa que os atributos baseados em LSF (MPCC e MPCEP) ou em LPC (MLPCC) serão obtidos em 100 Hz pela interpolação dos parâmetros LSF de 33 Hz para 100 Hz. A MFCC é gerada da voz original e reconstruída com uma duração de quadro de 25 ms (com sobreposição de quadro de forma que os atributos sejam gerados a cada 10 ms). Nenhuma interpolação então é necessária, pelo fato que a MFCC pode ser extraída diretamente da voz original e da voz decodificada na taxa de 100 Hz. Foram considerados apenas os atributos na escala MEL obtidos de LPC (MLPCC [5]) e de LSF (MPCC [1] e MPCEP [4]), pelo fato que os mesmos oferecem um desempenho muito melhor que o atingido para os atributos na escala linear (LPCC, PCC e PCEP) [75]. Com o objetivo de comparação foram também obtidos os atributos MFCC (*Mel-Frequency Cepstral Coefficients*) de voz original e voz reconstruída com o codificador ITU-T G.723.1 nas duas diferentes taxas de operação (6.3 kbit/s ou 5.3 kbit/s) [1].

É importante frisar que em todos os casos, os modelos são treinados com os mesmos tipos de atributos que serão utilizados nos testes. Isso significa que estará se trabalhando sempre em condições casadas.

Para garantir a confiabilidade estatística dos resultados, foi utilizada validação cruzada em todos os experimentos. A taxa Média de reconhecimento de

palavra ($\overline{WRR}$ - *Average Word Recognition Rates*), apresentada nas tabelas de resultado é obtida pela expressão [16]

$$\overline{WRR} = \frac{\sum_{i=1}^{N} WRR_i}{N} \tag{5.4}$$

onde N são o número de diferentes experimentos realizados ($N = 4$ para os cenários 1 e 2 e $N = 16$ para o cenário 3 de utilização da base) e WRR_i (*Word Recognition Rate*) no experimento i que é dado por

$$WRR_i = \left(1 - \frac{S + I + D}{W}\right) \cdot 100 \tag{5.5}$$

onde W é o número total de palavras na seqüência de teste e S, I e D são, respectivamente, o número total de erros por substituição (*substitution*), inserção (*insertion*) e supressão (*deletion*) na seqüência reconhecida. O desvio padrão (σ) é definido por [16]

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} \left(WRR_i - \overline{WRR}\right)^2}{N - 1}} \tag{5.6}$$

O Intervalo de Confiança de c% (neste caso 95%) para a média $\overline{WRR}$ é por definição, o intervalo

$$\left[\overline{WRR} - \delta(c), \overline{WRR} + \delta(c)\right] \tag{5.7}$$

Considerando que a distribuição de WRR_i pode ser considerada normal então $\delta(c)$ é dado por

$$\delta(c) = \frac{t_c(N)\ \sigma}{\sqrt{N}} \tag{5.8}$$

onde $t_c(N)$ é o valor tal que à esquerda está $(50+c/2)\%$ da área total sob a curva de densidade da distribuição t de Student com $N-1$ graus de liberdade.

Em todos os experimentos desta seção, o extrator de atributos gera um conjunto de 10 parâmetros, mais suas primeira e segunda derivadas, representando um total de 30 atributos de reconhecimento. O modelo acústico usa HMMs contínuas de três estados com a mistura de 20 Gaussianas por estado para modelar o fone. Como o silêncio é estacionário, um estado foi utilizado com o mesmo número de gaussianas. Os mesmos foram implementados usando o HTK (*HMM Toolkit*) software [77]. Trifones inter e intra palavra foram usados como unidades acústicas. O modelo de linguagem trigrama foi treinado usando o HTK (*HMM Toolkit*) software [77] com um léxico de 60.080 palavras obtidas de 240.000 frases extraídas de um grande corpus de textos do Ceten-Folha [25]. O modelo de linguagem trigrama foi implementado utilizando o ATK (*Application Toolkit for HTK*) [78].

O sistema foi simulado utilizando uma máquina Sun V880 com 4 processadores, 8Gb de memória RAM executando sistema operacional Solaris 10.

A. Interpolação Linear e Variação do uso das Bases de Vozes

Os resultados de desempenho são apresentados em três tabelas de acordo com a divisão realizada na base de vozes para a realização dos testes. Tab. 5.1 apresenta os resultados de reconhecimento para o sistema dependente de 100 locutores (Figura 1.2, do capítulo 1 – cenário 1), Tab. 5.2 traz os resultados de reconhecimento para o sistema independente de locutor, com todas as frases apresentadas no treino e teste (Figura 1.3, do capítulo 1 – cenário 2), e Tab. 5.3 apresenta os resultados para o sistema independente do locutor e das frases (Figura 1.4, do capítulo 1 – cenário 3).

Atributos	$\overline{WRR}$	σ	Intervalo de Confiança
MFCC - Voz Original	86,72%	1,02%	[85,52% ; 87,92%]
MFCC - Voz Reconstruída (5,3 Kbits/s)	72,10%	1,12%	[70,78% ; 73,42%]
MFCC - Voz Reconstruída (6,3 Kbits/s)	73,83%	1,07%	[72,57% ; 75,09%]
MPCC - Interp, 33Hz para 100 Hz	77,21%	1,01%	[76,02% ; 78,40%]
MPCEP - Interp, 33Hz para 100 Hz	78,11%	0,99%	[76,95% ; 79,28%]
MLPCC - Interp, 33Hz para 100 Hz	77,74%	1,01%	[76,55% ; 78,93%]

Tabela 5.1 – Tabela de desempenho de reconhecimento para sistema dependente de 100 locutores

Atributos	$\overline{WRR}$	σ	Intervalo de Confiança
MFCC - Voz Original	82,55%	1,37%	[80,94% ; 84,16%]
MFCC - Voz Reconstruída (5,3 Kbits/s)	68,02%	1,45%	[66,32% ; 69,73%]
MFCC - Voz Reconstruída (6,3 Kbits/s)	70,05%	1,39%	[68,44% ; 71,69%]
MPCC - Interp, 33Hz para 100 Hz	73,32%	1,37%	[71,71% ; 74,93%]
MPCEP - Interp, 33Hz para 100 Hz	74,27%	1,35%	[72,68% ; 75,86%]
MLPCC - Interp, 33Hz para 100 Hz	73,81%	1,35%	[72,22% ; 75,40%]

Tabela 5.2 – Tabela de desempenho de reconhecimento para sistema independente de locutor e com as mesmas frases para teste e treino

Atributos	$\overline{WRR}$	σ	Intervalo de Confiança
MFCC - Voz Original	76,82%	1,63%	[76,11% ; 77,54%]
MFCC - Voz Reconstruída (5,3 Kbits/s)	62,21%	1,74%	[61,45% ; 62,97%]
MFCC - Voz Reconstruída (6,3 Kbits/s)	63,94%	1,68%	[63,20% ; 64,68%]
MPCC - Interp, 33Hz para 100 Hz	66,31%	1,64%	[65,59% ; 67,03%]
MPCEP - Interp, 33Hz para 100 Hz	67,19%	1,61%	[66,49% ; 67,90%]
MLPCC - Interp, 33Hz para 100 Hz	66,83%	1,67%	[66,10% ; 67,56%]

Tabela 5.3 – Tabela de desempenho de reconhecimento para sistema independente de locutor e das frases

Comparando os resultados apresentados na Tab. 5.1 (cenário dependente de locutor) com os resultados da Tab. 5.2 (cenário independente de locutor com todas as frases usadas para treinamento), pode-se observar que a variabilidade do locutor é responsável pela redução de aproximadamente 4 % no $\overline{WRR}$ (*Average Word Recognition Rate*). Comparando também a Tab. 5.2 (cenário independente de locutor com todas as frases usadas para treinamento) com a Tab. 5.3 (cenário independente de locutor e das frases) verifica-se que o desempenho se reduz de entorno de 6 % da Tab. 5.2 para a Tab. 5.3. Isso mostra que além da redução de 4% no desempenho devido à variabilidade do locutor, um decréscimo adicional de 6% ocorre devido à variabilidade do texto. Isto ocorre pela diferença da realização do mesmo trifone (diferente contexto nas frases) durante o treinamento e teste.

Foi também obtido o desempenho da MFCC extraída de voz reconstruída e voz original. Comparando os resultados da MFCC nestas duas situações (Original vs Reconstruída), pode-se observar que a voz reconstruída provoca uma elevada degradação do desempenho (aproximadamente 14%) quando comparada com a voz original, e é pior também que os atributos de reconhecimento MPCC, MPCEP e MLPCC em todos os experimentos. A MFCC ainda tem uma sensibilidade em torno de 2% para a forma que a excitação é codificada e decodificada

(comparando a linha dois – operação em 5,3 kbits/s – e a linha três – operação em 6,3 kbits/s – de cada tabela). Isto mostra que a MFCC é muito sensível ao ruído de codificação. Os melhores resultados são obtidos dos atributos de reconhecimento MPCEP.

B. Interpolação Linear versus Interpolação com Filtro Digital

Os experimentos deste item foram conduzidos num cenário de independência de locutor e de texto (Fig. 1.4, do capítulo 1 – cenário 3), que é o cenário que melhor aproxima o uso prático de uso de sistemas de reconhecimento de voz distribuídos. A Tab. 5.4 (linhas 4 até 6) compara os resultados de reconhecimento para a interpolação linear e com filtro digital de atributos usando o codificador ITU-T G.723.1.

Atributos	Interpolação Linear			Interpolação com Filtro Digital		
	$\overline{WRR}$	σ	Intervalo de confiança	$\overline{WRR}$	σ	Intervalo de confiança
MFCC - Voz Original	76,82%	1,63%	[76,11% ; 77,54%]	76,82%	1,63%	[76,11% ; 77,54%]
MFCC - Voz Reconstruída (5,3 kbits/s)	62,21%	1,74%	[61,45% ; 62,97%]	62,21%	1,74%	[61,45% ; 62,97%]
MFCC - Voz Reconstruída (6,3 kbits/s)	63,94%	1,68%	[63,20% ; 64,68%]	63,94%	1,68%	[63,20% ; 64,68%]
MPCC - Interp. 33 Hz para 100 Hz	66,31%	1,64%	[65,59% ; 67,03%]	70,21%	1,59%	[69,51% ; 70,91%]
MPCEP - Interp. 33 Hz para 100 Hz	67,19%	1,61%	[66,49% ; 67,90%]	71,32%	1,57%	[70,63% ; 72,01%]
MLPCC - Interp. 33 Hz para 100 Hz	66,83%	1,67%	[66,10% ; 67,56%]	70,85%	1,60%	[70,15% ; 71,55%]

Tabela 5.4 – Tabela de desempenho de reconhecimento para interpolação linear e filtro digital

Comparando os resultados apresentados na Tab. 5.4, pode-se observar que a interpolação das LSFs decodificadas usando filtros digitais pode gerar uma melhoria considerável na taxa de reconhecimento em sistemas distribuídos. Este ganho de desempenho é de aproximadamente 4% quando comparado com o procedimento usual de interpolação linear. A melhor taxa de reconhecimento foi alcançada pelo atributo MPCEP (71,32%), usando interpolação de LSFs decodificadas com filtro digital. É importante lembrar que a MFCC é gerada da voz original e reconstruída com uma duração de quadro de 25 ms (com sobreposição de quadro de forma que os atributos sejam gerados a cada 10 ms). Nenhuma interpolação então é necessária, pelo fato que a MFCC pode ser extraída diretamente da voz original e da voz decodificada na taxa de 100 Hz. Desta tabela pode-se ainda observar que quando a MFCC é utilizada, no caso de voz reconstruída, existe uma grande degradação para a voz original. Ainda se pode

concluir que a MFCC de voz reconstruída é pior que os atributos obtidos dos parâmetros LSF e LPC (MPCC, MPCEP e MLPCC).

5.4. Resultados de Simulação para o Codec AMR-NB

Para facilitar o entendimento dos resultados e evitar que o leitor tenha que se referir aos capítulos anteriores desta tese, é apresentado aqui um breve resumo do codificador AMR-NB de forma a facilitar o entendimento sobre os resultados obtidos e a comparação com os resultados da seção anterior.

O codec de AMR-NB opera-se nas seguintes taxas de bits: 4,75, 5,15, 5,9, 6,7, 7,4, 7,95, 10,2 e 12,2 kbit/s. O AMR-NB é um codificador do tipo ACELP [40]. Opera sobre quadros de voz de 20 ms que correspondem a 160 amostras na frequência de amostragem de 8 kHz. A análise LP (*Linear Prediction*) é executada duas vezes por quadro para a taxa do codificador de 12,2 kbit/s e uma vez para as outras taxas. Para a taxa de 12,2 kbit/s, os dois conjuntos de parâmetros LP são convertidos para dois conjuntos de 10 LSFs os quais são conjuntamente quantizados usando-se um Split Matrix Quantization (SMQ) com 38 bits/quadro. Para as outras taxas, o único conjunto de parâmetros LP é convertido para 10 LSFs e quantizado com um Split Vector Quantization. Em 10,2, 7,4, 6,7 e 5,9 as LSFs são quantizadas com 26 bits/quadro e em 7,95 kbit/s as LSFs são codificadas com 27 bits/quadro. Nas taxas de 5,15 e 4,75 as LSFs são quantizadas com 23 bits/quadro. Note-se que as diferentes taxas de bits deste codec são geralmente chamadas de modos. A padronização do AMR-NB em 1999 [79] como o codec de voz do GSM representou uma melhoria grande da qualidade da voz para as redes móveis. O codec AMR-NB foi adotado também em 1999 por 3GPP como o codec de voz para o sistema de WCDMA (*Wideband Code Division Multiple Access*) 3G. O codec AMR foi desenvolvido conjuntamente pela Ericsson, Nokia e Siemens.

Em todos os experimentos desta seção, o extrator de atributos gera um conjunto de 10 parâmetros, mais suas primeira e segunda derivadas, representando um total de 30 atributos de reconhecimento. O modelo acústico usa HMMs contínuas de três estados com a mistura de 20 Gaussianas por estado para modelar o fone. Como o silêncio é estacionário, um estado foi utilizado com o mesmo

número de gaussianas. Os mesmos foram implementados usando o HTK (*HMM Toolkit*) software [77]. Trifones inter e intra palavra foram usados como unidades acústicas. O modelo de linguagem trigrama foi treinado usando o HTK (*HMM Toolkit*) software [77] com um léxico de 60.080 palavras obtidas de 240.000 frases extraídas de um grande corpus de textos do Ceten-Folha [25]. O modelo de linguagem trigrama foi implementado utilizando o ATK (*Application Toolkit for HTK*) [78].

Nas simulações com o codificador AMR-NB avaliou-se a técnica de interpolação utilizando filtros digitais, porém se restringindo ao uso da base de vozes no cenário mais próximo da utilização real de um sistema de reconhecimento de voz distribuído, que é o reconhecimento independente do locutor e do texto usado no treinamento (Fig. 1.4, do capítulo 1 – cenário 3). A Tab. 5.5 apresenta os resultados de reconhecimento para a interpolação de atributos usando o codificador AMR-NB.

É importante lembrar na análise dos resultados da Tab. 5.5 que o AMR-NB quando operando em 12,2 kbit/s gera LSFs em 100 Hz, o que evita a necessidade de interpolação das LSFs para esta taxa do codec (para as outras taxas do codec AMR-NB, as LSFs são geradas em 50 Hz e necessitam ser interpoladas para atingir os 100 Hz).

Obtive-se também o desempenho de reconhecimento da MFCC extraída da voz reconstruída e da voz original. Comparando os resultados de MFCC nas duas situações (original versus reconstruída), se pode observar que voz reconstruída tem uma degradação de desempenho elevado (entre 11% e 15%) em comparação à voz original, e é pior também que os atributos de reconhecimento MPCC, MPCEP e MLPCC em todos os experimentos. Isto mostra mais uma vez que o MFCC é muito sensível ao ruído de codificação. Os melhores resultados são obtidos com o atributo de reconhecimento MPCEP.

Atributos	$\overline{WRR}$	σ	Intervalo de Confiança
MFCC - Voz Original (8kHz,13bits)	76,41%	1,65%	[75,69% ; 77,13%]
12,2 kbits/s			
MFCC - Voz Reconstruída	65,23%	1,67%	[64,50% ; 65,96%]
MPCC - Sem Interpolação, 100 Hz	72,97%	1,55%	[72,29% ; 73,65%]
MPCEP - Sem Interpolação, 100 Hz	74,10%	1,53%	[73,42% ; 74,78%]
MLPCC - Sem Interpolação, 100 Hz	73,62%	1,56%	[72,94% ; 74,30%]
10,2 kbits/s			
MFCC - Voz Reconstruída	65,01%	1,70%	[64,26% ; 65,76%]
7,95 kbits/s			
MFCC - Voz Reconstruída	64,21%	1,70%	[63,46% ; 64,96%]
MPCC - Interp, 50 Hz para 100 Hz	71,75%	1,60%	[71,05% ; 72,45%]
MPCEP - Interp, 50 Hz para 100 Hz	72,89%	1,58%	[72,20% ; 73,58%]
MLPCC - Interp, 50 Hz para 100 Hz	72,41%	1,61%	[71,70% ; 73,12%]
7,4 kbits/s			
MFCC - Voz Reconstruída	63,97%	1,72%	[63,22% ; 64,72%]
6,7 kbits/s			
MFCC - Voz Reconstruída	62,71%	1,72%	[61,96% ; 63,46%]
5,9 kbits/s			
MFCC - Voz Reconstruída	62,33%	1,74%	[61,57% ; 63,09%]
10,2, 7,4, 6,7 e 5,9 kbits/s			
MPCC - Interp, 50 Hz para 100 Hz	71,74%	1,61%	[71,03% ; 72,44%]
MPCEP - Interp, 50 Hz para 100 Hz	72,87%	1,58%	[72,18% ; 73,56%]
MLPCC - Interp, 50 Hz para 100 Hz	72,41%	1,63%	[71,70% ; 73,12%]
5,15 kbits/s			
MFCC - Voz Reconstruída	62,02%	1,79%	[61,24% ; 62,80%]
4,75 kbits/s			
MFCC - Voz Reconstruída	61,94%	1,81%	[61,15% ; 62,73%]
5,15 e 4,75 kbits/s			
MPCC - Interp, 50 Hz para 100 Hz	71,37%	1,65%	[70,64% ; 72,10%]
MPCEP - Interp, 50 Hz para 100 Hz	72,53%	1,61%	[71,82% ; 73,24%]
MLPCC - Interp, 50 Hz para 100 Hz	72,11%	1,68%	[71,37% ; 72,85%]

Tabela 5.5 – Tabela de desempenho de reconhecimento para o AMR-NB

Comparando o ITU-T G.723.1 nas taxas 6,3 e 5,3 kbits/s (Tab. 5.4) com o codec AMR-NB (Tab. 5.5) nas taxas similares (6,7 e 5,9 kbits/s) a última fornece um ganho de 1.55% na $\overline{WRR}$. É importante notar que 95% dos intervalos de confiança nestes dois casos não se sobrepõem. É também importante observar que o AMR-NB (Tab. 5.5) em suas taxas mais baixas (5,15 e 4,75 kbits/s) supera o codec de ITU-T G.723.1, que opera em 6,3 e 5,3 kbits/s. Além disso, 95% de seus intervalos de confiança outra vez não se sobrepõem.

Finalmente, deve ser anotado que o LSFs no AMR-NB é codificada em uma taxa de bits mais elevada do que a usada pelo ITU-T G.723.1.

5.5. Resultados de Simulação para o Codec AMR-WB

Para facilitar o entendimento dos resultados e evitar que o leitor tenha que se referir aos capítulos anteriores desta tese, é apresentado aqui um breve resumo

do codificador AMR-WB de forma a facilitar o entendimento sobre os resultados obtidos e a comparação com os resultados da seção anterior.

O codec AMR-WB opera na faixa de 50 a 7000hz e nas seguintes taxas de bits: 23,85, 23,05, 19,85, 18,25, 15,85, 14,25, 12,65, 8,85 e 6,60 kbit/s. O AMR-WB é um codificador do tipo CELP [41]. Opera sobre quadros de voz de 20 ms que correspondem a 320 amostras na frequência de amostragem de 16 kHz. A análise LPC é executada uma vez por quadro. O único conjuntos de parâmetros LPC são convertidos para um conjunto de 16 ISFs os quais são quantizados vetorialmente usando-se um Split-Multistage Vector Quantization (S-MSVQ). Em 23,85, 23,05, 19,85, 18,25, 15,85, 14,25, 12,65 e 8,85 as ISFs são quantizadas com 46 bits/quadro e em 6,60 kbit/s as ISFs são codificadas com 36 bits/quadro. Esses dois cenários serão vistos posteriormente na Tab. 5.6.

Em todos os experimentos desta seção são utilizados o conjunto de 16 parâmetros ISF, mais suas primeira e segunda derivadas, representando um total de 48 atributos de reconhecimento. O modelo acústico usa HMMs contínuas de três estados com a mistura de 20 Gaussianas por estado para modelar o fone. Como o silêncio é estacionário, um estado foi utilizado com o mesmo número de gaussianas. Os mesmos foram implementados usando o HTK (*HMM Toolkit*) software [77]. Trifones inter e intra palavra foram usados como unidades acústicas. O modelo de linguagem trigrama foi treinado usando o HTK (*HMM Toolkit*) software [77] com um léxico de 60.080 palavras obtidas de 240.000 frases extraídas de um grande corpus de textos do Ceten-Folha [25]. O modelo de linguagem trigrama foi implementado utilizando o ATK (*Application Toolkit for HTK*) [78].

Nas simulações com o codificador AMR-WB utilizou-se a técnica de interpolação utilizando filtros digitais, se restringindo ao uso da base de vozes no cenário mais próximo da utilização real de um sistema de reconhecimento de voz distribuído, que é o reconhecimento independente do locutor e do texto usado no treinamento (Fig. 1.4, do capítulo 1 – cenário 3). A Tab. 5.6 apresenta os resultados de reconhecimento para a interpolação de ISFs usando o codificador AMR-WB.

Obtive-se também o desempenho de reconhecimento da MFCC extraída da voz reconstruída e da voz original. Comparando os resultados de MFCC nas duas situações (original versus reconstruída), se pode observar que voz reconstruída

tem uma degradação de desempenho elevado (entre 8% e 12%) em comparação à voz original, porém é melhor que os ISFs quando utilizados para reconhecimento em todos os experimentos.

Atributos	$\overline{WRR}$	σ	Intervalo de Confiança
MFCC - Voz Original (16kHz,14bits)	80,19%	1,11%	[79,70% ; 80,68%]
23,85, 23,05, 19,85, 18,25, 15,85, 14,25, 12,65 e 8,85 kbits/s			
ISF - Interp, 50 Hz para 100 Hz	51,13%	1,91%	[50,29% ; 51,97%]
23,85 kbits/s			
MFCC - Voz Reconstruída	72,15%	1,42%	[71,53% ; 72,77%]
23,05 kbits/s			
MFCC - Voz Reconstruída	72,02%	1,42%	[71,40% ; 72,64%]
19,85 kbits/s			
MFCC - Voz Reconstruída	71,56%	1,50%	[70,90% ; 72,22%]
18,25 kbits/s			
MFCC - Voz Reconstruída	71,19%	1,54%	[70,52% ; 71,87%]
15,85 kbits/s			
MFCC - Voz Reconstruída	70,65%	1,57%	[69,96% ; 71,34%]
14,25 kbits/s			
MFCC - Voz Reconstruída	70,03%	1,61%	[69,33% ; 70,74%]
12,65 kbits/s			
MFCC - Voz Reconstruída	69,83%	1,62%	[69,12% ; 70,54%]
8,85 kbits/s			
MFCC - Voz Reconstruída	68,97%	1,66%	[68,24% ; 69,70%]
6,60 kbits/s			
ISF - Interp, 50 Hz para 100 Hz	44,72%	2,02%	[43,84% ; 45,61%]
6,60 kbits/s			
MFCC - Voz Reconstruída	68,32%	1,68%	[67,59% ; 69,06%]

Tabela 5.6 – Tabela de desempenho de reconhecimento para o AMR-WB

Finalmente, pode se concluir que as ISFs são inadequados para uso como atributo de reconhecimento de voz, tendo seu desempenho sido superado inclusive pela MFCC de voz reconstruída.

5.6. Conclusão

No primeiro conjunto de testes para o ITU-T G.723.1, foram realizados alguns experimentos importantes para o cenário de reconhecimento de voz contínua distribuído em amplo vocabulário para o Português Brasileiro. Foi mostrado que apenas a independência do locutor deteriora em torno de 4% a taxa de reconhecimento. Adicionalmente, uma redução de 6% no desempenho foi observada pelo uso de diferentes frases no treinamento e testes do sistema. Mostrou-se também que a MFCC, obtida de voz reconstruída, é bastante sensível ao ruído de codificação, reduzindo o desempenho de 14% aproximadamente. Foi possível observar também que a MFCC de voz reconstruída tem uma sensibilidade em torno de 2% pela forma que a excitação é codificada e decodificada (comparando a linha dois – operação em 5,3 kbits/s – e a linha três –

operação em 6,3 kbits/s – de cada tabela). Os atributos obtidos dos parâmetros LSF ou LPC podem prover um melhor desempenho do que os obtidos da voz reconstruída (MFCC). O MPCEP é o melhor atributo para ser utilizado em um sistema LVDCSR (*Large Vocabulary Distributed Continuous Speech Recognition*) empregando o codificador ITU-T G.723.1. O mesmo oferece um melhor $\overline{WRR}$ (*Average Word Recognition Rate*) com menor complexidade.

Para o segundo conjunto de testes mostrou-se que a interpolação usando filtros digitais das LSFs decodificadas melhora significativamente o desempenho de todos os atributos de reconhecimento obtidos do codificador ITU-T G.723.1 quando comparados com a interpolação linear. O $\overline{WRR}$ melhorou de aproximadamente 4% em todas as situações onde os atributos são obtidos de parâmetros LSF e LPC.

Comparando o ITU-T G.723.1 nas taxas 6,3 e 5,3 o kbits/s com o codec AMR-NB nas taxas similares (6,7 e 5,9 kbits/s) a última fornece um ganho de 1.55% na $\overline{WRR}$. É importante notar que 95% dos intervalo de confiança nestes dois casos não se sobrepõem. É também importante observar que o AMR-NB em suas taxas mais baixas (5,15 e 4,75 kbits/s) supera o codec de ITU-T G.723.1, que opera em 6,3 e 5,3 kbits/s. Além disso, 95% de seus intervalos de confiança outra vez não se sobrepõem. Deve ser anotado que as LSFs no AMR-NB são codificadas em uma taxa de bits mais elevada do que a usada pelo ITU-T G.723.1.

Para o terceiro conjunto de testes evidenciou-se que os ISFs do codificador AMR-WB são inadequados para uso como atributo de reconhecimento de voz, tendo seu desempenho sido superado inclusive pela MFCC de voz reconstruída.

No próximo capítulo será abordado 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.

6
Perdas de Pacotes

O problema de perda de pacotes em rajadas nas redes IP e redes móveis é um dos fatores mais importantes a serem considerados na análise de sistemas de reconhecimento de voz distribuídos. Perdas de pacotes em rajadas causam uma redução drástica do desempenho do reconhecimento de voz. Neste capítulo, apresenta-se uma técnica nova para a reconstrução dos pacotes perdidos baseada em Redes Neurais e compara-se seu desempenho de reconhecimento com aqueles obtidos com as técnicas de inserção de zeros e interpolação linear.

Mesmo redes IP e redes móveis celulares sendo bastantes diferentes umas das outras, ambas sofrem de perdas de pacotes em rajadas. Em redes móveis celulares acontecem em momentos de atenuação acentuada do sinal (*signal fading*) enquanto em redes IP as perdas de pacote ou descartes ocorrem devido ao congestionamento.

Para levar em consideração as características do processo de perda em rajadas, o mesmo foi aproximado usando um processo Markoviano de dois estados, também conhecido como modelo de Gilbert [80]. Os estados se referem aos eventos de "PACOTES RECEBIDOS" e "PACOTES PERDIDOS". Como mostrado na Fig. 6.1, p denota a probabilidade de transição do estado de "PACOTES RECEBIDOS" para o estado de "PACOTES PERDIDOS" e q denota a probabilidade de transição do estado "PACOTES PERDIDOS" para o estado de "PACOTES RECEBIDOS".

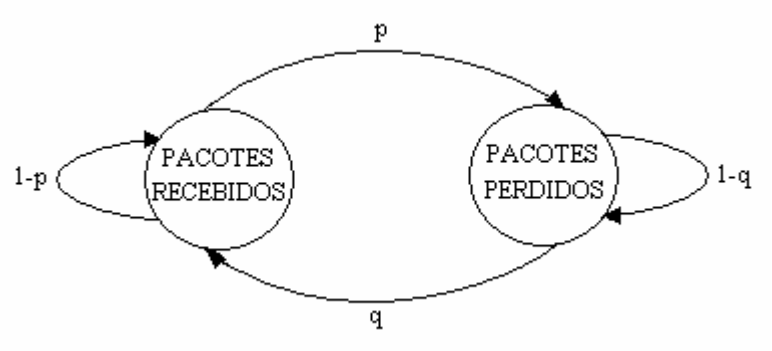

Figura 6.1 – Modelo de Gilbert

A Taxa de Perda de Pacote (TPP), também conhecida como Probabilidade de Perda Incondicional (PPI), é dada por

$$TPP = \frac{p}{p+q} \tag{6.1}$$

e a rajada de perda de pacote é medida pela Comprimento Médio da Rajada (CMR) que é dada por

$$CMR = \frac{1}{1-PCP} \tag{6.2}$$

onde PCP é a Probabilidade Condicional de Perda, que é a probabilidade de permanência no estado de "PACOTES PERDIDOS" (i.e. $PCP = 1-q$) [81]. O modelo de perda de pacote foi simulado neste capítulo com as condições de rede usadas em [9] e apresentadas na Tab. 6.1.

TPP(%)	PCP	CMR	p	q
0	-	-	0	0
10	0.15	1.18	0.10	0.85
20	0.30	1.43	0.20	0.70
30	0.35	1.54	0.30	0.65
40	0.50	2.00	0.30	0.50

Tabela 6.1 –Tabela de condições de rede para o modelo de Gilbert utilizado nas simulações

Na seção 6.1 deste capítulo é apresentada a reconstrução de pacotes perdidos usando Inserção de Zeros e Interpolação Linear. Na seção 6.2, é proposta uma nova técnica baseada em Redes Neurais, com o objetivo de reconstruir os pacotes perdidos. Na seção 6.3, são analisados os resultados de simulação. Finalmente, a seção 6.4 contém uma breve conclusão.

6.1. Inserção de Zeros e Interpolação Linear

Existem algumas aproximações para melhorar o desempenho do sistema de reconhecimento de voz, na presença de imperfeições do canal. Um exemplo é o apagamento de quadros. Uma solução simples para essa degradação é inserção dos zeros na posição dos pacotes perdidos. Uma outra aproximação é interpolação linear entre pacotes recebidos com sucesso (em nosso caso, quadros). O destino recebe, por exemplo, o primeiro conjunto de LSFs quantizadas. Entretanto, devido às imperfeições do canal, não é recebido o segundo conjunto. Na chegada do terceiro conjunto, o receptor pode aproximar o segundo pela interpolação linear do primeiro conjunto com o terceiro. Certamente, a interpolação de mais de um conjunto é praticável em troca de um incremento indesejável de atraso [80]. Para aplicações de redes IP, se n quadros consecutivos de duração t cada um, são perdidos, o atraso devido à interpolação é $D_i = nt + RTT/2$, onde RTT (*Round-Trip Time*) é o tempo para um pacote ir da fonte ao destino e então de volta à fonte. Valores típicos para RTT variam de 10 a 700 ms e, de acordo com [80], atrasos aceitáveis para aplicações de VoIP não devem exceder 800 ms.

É importante notar que a primeira técnica (Inserção Zero) ignora as características do sinal. Consequentemente, não explora o conhecimento do sinal para melhorar o desempenho do reconhecimento. Por outro lado, o uso da segunda técnica (Interpolação Linear) implica geralmente em longo atraso nos pacotes reconstruídos.

6.2.
Redes Neurais

Pelas razões expostas na seção anterior, foi proposta, nesta tese, uma nova técnica baseada em Redes Neurais [82] para reconstrução dos pacotes perdidos, com a vantagem de usar o conhecimento do comportamento do sinal e evitar o retardo significativo para a reconstrução do sinal. O atraso da técnica proposta é somente o tempo das Redes Neurais para computar a saída. Este cálculo está baseado nos quadros de LSFs recebidos antes do pacote perdido ou das LSFs interpoladas obtidas antes do pacote perdido que se deseja recuperar.

Na Fig. 6.2 é apresentada a topologia das Redes Neurais escolhida baseado em resultados de simulações obtidas em uma série de estudos preliminares e pela conhecido uso das Redes Neurais *Multi-Layer Perceptron* na previsão de séries temporais [82]. A camada escondida é composta de 3 neurônios (foram testados 1 à 8 neurônios na camada escondida, tendo o com 3 neurônios apresentado melhor desempenho) cuja função selecionada para o neurônio foi a tangente hiperbólica (não foi testada outra função ativação para a camada escondida) . A função linear foi selecionada para o neurônio da camada da saída (não foram testadas outras funções de ativação para o neurônio da camada de saída). Foram utilizadas 10 Redes Neurais com esta topologia, cada uma delas usada para cada uma das 10 LSFs de cada quadro. As 4 entradas de cada Rede Neural são os valores das LSFs ou de seus valores reconstruídos em $T-4$, $T-3$, $T-2$ e $T-1$ (foram testadas janelas de 1 à 8 amostras passadas – o que levou a determinação do número de entradas da rede neural, – tendo a de quatro apresentado melhor desempenho) onde T é o instante em que um quadro é perdido. A saída é a LSF reconstruída em T. Ressalta-se, portanto, que este valor da LSF será usado no sistema de reconhecimento de voz e como uma entrada da rede neural se a LSF de $T+1$ for perdida também. Cada uma das 10 Redes Neurais são treinadas inicialmente com a mesma base de dados usada no treinamento do HMMs (*Hidden Markov Models*). É interessante observar que quando são recebidos 5 quadros sucessivamente com sucesso, são usados os primeiros 4 pacotes como entradas das Redes Neurais e o quinto pacote como sua saída. Este procedimento tem como

única finalidade re-treinar (re-estimar) as Redes Neurais, sendo realizado durante a fase de teste (“on-line”).

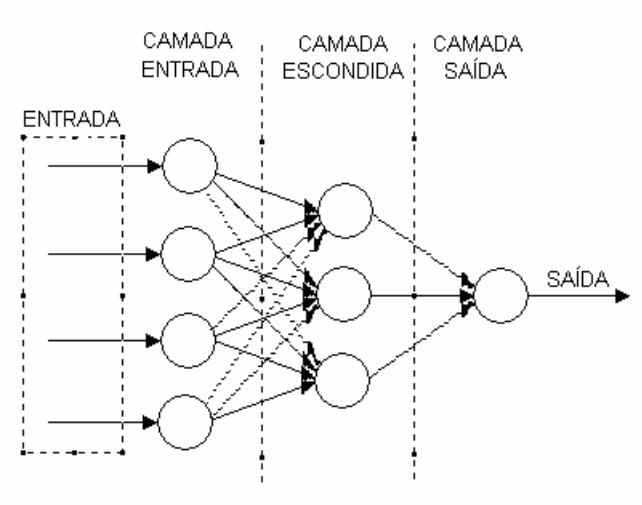

Figura 6.2 – Topologia da Rede Neural

Para o treinamento da rede Neural foi utilizado o algoritmo do Backpropagation (que utiliza a técnica do Gradiente Descendente para ajuste dos pesos da rede neural). Os parâmetros utilizados no treinamento foram:

- Taxa de aprendizado - 0,001
- Número de épocas de Treinamento – 5000
- Medida de erro - Mean Square Error (MSE) - valor alvo 0,0001

Não foi utilizado o fator de Momentum, pois não verificou-se durante o treinamento a captura da rede por mínimos locais que justificassem a utilização do mesmo. Não foi realizada validação cruzada pois devido ao grande volume de dados da base, avaliou-se que era pequeno o risco de overfitting (excesso de treinamento prejudicando a capacidade de generalização) da rede neural.

Não utilizou-se também o Filtro Digital, proposto no capítulo 5, para realizar a recuperação dos pacotes perdidos, pois o mesmo se baseia no espectro do sinal, sendo que neste caso o sinal já havia sido danificado pela perda de pacotes. Neste caso deseja-se recuperar esta informação e não aumentar a taxa com uma informação espectral danificada do sinal, por isso o uso das Redes

Neurais para a realização da restauração dos pacotes perdidos, tentando recuperar a informação perdida.

6.3. Resultados de Simulação para o Codec ITU-T G.723.1 e AMR-NB

Os resultados do desempenho são apresentados em cinco tabelas, onde em cada tabela são mostrados o desempenho de reconhecimento para o atributo MPCEP obtido das LSF em diversas taxas dos codificadores ITU-T G.723.1 e AMR-NB para diferentes condições de rede. A Tab. 6.2 mostra os resultados do reconhecimento para uma rede ideal sem perda dos pacotes. As Tab. 6.3, 6.4, 6.5 e 6.6 mostram os desempenhos de reconhecimento para redes reais com taxas da perda de pacotes *TPP* e comprimento médio das rajadas *CMR* dados por $TPP = 0,\ 10,\ 20,\ 30$ e 40% e $CMR = 0,\ 1{,}18,\ 1{,}43,\ 1{,}54$ e $2{,}00$, respectivamente. Deve-se observar que em cada caso de teste, os parâmetros do modelo são treinados com o mesmo tipo de atributos (o mesmo tipo de reconstrução), isto é, treinamento e teste estão casados neste sentido. É também importante lembrar que o AMR-NB quando operando em 12,2 kbit/s gera LSFs em 100 Hz, o que evita a necessidade de interpolação das LSFs para esta taxa do codec (para as outras taxas do codec AMR-NB, as LSFs são geradas em 50 Hz e necessitam ser interpoladas para atingir os 100 Hz).

Nas simulações com os codificadores ITU-T G.723.1 e AMR-NB se utilizou a nova técnica de interpolação de parâmetros, utilizando filtros digitais para a realização do aumento de taxa (Capítulo 5) e se restringiu ao uso da base de vozes no cenário mais próximo da utilização real de um sistema de reconhecimento de voz distribuído, que é o reconhecimento independente do locutor e do texto usado no treinamento (Fig. 1.4, do capítulo 1 – cenário 3).

Atributos	$\overline{WRR}$	σ	Intervalo de Confiança
MPCEP - ITU-T G723,1 (5,3 e 6,3 kbit/s)	71,32%	1,57%	[70,63% ; 72,01%]
MPCEP - AMR-NB (12,2 kbit/s)	74,10%	1,53%	[73,42% ; 74,78%]
MPCEP - AMR-NB (10,2, 7,40, 6,70 e 5,90 kbit/s)	72,87%	1,58%	[72,18% ; 73,56%]
MPCEP - AMR-NB (7,95 kbit/s)	72,89%	1,58%	[72,20% ; 73,58%]
MPCEP - AMR-NB (5,15 e 4,75 kbit/s)	72,53%	1,61%	[71,82% ; 73,24%]

Tabela 6.2 – Tabela de desempenho de reconhecimento para redes sem perdas de pacotes (TPP= 0% e CMR=0)

Atributos	$\overline{WRR}$	σ	Intervalo de Confiança
Inserção de Zeros			
MPCEP - ITU-T G723,1 (5,3 e 6,3 kbit/s)	66,21%	1,59%	[65,51% ; 66,91%]
MPCEP - AMR-NB (12,2 kbit/s)	70,01%	1,54%	[69,34% ; 70,69%]
MPCEP - AMR-NB (10,2, 7,40, 6,70 e 5,90 kbit/s)	68,12%	1,60%	[67,42% ; 68,82%]
MPCEP - AMR-NB (7,95 kbit/s)	68,15%	1,59%	[67,45% ; 68,85%]
MPCEP - AMR-NB (5,15 e 4,75 kbit/s)	67,09%	1,64%	[66,37% ; 67,81%]
Interpolação Linear			
MPCEP - ITU-T G723,1 (5,3 e 6,3 kbit/s)	67,52%	1,59%	[66,82% ; 68,22%]
MPCEP - AMR-NB (12,2 kbit/s)	71,45%	1,54%	[70,78% ; 72,13%]
MPCEP - AMR-NB (10,2, 7,40, 6,70 e 5,90 kbit/s)	69,53%	1,59%	[68,83% ; 70,23%]
MPCEP - AMR-NB (7,95 kbit/s)	69,55%	1,59%	[68,85% ; 70,25%]
MPCEP - AMR-NB (5,15 e 4,75 kbit/s)	68,51%	1,63%	[67,80% ; 69,23%]
Redes Neurais			
MPCEP - ITU-T G723,1 (5,3 e 6,3 kbit/s)	67,54%	1,58%	[66,85% ; 68,23%]
MPCEP - AMR-NB (12,2 kbit/s)	71,49%	1,54%	[70,82% ; 72,17%]
MPCEP - AMR-NB (10,2, 7,40, 6,70 e 5,90 kbit/s)	69,54%	1,59%	[68,84% ; 70,24%]
MPCEP - AMR-NB (7,95 kbit/s)	69,58%	1,59%	[68,88% ; 70,28%]
MPCEP - AMR-NB (5,15 e 4,75 kbit/s)	68,52%	1,62%	[67,81% ; 69,23%]

Tabela 6.3 – Tabela de desempenho de reconhecimento para rede com TPP= 10% e CMR=1,18

Atributos	$\overline{WRR}$	σ	Intervalo de Confiança
Inserção de Zeros			
MPCEP - ITU-T G723,1 (5,3 e 6,3 kbit/s)	61,57%	1,64%	[60,85% ; 62,29%]
MPCEP - AMR-NB (12,2 kbit/s)	65,82%	1,58%	[65,13% ; 66,51%]
MPCEP - AMR-NB (10,2, 7,40, 6,70 e 5,90 kbit/s)	63,71%	1,65%	[62,99% ; 64,43%]
MPCEP - AMR-NB (7,95 kbit/s)	63,77%	1,64%	[63,05% ; 64,49%]
MPCEP - AMR-NB (5,15 e 4,75 kbit/s)	62,22%	1,70%	[61,48% ; 62,97%]
Interpolação Linear			
MPCEP - ITU-T G723,1 (5,3 e 6,3 kbit/s)	62,63%	1,63%	[61,92% ; 63,35%]
MPCEP - AMR-NB (12,2 kbit/s)	66,84%	1,58%	[66,15% ; 67,53%]
MPCEP - AMR-NB (10,2, 7,40, 6,70 e 5,90 kbit/s)	64,72%	1,65%	[64,00% ; 65,44%]
MPCEP - AMR-NB (7,95 kbit/s)	64,78%	1,63%	[64,07% ; 65,50%]
MPCEP - AMR-NB (5,15 e 4,75 kbit/s)	63,21%	1,70%	[62,47% ; 63,96%]
Redes Neurais			
MPCEP - ITU-T G723,1 (5,3 e 6,3 kbit/s)	63,12%	1,61%	[62,42% ; 63,83%]
MPCEP - AMR-NB (12,2 kbit/s)	67,37%	1,56%	[66,69% ; 68,06%]
MPCEP - AMR-NB (10,2, 7,40, 6,70 e 5,90 kbit/s)	65,28%	1,64%	[64,56% ; 66,00%]
MPCEP - AMR-NB (7,95 kbit/s)	65,34%	1,61%	[64,64% ; 66,05%]
MPCEP - AMR-NB (5,15 e 4,75 kbit/s)	63,71%	1,69%	[62,97% ; 64,45%]

Tabela 6.4 – Tabela de desempenho de reconhecimento para rede com TPP= 20% e CMR=1,43

Atributos	$\overline{WRR}$	σ	Intervalo de Confiança
Inserção de Zeros			
MPCEP - ITU-T G723,1 (5,3 e 6,3 kbit/s)	57,79%	1,68%	[57,06% ; 58,53%]
MPCEP - AMR-NB (12,2 kbit/s)	62,01%	1,63%	[61,30% ; 62,73%]
MPCEP - AMR-NB (10,2, 7,40, 6,70 e 5,90 kbit/s)	59,64%	1,69%	[58,90% ; 60,38%]
MPCEP - AMR-NB (7,95 kbit/s)	59,72%	1,68%	[58,99% ; 60,46%]
MPCEP - AMR-NB (5,15 e 4,75 kbit/s)	58,10%	1,75%	[57,33% ; 58,87%]
Interpolação Linear			
MPCEP - ITU-T G723,1 (5,3 e 6,3 kbit/s)	58,57%	1,68%	[57,84% ; 59,31%]
MPCEP - AMR-NB (12,2 kbit/s)	62,81%	1,63%	[62,10% ; 63,53%]
MPCEP - AMR-NB (10,2, 7,40, 6,70 e 5,90 kbit/s)	60,27%	1,69%	[59,53% ; 61,01%]
MPCEP - AMR-NB (7,95 kbit/s)	60,37%	1,68%	[59,64% ; 61,11%]
MPCEP - AMR-NB (5,15 e 4,75 kbit/s)	58,68%	1,75%	[57,91% ; 59,45%]
Redes Neurais			
MPCEP - ITU-T G723,1 (5,3 e 6,3 kbit/s)	59,93%	1,66%	[59,20% ; 60,66%]
MPCEP - AMR-NB (12,2 kbit/s)	64,49%	1,60%	[63,79% ; 65,19%]
MPCEP - AMR-NB (10,2, 7,40, 6,70 e 5,90 kbit/s)	63,91%	1,67%	[63,18% ; 64,64%]
MPCEP - AMR-NB (7,95 kbit/s)	63,99%	1,66%	[63,26% ; 64,72%]
MPCEP - AMR-NB (5,15 e 4,75 kbit/s)	62,41%	1,73%	[61,65% ; 63,17%]

Tabela 6.5 – Tabela de desempenho de reconhecimento para rede com TPP= 30% e CMR=1,54

Atributos	$\overline{WRR}$	σ	Intervalo de Confiança
Inserção de Zeros			
MPCEP - ITU-T G723,1 (5,3 e 6,3 kbit/s)	49,20%	1,75%	[48,43% ; 49,97%]
MPCEP - AMR-NB (12,2 kbit/s)	56,40%	1,70%	[55,66% ; 57,15%]
MPCEP - AMR-NB (10,2, 7,40, 6,70 e 5,90 kbit/s)	52,99%	1,77%	[52,22% ; 53,77%]
MPCEP - AMR-NB (7,95 kbit/s)	53,13%	1,75%	[52,36% ; 53,90%]
MPCEP - AMR-NB (5,15 e 4,75 kbit/s)	51,06%	1,84%	[50,25% ; 51,87%]
Interpolação Linear			
MPCEP - ITU-T G723,1 (5,3 e 6,3 kbit/s)	49,31%	1,75%	[48,54% ; 50,08%]
MPCEP - AMR-NB (12,2 kbit/s)	56,59%	1,70%	[55,85% ; 57,34%]
MPCEP - AMR-NB (10,2, 7,40, 6,70 e 5,90 kbit/s)	53,27%	1,76%	[52,50% ; 54,04%]
MPCEP - AMR-NB (7,95 kbit/s)	53,37%	1,75%	[52,60% ; 54,14%]
MPCEP - AMR-NB (5,15 e 4,75 kbit/s)	51,32%	1,84%	[50,51% ; 52,13%]
Redes Neurais			
MPCEP - ITU-T G723,1 (5,3 e 6,3 kbit/s)	52,22%	1,71%	[51,47% ; 52,97%]
MPCEP - AMR-NB (12,2 kbit/s)	59,47%	1,65%	[58,75% ; 60,19%]
MPCEP - AMR-NB (10,2, 7,40, 6,70 e 5,90 kbit/s)	56,04%	1,72%	[55,29% ; 56,80%]
MPCEP - AMR-NB (7,95 kbit/s)	56,14%	1,70%	[55,40% ; 56,89%]
MPCEP - AMR-NB (5,15 e 4,75 kbit/s)	54,07%	1,79%	[53,29% ; 54,86%]

Tabela 6.6 – Tabela de desempenho de reconhecimento para rede com TPP= 40% e CMR=2,00

Dos resultados da simulação fica claro que a Inserção de Zeros é definitivamente a pior aproximação para a solução da perda de pacotes. Agora comparando a Inserção de Zeros, a Interpolação Linear e a proposta de Redes Neurais para a reconstrução de pacotes perdidos de LSFs nas Tab. 6.3, 6.4, 6.5 e 6.6, pode-se ver que a técnica proposta que usa Redes Neurais supera as duas outras técnicas em todos os casos. Entretanto, as melhorias são somente significativas nas Tab. 6.5 e 6.6, correspondendo à perda de pacotes - TPP - de 30% e 40%, respectivamente, onde as rajadas de perdas de pacotes das redes IP e das redes móveis celulares são mais severas. No caso onde $TPP = 40\%$ (Tab. 6.5), o esquema novo fornece ganhos de reconhecimento de aproximadamente 3% quando comparado com a técnica da Interpolação Linear.

Comparando o codec ITU-T G.723.1 nas taxas 6,3 e 5,3 kbit/s com o codec AMR-NB nas taxas similares (6,7 e 5,9 kbit/s), em toda as condições de rede, o AMR-NB fornece um ganho em torno de 1,50%. É também muito significativo notar que os intervalos de confiança de 95% não se sobrepõem. Além disso, é importante observar que o AMR-NB em suas taxas mais baixas (5,15 e 4,75 kbit/s) supera o codec de ITU-T G.723.1 que opera em 6,3 e 5,3 kbit/s para todos os valores de TPP. Outra vez, seus intervalos de confiança de 95% não se sobrepõem. Note que as LSFs, de onde os atributos de reconhecimento são extraídos, são codificados em uma taxa de bits mais elevada pelo AMR-NB em

comparação ao ITU-T G.723.1. Finalmente, está claro que as Redes Neurais são uma técnica atrativa para a reconstrução de pacotes perdidos para ambos os codificadores de voz.

6.4. Conclusão

Neste capítulo, foram realizadas diversas experiências importantes em Reconhecimento de Voz Contínuo Distribuído com amplo vocabulário no Português Brasileiro. Foi proposto o uso de Redes Neurais para a reconstrução de pacotes perdidos em sistemas Móveis e redes IP. Comparando com a Inserção de Zeros e a técnica de Interpolação Linear, as Redes Neurais mostraram ser o melhor método para reconstruir pacotes perdidos em sistemas de Reconhecimento de Voz Distribuído que empreguem os codecs ITU-T G.723.1 ou AMR-NB, especialmente em condições severas da perda do pacote. Além disso, foi mostrado que o AMR-NB que opera em uma taxa de bits mais baixa supera o codec ITU-T G.723.1 nas taxas de reconhecimento, sem sobreposição dos seus intervalos de confiança em 95%, em todas as condições da rede.

No próximo capítulo serão apresentadas as conclusões finais e as sugestões para trabalhos futuros.

7
Conclusões e Sugestões para Trabalhos Futuros

A grande motivação para o presente trabalho foi a de propor técnicas que fossem capazes de superar alguns dos obstáculos encontrados para que os reconhecedores de voz contínua distribuída com amplo vocabulário tenham um bom desempenho.

7.1.
Conclusões

A construção da base de vozes (base Alcaim – Alencar) foi a primeira contribuição relevante para o desenvolvimento de sistemas de reconhecimento de voz contínua para o Português Brasileiro obtido nesta tese. A Base de voz construída 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]). 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 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. Os locutores desta base têm idades variando de 17 à 65 anos, sendo alguns deles profissionais na realização de locuções, outros treinados para a gravação de livros falados e outros sem nenhum conhecimento teórico ou prático sendo orientados durante a gravação. 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 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.).

Para um bom funcionamento dos Sistemas Automáticos de Reconhecimento de voz é necessário que os atributos de reconhecimento sejam obtidos a uma taxa elevada, porém os codificadores de Voz para Telefonia IP e Móvel Celular, usados em cenários distribuídos, normalmente geram seus parâmetros a taxas mais baixas, o que degrada o desempenho do reconhecedor. Usualmente é utilizada a interpolação linear no domínio das LSFs para resolver este problema. Nesta Tese foi proposta a realização da interpolação com a utilização de um Filtro Digital Interpolador que demonstrou ter um desempenho de reconhecimento muito superior ao da interpolação linear.

No primeiro conjunto de testes para o ITU-T G.723.1, foram realizados alguns experimentos importantes para o cenário de reconhecimento de voz contínua distribuído em amplo vocabulário para o Português Brasileiro. Foi mostrado que apenas a independência do locutor deteriora em torno de 4% a taxa de reconhecimento. Adicionalmente, uma redução de 6% no desempenho foi observada pelo uso de diferentes frases no treinamento e testes do sistema. Mostrou-se também que a MFCC, obtida de voz reconstruída, é bastante sensível ao ruído de codificação, reduzindo o desempenho de 14% aproximadamente. Foi possível observar também que a MFCC de voz reconstruída tem uma sensibilidade em torno de 2% pela forma que a excitação é codificada e decodificada (comparando a linha dois – operação em 5,3 kbits/s – e a linha três – operação em 6,3 kbits/s – de cada tabela do capítulo 5). Os atributos obtidos dos parâmetros LSF ou LPC podem prover um melhor desempenho do que os obtidos da voz reconstruída (MFCC). O MPCEP é o melhor atributo para ser utilizado em um sistema LVDCSR (*Large Vocabulary Distributed Continuous Speech Recognition*) empregando o codificador ITU-T G.723.1. O mesmo oferece uma melhor $\overline{WRR}$ (*Average Word Recognition Rate*) com menor complexidade.

Para o segundo conjunto de testes mostrou-se que a interpolação usando filtros digitais das LSFs decodificadas melhora significativamente o desempenho de todos os atributos de reconhecimento obtidos do codificador ITU-T G.723.1 quando comparados com a interpolação linear. O $\overline{WRR}$ melhorou de

aproximadamente 4% em todas as situações onde os atributos são obtidos de parâmetros LSF e LPC.

Comparando o ITU-T G.723.1 nas taxas 6,3 e 5,3 o kbits/s com o codec AMR-NB nas taxas similares (6,7 e 5,9 kbits/s) a última fornece um ganho de 1.55% na $\overline{WRR}$. É importante notar que os intervalo de confiança de 95% nestes dois casos não se sobrepõem. É também importante observar que o AMR-NB em suas taxas mais baixas (5,15 e 4,75 kbits/s) supera o codec de ITU-T G.723.1, que opera em 6,3 e 5,3 kbits/s. Além disso, seus intervalos de confiança de 95% outra vez não se sobrepõem. Deve ser anotado que as LSFs no AMR-NB são codificadas em uma taxa de bits mais elevada do que a usada pelo ITU-T G.723.1.

Para o terceiro conjunto de testes evidenciou-se que os ISFs do codificador AMR-WB são inadequados para uso como atributo de reconhecimento de voz, tendo seu desempenho sido superado inclusive pela MFCC de voz reconstruída.

O problema de perda de pacotes em rajadas nas redes IP e redes móveis é um dos fatores mais importantes a serem considerados na análise de sistemas de reconhecimento de voz distribuídos. Perdas de pacotes em rajadas causam uma redução drástica do desempenho do reconhecimento de voz.

Existem algumas aproximações para melhorar o desempenho do sistema de reconhecimento de voz na presença de imperfeições do canal, tais como apagamentos dos quadros. Uma solução simples é inserção dos zeros na posição dos pacotes perdidos. Uma outra aproximação é interpolação linear, entre pacotes recebidos com sucesso (em nosso caso, quadros). O destino recebe, por exemplo, o primeiro conjunto de LSFs quantizadas. Entretanto, devido às imperfeições do canal, não é recebido o segundo conjunto. Na chegada do terceiro conjunto, o receptor pode aproximar o segundo pela interpolação linear do primeiro conjunto com o terceiro. Certamente, a interpolação de mais de um conjunto é praticável em troca de um incremento indesejável de atraso [80]. Para aplicações de redes IP, se n quadros consecutivos de duração t cada um, é perdido, o atraso devido à interpolação é $D_i = nt + RTT/2$, onde RTT (*Round-Trip Time*) é o tempo para um pacote ir da fonte ao destino e então de volta à fonte. Anote que valores típicos para RTT variam de 10 a 700 ms e de acordo com [80], atrasos aceitáveis para aplicações de VoIP não devem exceder 800 ms.

É importante notar que a primeira técnica (Inserção Zero) ignora as características do sinal. Consequentemente, não explora o conhecimento do sinal para melhorar o desempenho do reconhecimento. Por outro lado, o uso da segunda técnica (Interpolação Linear) implica geralmente em longo atraso nos pacotes reconstruídos.

Foi proposta uma técnica nova para a reconstrução dos pacotes perdidos baseada em Redes Neurais e comparou-se seu desempenho de reconhecimento com os aqueles obtidos com as técnicas de inserção de zeros e interpolação linear.

Dos resultados da simulação fica claro que a Inserção de Zeros é definitivamente a pior aproximação para a solução da perda de pacotes. Comparando a Inserção de Zeros, a Interpolação Linear e a Redes Neurais para a reconstrução de pacotes perdidos de LSFs, pode-se ver que a técnica proposta que usa Redes Neurais supera as duas outras técnicas em todos os casos. Entretanto, as melhorias são somente significativas nas perdas de pacotes - *TPP* - de 30% e 40%, respectivamente, onde as condições das redes IP e das redes móveis são mais severas. No caso onde $TPP = 40\%$, o esquema novo fornece ganhos de reconhecimento de aproximadamente 3% quando comparado com a técnica da Interpolação Linear.

Agora comparando o ITU-T G.723.1 nas taxas 6,3 e 5,3 kbit/s com o codec de AMR-NB nas taxas similares (6,7 e 5,9 kbit/s), em toda as condições de rede, o AMR-NB fornece um ganho em torno de 1,50%. É também muito significativo notar que os intervalos de confiança de 95% não se sobrepõem. Além disso, é importante observar que o AMR-NB em suas taxas mais baixas (5,15 e 4,75 kbit/s) supera o codec de ITU-T G.723.1 que opera em 6,3 e 5,3 kbit/s para todos os valores de *TPP*. Outra vez, seus intervalos de confiança de 95% não se sobrepõem. Note que as LSFs de onde os atributos de reconhecimento são extraídos, são codificados em uma taxa de bits mais elevada pelo AMR-NB em comparação ao ITU-T G.723.1. Finalmente, ficou claro que as Redes Neurais representam uma técnica atrativa para a reconstrução de pacotes perdidos para ambos os codificadores de voz.

7.2. Sugestões para Trabalhos Futuros

Uma primeira sugestão interessante seria obter matematicamente e testar atributos de reconhecimento para o codificador AMR-WB diretamente dos parâmetros ISFs, que eventualmente apresentassem um desempenho semelhante ou superior ao obtido com os atributos obtidos de LSFs.

Fica também como sugestão, para trabalhos futuros, a busca por técnicas que possam melhorar o desempenho de sistemas de reconhecimento de voz distribuído na presença de ruído ambiente.

Uma experiência interessante a ser realizada consiste em verificar o comportamento de atributos robustos ao ruído, como por exemplo o ZCPA [10], quando obtidos de voz reconstruída por um decodificador padrão e compará-los com os melhores atributos aqui obtidos.

Um outro caminho a ser validado para a melhoria do desempenho do decodificador é a fusão de diferentes atributos de reconhecimento que possam de alguma forma agregar informações diferentes sobre a voz a ser reconhecida.

Referências Bibliográficas

[1] Choi, H. S., Kim, H. K. and Lee, H. S., “Speech Recognition Using Quantized LSP Parameters and their Transformations in Digital Communication,” Speech Communication, vol. 30, pp. 223-233, 2000.

[2] Cirigliano, R. J. R., Monteiro, C., Barbosa, F. L., Resende Junior, F. G. V., Couto, L. R., Moraes, J. A., “Um Conjunto de 1000 Frases Foneticamente Balanceadas para o Português Brasileiro obtido utilizando a abordagem de Algoritmos Genéticos”, pp. 544-549 XXII Simpósio Brasileiro de Telecomunicações, 2005.

[3] Peinado, A. M., Sánchez, V., Pérez-Córdoba, J. L. and Torre, A., “HMM – based channel error mitigation and its application to distributed speech recognition,” Speech Communication, vol.41, pp. 549-561, March, 2003.

[4] Kim, H. K., Choi, S. H. and Lee, H. S., “On Approximating Line Spectral Frequencies to LPC Cepstral Coefficients,” IEEE Trans. Speech and Audio Processing, vol. 8, pp. 195 – 199, March 2000.

[5] Ohshima, Y., “Environmental Robustness in Speech Recognition using Physiologically-Motivated Signal Processing,” PH. D. Thesis, Carnegie Mellon University, Pittsburgh, Pennsylvanya, December 1993.

[6] Kim, H. K., and Cox, R. V., “A Bitstream-Based Front-End for Wireless Speech recognition on IS-136 Communications System,” IEEE Trans. Speech and Audio Processing, vol. 9, pp. 558-568, July, 2001.

[7] Davis, S. B. and Mermelstein. P., “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” IEEE Trans. Acoust., Speech, Signal Processing, vol. 28, pp. 357-366, 1980.

[8] Shannon, B. J., Paliwal, K. K., “Feature extraction from higher-lag autocorrelation coefficients for robust speech recognition,” Speech Communication, vol. 48, pp. 1458-1485, 2006.

[9] You, K.-H. e Wang, H.-C., "Robust Features Derived from Temporal Trajectory Filtering for Speech Recognition under the Corruption of Additive and Convolutional Noises," ICASSP'98, pp.577-580, 1998.

[10] Kim, D.-S., Lee, S.-Y. and Kil, R. M., “Auditory Processing of Speech Signals for Robust Speech Recognition in Real-World Noisy Environments,” IEEE Trans. Speech and Audio Processing, vol. 7, pp. 55- 69, January, 1999.

[11] Ghulam, M., et al, “Pitch-synchronous ZCPA (PS-ZCPA)-based feature extraction with auditory masking,” Proc. ICASSP05, vol. 1, pp. 517-520, 2005.

[12] Ghulam, M., Horikama, J., and Nitta, T. “A Pitch-synchronous peak-amplitude based feature extraction method for noise robust ASR,” Proc. ICASSP06, vol. 1, pp. 505-608, 2006

[13] Gajic, B. and Paliwal, K. K., "Robust feature extraction using subband spectral centroid histograms," IEEE Int. Conf. on Acoust., Speech and Signal Processing, vol. 1, pp. 85-88, Salt Lake City, USA, May, 2001.

[14] Baker, J. K., "The Dragon System – an overview," IEEE Trans. ASSP, vol. 23(1), pp. 24-29, 1975.

[15] Jelinek, F., "Continuous Speech Recognition by Statistical Methods," Proc. IEEE, vol. 64 (4), pp. 532-556, 1976.

[16] Tevah, R. T., "Implementação de um Sistema de Reconhecimento de Fala Contínua com Amplo Vocabulário para o Português Brasileiro," Dissertação de Mestrado, Junho de 2006.

[17] Cole, R. A., Mariani, J., Uszkoreit, H., Zaenen, A. and Zue, V., "Survey of the State of the Art in Human Language Technology," Cambridge University Press, Cambridge, UK, 1997, (http://cslu.cse.ogi.edu/HLTsurvey).

[18] Huang, X., Acero, A., Hon, H.-W., " Spoken Language Processing, A guide to Theory, Algorithm and System Development," Prentice-Hall, 2001.

[19] Baum, L. E., "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes," Inequalities, vol. 1, pp. 1-8, 1972.

[20] Hwang, M. Y. and Huang, X., "Shared Distribution Hidden Markov Models for Speech Recognition," IEEE Trans Speech and Audio Processing, vol. 1(4), pp414-420, 1993.

[21] Young, S. J. and Woodland, P. C., "State Clustering in HMM-based Continuous Speech Recognition," Computer Speech and Language, vol. 8(4), pp. 369-384, 1994.

[22] Bahl, L. R., Souza, P. V., Gopalakrishnan, P. S., Nahamoo, D. and Picheny, M. A., "Context Dependent Modeling of Phones in Continuous Speech Using Decision Trees," Proc DARPA Speech and Natural Language Processing Workshop, pp. 264-270, Pacific Grove, Calif., Feb, 1991.

[23] Kannan, A., Ostendorf, M. and Rohlicek, J. R., "Maximum Likelihood Clustering of Gaussians for Speech Recognition," IEEE Trans. on Speech and Audio Processing, Vol. 2(3), pp. 453-455, 1994.

[24] Young, S. J., Odell, J. J. and Woodland, P. C., "Tree-Based State Tying for High Accuracy Acoustic Modeling", Proc. Human Language Technology Workshop, pp. 307-312, Plainsboro NJ, Morgan Kaufman Publishers, March, 1994.

[25] "Corpus de Extractus de Textos Eletrônicos Nilc/ Folha de São Paulo (Ceten-Folha)", (http://acdc.linguateca.pt/cetenfolha/), 14 November 2005.

[26] Katz, S. M., "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer", IEEE Trans ASSP, Vol. 35(3), pp. 400-401, 1987.

[27] Ney, H., Essen, U., Kneser, R., "On Structuring Probabilistic Dependences in Stochastic Language Modeling", Computer Speech and Language, Vol. 8(1), pp. 1-38, 1994.

[28] Jelinek, F., “Up from Trigrams: the Struggle for Improved Language Models”, Proc. Eurospeech, pp. 1037-1040, Genoa, 1991.

[29] Bahl, L. R., Brown, P. F., de Souza, P. V., Mercer, R. L., “A Tree-Based Statistical Language Model for Natural Language Speech Recognition”, IEEE Trans ASSP, Vol. 37(7), pp. 507-514, 1989.

[30] Waegner, N. P., Young, S. J., “A Trellis-based Language Model for Speech Recognition”, Proc ICSLP, pp. 245-248, Banff, Canada, October, 1992.

[31] Lau, R., Rosenfeld, R., Roukos, S., “Trigger-based Language Models: a Maximum Entropy Approach”, Proc ICASSP’93, Vol. 2, pp. 45-48, Minneapolis, 1993.

[32] Black, E., Jelinek, F., Lafferty, J., Margeman, D. M., Mercer, R., Roukos, S., “Towards History-based Grammars: Using Richer Models for Probabilistic Parsing”, Proc DARPA, Spoken Language Workshop, pp.31-34, February, 1992.

[33] Deligne, S., Bimbot, F., “Language Modeling by Variable Length Sequences: Theoretical Formulation and Evaluation of Multigrams”, Proc ICASSP, Vol. 1, pp. 169-172, Detroit, 1995.

[34] Forney JR., G. D., "The Viterbi Algorithm," Proc. IEEE, vol. 61, pp. 268-278, March 1973.

[35] Vintsyuk, T. K., “Speech Discrimination by Dynamic Programming”, Kibernetika (Cybernetics), vol. 4 (1), pp. 81-88, Jan.-Feb. 1968.

[36] Lowerre, B. T., “The HARPY Speech Recognition System”, PhD Thesis in Computer Science Department, Carnegie Mellon University, 1976.

[37] Ney, H. J. and Ortmanns, S., “Dynamic Programming Search for Continuous Speech Recognition”, IEEE Signal Processing Magazine, pp. 64-83, 1999.

[38] Schwartz, R., et al., "Context-Dependent Modeling for Acoustic-Phonetic Recognition of Speech Signals," Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing, 1985, Tampa, FLA pp. 1205-1208.

[39] ITU-T Recommendation G.723.1, “Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 kbit/s,” (http://www.itu.int/en/pages/default.aspx), March, 1996.

[40] 3GPP TS 26.071 V6.0.0, “Mandatory speech CODEC speech processing functions, AMR speech CODEC - General description,” (http://www.3gpp.org/), December, 2004.

[41] 3GPP TS 26.171 V6.0.0, “Speech codec speech processing functions, Adaptive Multi-Rate - Wideband (AMR-WB) speech codec - General description,” (http://www.3gpp.org/), December, 2004.

[42] ITU-T Recommendation G.722.2, “Wideband coding of speech at around 16 k/bits using adaptive multi-rate wideband (AMR-WB),” (http://www.itu.int/en/pages/default.aspx), Jully, 2003.

[43] ITU-T Recommendation G.712, “Transmission performance characteristics of pulse code modulation channels,” (http://www.itu.int/en/pages/default.aspx), November, 2001.

[44] ITU-T Recommendation G.723.1, “Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 kbit/s - Annex A: Silence compression scheme,” (http://www.itu.int/en/pages/default.aspx), November, 1996.

[45] ITU-T Recommendation G.723.1, “Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 kbit/s - Annex B: Alternative specification based on floating point arithmetic,” (http://www.itu.int/en/pages/default.aspx), November, 1996.

[46] 3GPP TS 46.001 V8.0.0, “Full rate speech; Processing functions,” (http://www.3gpp.org/), December, 2008.

[47] 3GPP TS 46.002 V8.0.0, “Half rate speech; Half rate speech processing functions,” (http://www.3gpp.org/), December, 2008.

[48] 3GPP TS 46.051 V8.0.0, “Enhanced Full Rate (EFR) speech processing functions; General description,” (http://www.3gpp.org/), December, 2008.

[49] 3GPP TS 26.094 V6.1.0, “Mandatory speech CODEC speech processing functions, Adaptive Multi-Rate (AMR) speech codec - Voice Activity Detector (VAD),” (http://www.3gpp.org/), July, 2006.

[50] 3GPP TS 26.092 V6.0.0, “Mandatory speech CODEC speech processing functions, Adaptive Multi-Rate (AMR) speech codec - Comfort noise aspects,” (http://www.3gpp.org/), December, 2004.

[51] 3GPP TS 26.091 V6.0.0, “Mandatory speech CODEC speech processing functions, Adaptive Multi-Rate (AMR) speech codec - Error concealment of lost frames,” (http://www.3gpp.org/), December, 2004.

[52] Hammer F., Reichl P., Nordstrom T., Kubin G., "Corrupted Speech Data Considered Useful", Acta Acustica, Vol. 90(6), pp. 1052-1060, November/December, 2004.

[53] Kondoz, A. M., “Coding for Low Bit Rate Communication Systems,” Digital Speech, John Wiley & Sons, Ltd. Chichester, UK, 1999.

[54] 3GPP TS 26.090 V6.0.0, “Mandatory speech CODEC speech processing functions, Adaptive Multi-Rate (AMR) speech codec – Transcoding functions,” (http://www.3gpp.org/), December, 2004.

[55] 3GPP TS 26.104 V6.1.0, “ANSI C code for the floating-point Adaptive Multi Rate (AMR) speech codec,” (http://www.3gpp.org/), March, 2004.

[56] 3GPP TS 26.073 V6.0.0, “ANSI C code for the Adaptive Multi Rate (AMR) speech codec,” (http://www.3gpp.org/), December, 2004.

[57] 3GPP TS 26.194 V6.0.0, “Speech codec speech processing functions, Adaptive Multi-Rate - Wideband (AMR-WB) speech codec - Voice Activity Detector (VAD),” (http://www.3gpp.org/), December, 2004.

[58] 3GPP TS 26.192 V6.0.0, “Speech codec speech processing functions, Adaptive Multi-Rate - Wideband (AMR-WB) speech codec - Comfort noise aspects,” (http://www.3gpp.org/), December, 2004.

[59] 3GPP TS 26.191 V6.0.0, “Speech codec speech processing functions, Adaptive Multi-Rate - Wideband (AMR-WB) speech codec - Error

concealment of erroneous or lost frames," (http://www.3gpp.org/), December, 2004.

[60] Bistritz, Y. and Pellerm, S. "Immittance Spectral Pairs (ISP) for speech encoding," IEEE Int. Conf. Acoustics, Speech, Signal Processing, Vol. 2, pp. 9-12, 1993.

[61] 3GPP TS 26.190 V6.1.1, "Speech codec speech processing functions, Adaptive Multi-Rate - Wideband (AMR-WB) speech codec – Transcoding functions," (http://www.3gpp.org/), July, 2005.

[62] 3GPP TS 26.204 V6.0.0, "Speech codec speech processing functions, Adaptive Multi-Rate - Wideband (AMR-WB) speech codec - ANSI-C code," (http://www.3gpp.org/), December, 2004.

[63] 3GPP TS 26.173 V6.0.0, "ANSI-C code for the Adaptive Multi Rate-Wideband (AMR-WB) speech codec," (http://www.3gpp.org/), December, 2004.

[64] Kutwak, A. B. "Análise da codificação LPC para sinais de fala," Projeto Final de Curso, UFRJ, 1999.

[65] Deller, J. R., Proakis, J. G. And Hansen, J. H. "Discrete-time processing of speech signals," MacMillan, 1993.

[66] Rabiner, L. R., and Juang, B. H. "Fundamentals of speech recognition," Prentice Hall, 1993.

[67] Oppenheim, A. V., and Johnson, D. H. "Discrete Representation of signals," Proc. IEEE, vol.60, pp.681- 691, June, 1972.

[68] Mitra, S. K., Digital Signal Processing: A Computer-Based Approach, McGraw-Hill International Editions, 1998.

[69] Wölfel, M., McDonough, J., and Waibel, A. "Minimum Variance Distortionless Response on a Warped Frequency Scale," Eurospeech, pp. 1021–1024, Geneva, 2003.

[70] Kleijn, W. B., and Paliwal, K. K. "Speech Coding and Synthesis", pp. 774 Amsterdam, The Netherlands: Elsevier, 1995.

[71] Itakura, F. "Line Spectrum Representation of Linear Predictive Coefficients of Speech Signals," J. Acoustic Soc. America, Vol. 57, S35(A), 1975.

[72] Alencar, V. F. S. and Alcaim, A. "Transformations of LPC and LSF Parameters to Speech Recognition Features", Proceedings of the ICAPR, pp. 522-528, Bath, UK, August 2005.

[73] Gurgen, F. S., Sagayama, S., and Furui, S. "Line spectrum frequency-based distance measures for speech recognition," ICSLP, pp.521-524, Kobe, Japan, November, 1990.

[74] Stevens, S. S., and J. Volkman, "The relation of pitch of frequency: A revised scale", Am. J. Psychol., vol.53, pp.329-353, 1940.

[75] Alencar, V. F. S., and Alcaim, A. "Features Interpolation Domain for Distributed Speech Recognition and Performance for ITU-T G.723.1 CODEC", Proceedings of the ICSLP, pp.1142-1145, Antwerp, BE, August 2007.

[76] Oetken, G., Parks, T. W., and Schüssler, H. W. “New Results in Design of Digital Interpolators,” IEEE Trans. On Acoustics, Speech and Signal Processing, ASSP-23, pp. 301-309, June 1975.

[77] Young, S. et al., “The HTK Book (for HTK Version 3.4),” (http://htk.eng.cam.ac.uk/), December 2006.

[78] Young, S. “An Application Toolkit for HTK (ATK 1.6),” (http://htk.eng.cam.ac.uk/develop/atk.shtml), June 2007.

[79] Järvinen, K. “Standardisation of the Adaptive Multi-rate Codec,” European Signal Processing Conference (EUSIPCO), pp. 1313-1316, Tampere, Finland, 4–8 Setembro 2000.

[80] Wang, J., Gibson, J., “Parameter interpolation to enhance the frame erasure robustness of CELP coders in packet networks”, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 745-748, 2001.

[81] Bolot, J. C., “Characterizing end-to-end packet delay and loss in the Internet”, Proc. ACM SIGCOMM, pp. 289-298, September, 1993.

[82] Haykin, S., "Neural Networks: A Comprehensive Foundation," ed.2, Prentice Hall, 1999.

Apêndice

Neste apêndice iremos apresentar na seção A.1., os dados técnicos dos equipamentos usados para a gravação da base e na seção A.2., as publicações relacionadas a esta tese doutorado até o momento da sua submissão.

A.1.
Informações Técnicas da Gravação da Base Alcaim - Alencar

As características técnicas dos equipamentos usados para a gravação da base de vozes utilizada nesta tese.

- Microfone: SM 58-LC - Shure
- Filtro anti-puff: Shure
- Pré-amplificador: Mic200 Phantom Power
- Placa de Som: Sound Blaster X-Fi Xtreme Áudio
- Software para gravação e edição: Sony Sound Forge 8.0

A sala de gravação possui tratamento acústico, e os equipamentos de gravação, como computadores, pré-amplificadores, que podem emitir algum tipo de ruído, ficam fora da mesma. A sala permite a permanência apenas da pessoa que está realizando a gravação, minimizando assim também outras fontes de ruído.

Os Locutores desta base têm idades variando de 17 à 65 anos, sendo alguns deles profissionais na realização de locuções, outros treinados para a gravação de livros falados e outros sem nenhum conhecimento teórico ou prático sendo orientados durante a gravação.

As frases utilizadas pelos locutores são todas afirmativas, não existindo frases interrogativas, exclamativas, etc.

A.2.
Publicações Relacionadas à Tese

- Alencar, V. F. S., e Alcaim, A., "LSF and LPC - Derived Features for Large Vocabulary Distributed Continuous Speech Recognition in Brazilian Portuguese," Asimolar, pp. 1237-1241, 2008.
- Alencar, V. F. S., e Alcaim, A., "Digital filter interpolation of decoded LSFs for distributed continuous speech recognition," Electronics Letters, Vol. 44, N. 17, pp. 1039-1040, August, 2008.
- Alencar, V. F. S., e Alcaim, A., "On the Performance of ITU-T G.723.1 and AMR-NB Codecs for Large Vocabulary Distributed Speech Recognition in Brazilian Portuguese" In: 2nd International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2009), Londres, 2009.
- Alencar, V. F. S., e Alcaim, A., "Recuperação de Pacotes Perdidos em Sistemas de Reconhecimento de Voz Distribuído Usando Redes Neurais," In: XXVIII Simpósio Brasileiro de Telecomunicações, Blumenau - SC, 2009.

LSF and LPC - Derived Features for Large Vocabulary Distributed Continuous Speech Recognition in Brazilian Portuguese

V. F. S. Alencar and A. Alcaim
vladimir@cetuc.puc-rio.br, alcaim@cetuc.puc-rio.br
CETUC – PUC/RIO
22453-900, Rio de Janeiro/RJ, Brazil

Abstract - In this paper, we describe several important experiments concerning Large Vocabulary Distributed Continuous Speech Recognition (LVDCSR) systems in Brazilian Portuguese using LSF and LPC - Derived Features. The ITU-T G.723.1 codec is employed and investigated as a case of practical use of this technology. Results are presented for both speaker dependent and independent modes as well as the situations where the same text or different texts were used for training and testing.

I. Introduction

The growth of the Internet and cellular mobile communication networks, along with the increasing interest in more natural Automatic Speech Recognition (ASR) systems, have stimulated the development of Large Vocabulary Distributed Continuous Speech Recognition (LVDCSR) services. Such services perform ASR in a server system, based on the acoustic parameters extracted at the user terminal. This procedure allows that the high complexity and large memory requirements of ASR systems be distributed between the simple/low power client devices and the remote server.

Most speech coders employed in mobile communication systems and IP networks operate at low bit rates and utilize, in general, LPC (Linear Predictive Coding) algorithms based on a speech production model. In this model, an excitation signal is applied to an all-pole filter (characterized by the LPC parameters), that represents the spectral envelope information of the speech signal. Usually, the LPC parameters are transformed to LSF (Line Spectral Frequencies), due to attractive properties of the latter to the quantization and interpolation procedures. It is also known that in LVDCSR systems, extracting speech recognition features from the parameters of a speech coder provides better recognition performance than obtaining the features from the decoded/reconstructed signal [1]. However, the parameters of a speech coder are not the most adequate ones for the remote recognition system. For this reason, different codec parameter transformations have to be considered in order to improve recognition accuracy.

Another important remark is that for satisfactory operation of the ASR system, the recognition features have to be obtained at a high rate (typically 100 Hz). However, speech coders for mobile telephony and IP networks generate their parameters at lower rates (e.g., 50 Hz or 33 Hz). In a recent study on the efficiency of recognition features for distributed speech recognition [2], it was shown that low rates significantly degrade the performance of the recognizer. Hence, it is paramount to interpolate the speech features in order to achieve a recognition performance which is closer to the one obtained when the features are extracted at a high rate. In this paper we have used linear interpolation on the LSF domain [3] to obtain features at 100 Hz from the ITU-T G.723.1 codec [4] in a LVDCSR scenario in order to improve the recognition performance of the system.

In Section II, we present a brief description of the recognition features used in the experiments reported in this paper. Section III describes the database and the experimental procedures. Simulation results are presented and discussed in section IV. Finally, Section V concludes de paper.

II. Recognition Features

The recognition features can be extracted directly from the LPC parameters, without the need to reconstruct the speech signal. In speech decoders, these parameters are obtained in a stage before speech reconstruction. This means that recognition features extracted in this stage are less complex than the ones obtained from the reconstructed speech, since they avoid the need of speech recovery. Moreover, it is important to remark that generating features from the reconstructed speech at the decoder yields worse recognition performance than directly extracting them from the codec parameters. Recognition features that can be obtained from the LPC parameters are the LPCC (LPC Cepstrum) and MLPCC (Mel-Frequency LPCC) [5].

The Line Spectral Frequencies (LSFs) are often used in speech coders due to their high coding efficiency and attractive interpolation properties [6]. Extracting recognition features from the LSFs avoids a speech decoding operation, as well as a conversion of LSF to LPC. A distributed speech recognition system that adopts this strategy becomes computationally more efficient than any other one based either on speech reconstruction or on LPC parameter

transformations. The recognition features which can be obtained from the LSFs are the PCC (Pseudo-Cepstral Coefficients) [7], MPCC (Mel-Frequency PCC) [7], PCEP (Pseudo-Cepstrum) [1] and MPCEP (Mel-Frequency PCEP) [1]. It is worth to mention that these features, which are directly obtained from LSFs, correspond to approximations of the LPCC and MLPCC features obtained from LPC parameters. Note that the use of these approximations avoid the need to recover LPC parameters to obtain the recognition features.

In this paper, we will consider only the MEL scale features (MLPCC, MPCC and MPCEP), since they provide a much better performance than the ones achieved with the linear scale features (LPCC, PCC and PCEP) [2]. The MFCC (Mel-Frequency Cepstral Coefficients) features [8]-[9] will be also obtained from voice reconstructed with the ITU-T G.723.1 codec at the two different operation rates (6.3 kbit/s or 5.3 kbit/s) [4]. The difference between the two rates of ITU-T G723.1 affects the MFCC because only the excitation encoding differ between the two rates. Therefore, this difference affects only the features derived from reconstructed speech – the MFCC.

A. Mel-Frequency LPCC (MLPCC)

The extraction process of the LPCC features from the LPC coefficients is formulated in the z-transform domain, using the complex logarithm of the LPC system transfer function, which is analogous to the cepstrum computation from the discrete Fourier transform of the speech signal [10]. The *i*-th LPCC parameter is given by the following recursive equation

$$c_i = \begin{cases} \ln(G) & i = 0 \\ a_1 & i = 1 \\ a_i + \sum_{j=1}^{i-1} \frac{i-j}{i} c_{i-j} a_j & 1 < i \le p \\ \sum_{j=1}^{p} \frac{i-j}{i} c_{i-j} a_j & i > p \end{cases} \quad (1)$$

where a_i is the *i*-th LPC parameter, *p* is the LPC system order and *G* is the gain factor of the system.

The MLPCC feature is obtained by transforming the real frequency axis of the LPCC to the mel frequency scale. This is performed by a bank of n first-order all-pass filters, where *n* is the number of LPCC features [11]. The filters have their first-order all-pass transfer function $\psi(z)$ [10] given by

$$\psi(z) = \frac{z^{-1} - a^*}{1 - az^{-1}} \quad (2)$$

where *a* is the all-pass filter coefficient and a^* is the complex conjugate of *a*. Each LPCC parameter, c_i, is processed by a different filter.

Since the purpose of each filtering operation is to approximate the mel scale frequency, it is important to analyze the relationship of the transfer function given by (2) and the transformation of the frequency axis. In order to simplify the filter implementation, let *a* be a real number [12]. Now rewrite ψ, as a function of $e^{j\Omega}$, as

$$\psi\left(e^{j\Omega}\right) = e^{-j\theta(\Omega)} \quad (3)$$

where Ω is the real frequency. From (2) and (3), we can derive the mel scale frequency as a function of the real frequency Ω:

$$\theta(\Omega) = \arctan\left[\frac{\left(1-a^2\right)\operatorname{sen}\Omega}{\left(1+a^2\right)\cos\Omega - 2a}\right] \quad (4)$$

Changing the value of *a* it is possible to adjust $\theta(\Omega)$ to the mel scale curve. At an 8 kHz sampling frequency, the value of *a* that best approximates the mel scale curve is 0.3624 [12].

The outputs of the filter bank are the MLPCC features.

B. Mel-Frequency PCC (MPCC)

The PCC is computed directly from the LSFs. However, its derivation is based on the LPCC. Mathematical manipulations and approximations allow it to be expressed in terms of the LSFs [7]. The n-th PCC is given by the equation

$$\hat{c}_n = \frac{1}{2n}\left(1 + (-1)^n\right) + \frac{1}{n}\sum_{i=1}^{p} \cos n w_i \quad (5)$$

where w_i is the i-th LSF parameter.

To obtain the MPCC features from the PCC [7], the LSFs w_i are replaced by w_i^m, wich are defined by the transformation

$$w_i^m = w_i + 2\tan^{-1}\left(\frac{0.45\,sin\,w_i}{1 - 0.45\cos w_i}\right) \quad (6)$$

This expression transforms the frequency axis of a particular set of parameters to the mel scale frequency axis [13]. The MPCC features are expressed by

$$\hat{c}_n^m = \frac{1}{2n}\left(1 + (-1)^n\right) + \frac{1}{n}\sum_{i=1}^{p} \cos n w_i^m \quad (7)$$

where $\hat{c}_n^m$ is the *n*-th MPCC.

C. Mel-Frequency PCEP (MPCEP)

Using the mathematical expression of the PCC features, it is somewhat trivial to obtain the PCEP [1]. They are derived from the PCC by eliminating the $\frac{1}{2n}\left(1+(-1)^n\right)$ term. Note that this term does not depend on the speech signal, i.e., it does not depend on the LSF parameters. The n-th PCEP expression is given by

$$\hat{d}_n = \frac{1}{n}\sum_{i=1}^{p}\cos nw_i \qquad (8)$$

It is fair to expect a good spectral performance of the PCEP because they provide a spectral envelope very similar to the one provided by the Cepstrum, wich is generated from the original speech signal [1]. The PCEP features have the advantage of presenting a computational load even lower than the PCC.

Following the same procedure described for the MPCC, we can express the MPCEP features by

$$\hat{d}_n^m = \frac{1}{n}\sum_{i=1}^{p}\cos nw_i^m \qquad (9)$$

where $\hat{d}_n^m$ is the n-th MPCEP.

III. Description of the Database and Experiments

The new database used in this paper was designed based on [14] that specifies a phonetically balanced set of 1000 sentences for the Brazilian Portuguese. The speech database is composed of 50 male speakers and 50 female speakers, each one repeating once all the 1000 sentences (3528 words). The database was recorded in a studio with 16 kHz sample rate and 16 bits per sample with a bandwith from 50 – 7000 Hz. This database was filtered and down sampled [10] to match the ITU-T G723.1 [4] requirements. Fig. 1 is a graphical representation of this database and will be used to explain the set of experiments carried out in this paper.

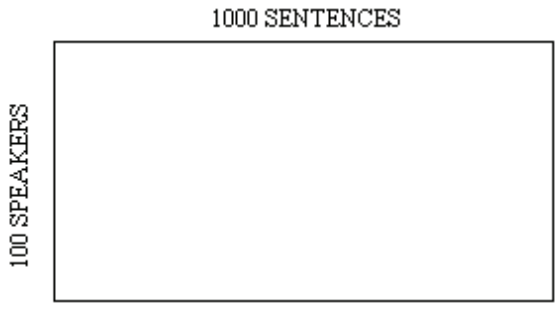

Fig. 1. Graphical representation of the constructed database

In the experiments, we consider LVDCSR systems using the ITU-T G.723.1 codec, as illustrated in Fig. 2. The ITU-T G.723.1 codec is one of the most widely used standards for IP networks nowadays. It allows speech encoding at 6.3 kbit/s or 5.3 kbit/s. In our experiments we have considered both operation modes. The G.723.1 codec employs 30 ms-frames, 8 kHz sampling rate, and 10 LSFs per frame. The LSFs are quantized with a 24 bit predictive split vector quantizer and transmitted at a 33 Hz rate (one every 30 ms). The 100 Hz frame rate was chosen because this is the usual value employed by speech recognizers to provide good performance. Therefore, interpolating the LSFs from 1 per 30 ms to 1 per 10 ms is equivalent to a linear interpolation by a factor of 3. Based on the results presented in [3], we have only considered interpolation in the LSF domain. This means that the features based on LSF (MPCC and MPCEP) or LPC (MLPCC) will be obtained at 100 Hz by the linear interpolation of the LSF parameters from 33Hz to 100 HZ. The MFCC feature is generated from the original and reconstructed speech with 25 ms frame duration (with frames overlay so that parameters are generated at each 10 ms). Hence, no interpolation will be required, since this feature can be directly extracted from the original and decoded speech respectively at the 100 Hz rate.

It should be remarked that in all cases, the model parameters are trained with the same type of features that will be used in tests. This means that we are working in a matched condition.

The database was divided in three different ways to produce three different sets of experiments. The first set of experiments, represented by Fig. 3, can be considered a 100 speaker dependent scenario. The second set of experiments, represented by Fig. 4, can be considered a speaker independent scenario with all the sentences used in the training of the system (same text for training and testing). The third set of experiments, represented by Fig. 5, is speaker and text independent, and characterizes a scenario that best approximates a practical use of the LVDCSR system. A distribution of 75% and 25% of the speech database was used for training and testing in the experiments sets 1 and 2. The experiment set 3 used a distribution of 56.25% for training, 6.25% for testing and 37.5% of the database was not used.

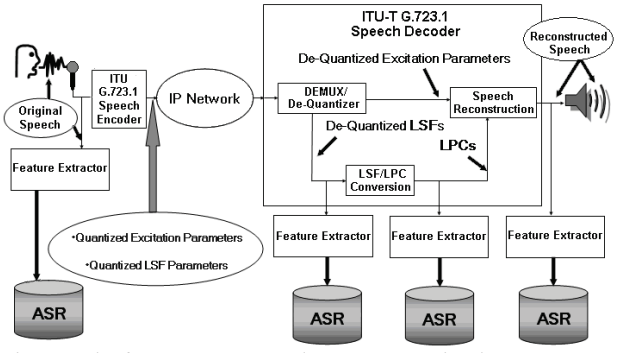

Fig. 2. Codec features extractors and ASR systems using the ITU-T G.723.1

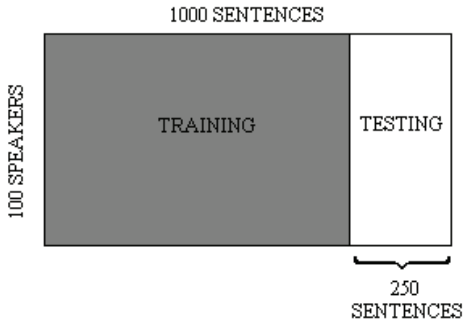

Fig. 3. Graphical representation of database used in Experiment set 1 (speaker dependent)

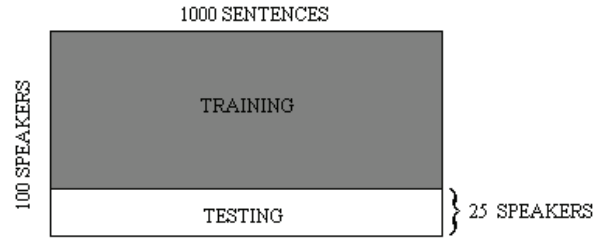

Fig. 4. Graphical representation of database used in Experiment set 2 (speaker independent with same text for training and testing)

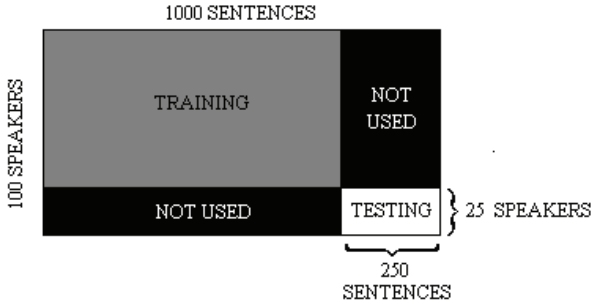

Fig. 5. Graphical representation of database used in Experiment set 3 (speaker and text independent)

To guarantee the statistical confidence of the results, cross-validation was employed in all experiments. Average Word Recognition Rates ($\overline{WRR}$), Standard Deviations (σ) and 95% Confidence Intervals are presented in the next section.

In all experiments in this work, the feature extractors generate one set of 10 parameters plus its first and second derivatives, representing a total of 30 recognition features. The Acoustic Model uses three states continuous observation HMMs (Hidden Markov Models) with a mixture of twenty Gaussians per state for phone modeling. Because silence is stationary, one state was used with the same number of Gaussians. They were implemented with the HTK (HMM Toolkit) software [8]. Inter- and intraword triphones are used as acoustic units. A Trigram language model was trained with the HTK (HMM Toolkit) software [8] with a lexicon of 60,080 words with perplexity of 307 obtained from 240,000 sentences extracted from a large text corpus of Ceten-Folha [15]. The Trigram language model was tested using the ATK (Application Toolkit for HTK) [16].

The system was simulated using a Sun V880 with 4 processors, 8Gb of RAM memory executing Solaris 10 operating system.

IV. Simulations Results

.

Performance results are given in three tables according to the experiments sets described in Section III. Table I shows the recognition results of experiment set 1 (the 100 speaker dependent scenario), Table II presents the recognition results of experiment set 2 (the speaker independent scenario with all the sentences used in the training and testing of the system) and Table III shows the recognition results of experiment set 3 (the speaker and sentences independent scenario). It should be remarked that in each test case, the model parameters are trained with the same type of features (same type of interpolation), i.e., training and testing are matched in this sense.

Comparing the results presented in Table I (speaker dependent scenario) with results presented in Table II (speaker independent scenario with all the sentences used in the training), we can see that the variability of the speaker yields a reduction of around 4% in the $\overline{WRR}$. It can also be seen that the performance drops near 6% from Table II (speaker independent scenario with all the sentences used in training) to Table III (speaker and sentences independent scenario). This shows that besides the 4% performance reduction due to speaker variability, an additional 6% occurs due to text variability. This is a result of different realization of the same triphone (different sentence contexts) during training and testing.

We have also obtained the recognition performance of the MFCC feature extracted from the Reconstructed Speech and Original Speech. Comparing the results of MFCC in theses two situations (Original vs Reconstructed), we can observe that the Reconstructed speech has high performance degradation (around 14%) as compared to the Original Speech, and is worse than the ones obtained with any of the recognition features MPCC, MPCEP and MLPCC in all experiments sets. The MFCC has also sensitivity of around of 2% by how the excitation is encoded and decoded (comparing row two – operation at 5.3 kbits/s – and row three – operation at 6.3 kbits/s – of each table). This shows that the MFCC is very sensitive to the encoding noise. The best results are obtained with the MPCEP recognition feature.

TABLE I
Recognition Accuracy in Experiment Set 1

Features	$\overline{WRR}$	σ	Confidence Interval
MFCC - Original Speech	86.72%	1.02%	[85.52% ; 87.92%]
MFCC - Recons. Speech (5.3 Kbits/s)	72.10%	1.12%	[70.78% ; 73.42%]
MFCC - Recons. Speech (6.3 Kbits/s)	73.83%	1.07%	[72.57% ; 75.09%]
MPCC - Interp. 33Hz to 100 Hz	77.21%	1.01%	[76.02% ; 78.40%]
MPCEP - Interp. 33Hz to 100 Hz	78.11%	0.99%	[76.95% ; 79.28%]
MLPCC - Interp. 33Hz to 100 Hz	77.74%	1.01%	[76.55% ; 78.93%]

TABLE II.
RECOGNITION ACCURACY IN EXPERIMENT SET 2

Features	$\overline{WRR}$	σ	Confidence Interval
MFCC - Original Speech	82.55%	1.37%	[80.94% ; 84.16%]
MFCC - Recons. Speech (5.3 Kbits/s)	68.02%	1.45%	[66.32% ; 69.73%]
MFCC - Recons. Speech (6.3 Kbits/s)	70.05%	1.39%	[68.44% ; 71.69%]
MPCC - Interp. 33Hz to 100 Hz	73.32%	1.37%	[71.71% ; 74.93%]
MPCEP - Interp. 33Hz to 100 Hz	74.27%	1.35%	[72.68% ; 75.86%]
MLPCC - Interp. 33Hz to 100 Hz	73.81%	1.35%	[72.22% ; 75.40%]

TABLE III
RECOGNITION ACCURACY IN EXPERIMENT SET 3

Features	$\overline{WRR}$	σ	Confidence Interval
MFCC - Original Speech	76.82%	1.63%	[76.11% ; 77.54%]
MFCC - Recons. Speech (5.3 Kbits/s)	62.21%	1.74%	[61.45% ; 62.97%]
MFCC - Recons. Speech (6.3 Kbits/s)	63.94%	1.68%	[63.20% ; 64.68%]
MPCC - Interp. 33Hz to 100 Hz	66.31%	1.64%	[65.59% ; 67.03%]
MPCEP - Interp. 33Hz to 100 Hz	67.19%	1.61%	[66.49% ; 67.90%]
MLPCC - Interp. 33Hz to 100 Hz	66.83%	1.67%	[66.10% ; 67.56%]

V. Conclusion

In this paper, we have carried out several important experiments with Large Vocabulary Distributed Continuous Speech Recognition in Brazilian Portuguese. We have shown that only the independency of the speaker deteriorates in around 4% the recognition rate. An additional 6 % performance reduction is due to the use of different sentences during training and testing. We have also shown that the MFCC feature, which is obtained from the reconstructed speech, is highly sensitive to the encoding noise. The performance drops around 14%. We have observed that the MFCC of reconstructed speech has a sensitivity of around of 2% by how the excitation is encoded and decoded (comparing row two – operation at 5.3 kbits/s – and row three – operation at 6.3 kbits/s – of each table). The features obtained from the LSF or LPC parameters can provide much better recognition accuracy than the one obtained from the reconstructed speech (MFCC). The MPCEP feature is the best parameter to be used in an LVDCSR system employing the ITU-T G.723.1 codec. It yields the highest $\overline{WRR}$ with the lowest complexity.

Acknowledgment

The database developed to be used in this work was only possible due a big effort of Salt and Light (a non governmental organization in Brazil, that records books for blind people) and the support given by CNPQ (a governmental organization that supports research in Brazil).

References

[1] H. S. Choi, H. K. Kim, and H. S. Lee, "Speech Recognition Using Quantized LSP Parameters and their Transformations in Digital Communication", vol. 30, pp. 223-233, Speech Communication, 2000.

[2] V. F. S. Alencar and A. Alcaim, "Transformations of LPC and LSF Parameters to Speech Recognition Features", Proceedings of the ICAPR, Bath, UK, August 2005.

[3] V. F. S. Alencar and A. Alcaim, "Features Interpolation Domain for Distributed Speech Recognition and Performance for ITU-T G.723.1 CODEC", Proceedings of the ICSLP, Antwerp, BE, August 2007.

[4] ITU-T Recommendation G.723.1, "Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 Kbit/s," March 1996.

[5] Y. Ohshima, "Environmental Robustness in Speech Recognition using Physiologically-Motivated Signal Processing," PH. D. Thesis, Carnegie Mellon University, Pittsburgh, Pennsylvanya, December 1993.

[6] W. B. Kleijn and K. K. Paliwal, Speech Coding and Synthesis, Amsterdam, The Netherlands: Elsevier, 1995.

[7] H. K. Kim, S. H. Choi and H. S., Lee, "On Approximating Line Spectral Frequencies to LPC Cepstral Coefficients," IEEE Trans. Speech and Audio Processing, vol. 8, pp. 195 – 199, March 2000.

[8] S. Young, et al., The HTK Book (for HTK Version 3.4), (http://htk.eng.cam.ac.uk/), December 2006.

[9] S. B. Davies and P. Mermelstein, "Comparasion of Parametric Representations for Mono syllabic Word Recognition in Continuously Spoken Sentences," vol.28, pp.357-366, IEEE Trans. ASSP, August 1980.

[10] S. K. Mitra, Digital Signal Processing: A Computer-Based Approach, McGraw-Hill International Editions, 1998.

[11] A. V. Oppenheim e D. H., Johnson, "Discrete Representation of Signals," Proc. IEEE, vol. 60, pp.681- 691, June 1972.

[12] M. Wölfel, J. McDonough, e A., Waibel, "Minimum Variance Distortionless Response on a Warped Frequency Scale," Eurospeech, Geneva, 2003.

[13] F. S. Gurgen, S. Sagayama, e S. Furui, "Line Spectrum Frequency-Based Distance Measures for Speech Recognition," pp.521-524, Proc. ICSLP, Kobe, Japan, November 1990.

[14] R. J. R. Cirigliano, et al., "Um Conjunto de 1000 Frases Foneticamente Balanceadas para o Português Brasileiro obtido utilizando a abordagem de Algoritmos Genéticos", XXII Simpósio Brasileiro de Telecomunicações – SBrT, 2005.

[15] "Corpus de Extractus de Textos Eletrônicos Nilc/ Folha de São Paulo (Ceten-Folha)", (http://acdc.linguateca.pt/cetenfolha/), 14 November 2005.

[16] S. Young, An Application Toolkit for HTK (ATK 1.6), (http://htk.eng.cam.ac.uk/develop/atk.shtml), June 2007.

Digital filter interpolation of decoded LSFs for distributed continuous speech recognition

V.F.S. de Alencar and A. Alcaim

A digital filter interpolation of decoded line spectral frequencies (LSFs) that significantly outperforms linear interpolation for large vocabulary distributed continuous speech recognition systems is presented. Experiments were conducted using linear predictive coding (LPC) and LSF-derived speech recognition features, CDHMM acoustic models, triphone units and trigram language models for Brazilian Portuguese.

Introduction: The last few years have witnessed a considerable change in the way speech is carried over digital communication networks. On the other hand, the rapid growth of both mobile and IP networks motivated the integration of automatic speech recognition (ASR) technologies into these networks. Owing to the high complexity of ASR, distributed recognition systems, where acoustic parameters are extracted at the user terminal and recognition is performed at the remote server, are used. In this Letter, we focus on a user terminal where speech is encoded by the ITU-T G.723.1 codec [1]. It is one of the most commonly used codecs for VoIP transmission, owing to its high compression rates (5.3 or 6.3 kbit/s) and the quality of the decoded speech.

The ITU-T G.723.1 utilises linear predictive coding (LPC) algorithms based on a speech production model. In this model, an excitation signal is applied to an all-pole filter (characterised by the LPC parameters), that represents the spectral envelope information of the speech signals. Usually, the LPC parameters are transformed to line spectral frequencies (LSF), owing to the attractive properties of the latter to the quantisation and interpolation procedures. It is also known that extracting speech recognition features from the parameters of a speech coder provides better recognition performance than obtaining the features from the decoded/reconstructed signal [2]. However, the parameters of a speech coder are not the most adequate ones for the remote recognition system. For this reason, different codec parameter transformations have to be considered so as to improve recognition accuracy.

Another important remark is that, for satisfactory operation of the ASR system, the recognition features have to be obtained at a high rate (typically 100 Hz). However, speech coders for mobile telephony and IP networks generate their parameters at lower rates (e.g. 33 Hz). In a recent study on the efficiency of recognition features for distributed speech recognition [3], it was shown that low rates significantly degrade the performance of the recogniser. Usually the linear interpolation in the LSF domain [4] is used to solve this problem. In this Letter we propose a digital filter interpolation technique that results in better speech recognition performance.

Experimental procedures: The new database used in this Letter was designed on the basis of a phonetically balanced set of 1000 sentences for Brazilian Portuguese [5]. The speech database comprises 50 male speakers and 50 female speakers, each one repeating once all the 1000 sentences (3528 words). The database was recorded in a studio with 16 kHz sample rate and 16 bits per sample with a bandwith 50–7000 Hz. This database was filtered and down-sampled to match the ITU-T G723.1 [1] requirements.

The experiments were carried out in a speaker and text independent mode, and characterise a scenario that best approximates a practical use of distributed speech recognition system. We used a distribution of 56.25% of the database for training (75 speakers, each uttering 750 sentences), 6.25% for testing (25 different speakers, each uttering 250 different sentences) and 37.5% of the database was not used. To guarantee the statistical confidence of the results, cross-validation was employed in all experiments. Performance was measured in terms of average word recognition rates ($\overline{WRR}$), standard deviations (σ) and 95% confidence intervals.

The feature extractors generate one set of 10 parameters plus its first and second derivatives, representing a total of 30 recognition features. The acoustic model uses three states continuous observation hidden Markov models (HMMs) with a mixture of 20 Gaussians per state for phone modelling. Because silence is stationary, one state was used with the same number of Gaussians. Inter- and intraword triphones are used as acoustic units. A trigram language model for Brazilian Portuguese was trained with a lexicon of 60 080 words with perplexity of 307 obtained from 240 000 sentences extracted from a large text corpus of Ceten-Folha [6].

In this Letter, we consider the MEL scale features obtained from LPC (the MLPCC [7]) and from LSF (the MPCC [8] and MPCEP [2]), since they provide much better performance than those achieved with the linear scale features (LPCC, PCC and PCEP) [3]. For comparative purposes we have also obtained the mel-frequency cepstral coefficients (MFCC) features from the original speech and from voice reconstructed with the ITU-T G.723.1 codec at the two different operation rates (6.3 or 5.3 kbit/s) [1].

LSFs interpolation: The linear interpolation is a technique usually employed in a variety of applications to obtain a signal at a higher rate. In particular, it has been used in distributed speech recognition systems to interpolate the decoded LSFs from the bitstream of the IS-641 speech coder in [9] and from the bitstream of the ITU-T G.723.1 codec in [4]. In the latter case the LSFs are obtained at the higher rate of 100 Hz from the ITU-T G.723.1 codec which operates at 33 Hz (one set of parameters every 30 ms) [4, 9].

In this Letter, we propose an interpolation technique that is designed using an up-sampler followed by a lowpass digital filter $H(z)$. The up-sampler with factor $r > 1$ (where r is the interpolation factor that in this case is 3), inserts $r - 1$ equidistant zero-valued samples between two consecutive samples. The lowpass filter $H(z)$ eliminates the insertion of images (in this case two images) of the original spectrum compressed by a factor r. It should be noted that in this interpolation application, one important requirement is to ensure that the sequence of input samples (in our case, the decoded LSFs) are not changed at the output. The input signal is assumed to be finite energy and band limited to the frequency range $0 \leq \omega \leq \alpha$, where α must be equal or smaller than 0.5 (in our case, α is assumed to be equal to 0.5). The filter $H(z)$ is a symmetric FIR filter that determines the missing samples by minimising the mean square errors using the orthogonality principle [10]. The length of the FIR filter used in this Letter is $2rL + 1$, where L is an integer that determines the length of the filter (in this Letter L was made equal 4). To reduce the complexity of directly computing the coefficients of this filter of length $2rL + 1 = 25$, we used the techniques proposed in [10] to design optimum interpolators with lower complexity and consuming less hardware. The procedure corresponds to replacing the design of a filter of length 25 by $r = 3$ subfilters of length $2L + 1 = 9$.

Simulation results: Table 1 (rows 4 to 6) shows the recognition results for the interpolation of the recognition features using the ITU-T G723.1 speech codec. It should be noted that in each test case the model parameters are trained with the same type of features and interpolation, i.e. training and testing are matched. The interpolation is carried out in the LSF domain, the MPCEP and MPCC features are directly obtained from the interpolated LSFs. On the other hand, the MLPCC features are obtained from the LPC parameters which, in turn, are generated from the interpolated LSFs.

Table 1: Recognition accuracy for linear and digital filter interpolation techniques

Features	Linear interpolation			Digital filter interpolation		
	$\overline{WRR}$ (%)	σ(%)	Confidence interval (%)	$\overline{WRR}$ (%)	σ(%)	Confidence interval (%)
MFCC–original speech	76.82	1.63	[76.11; 77.54]	76.82	1.63	[76.11; 77.54]
MFCC–reconst. (5.3 kbits/s)	62.21	1.74	[61.45; 62.97]	62.21	1.74	[61.45; 62.97]
MFCC–reconst. (6.3 kbits/s)	63.94	1.68	[63.20; 64.68]	63.94	1.68	[63.20; 64.68]
MPCC–interp. 33–100 Hz	66.31	1.64	[65.59; 67.03]	70.21	1.59	[69.51; 70.91]
MPCEP–interp. 33–100 Hz	67.19	1.61	[66.49; 67.90]	71.32	1.57	[70.63; 72.01]
MLPCC–interp. 33–100 Hz	66.83	1.67	[66.10; 67.56]	70.85	1.60	[70.15; 71.55]

Comparing the results presented in Table 1 we can see that the digital filter interpolation of decoded LSFs can provide remarkable recognition improvement in the distributed systems. It yields an average performance gain of around 4% when compared to the usual linear interpolation procedure. The best word recognition rate was achieved by the MPCEP

feature (71.32% recognition rate), using the digital filter interpolation of the decoded LSFs. From the Table it can also be observed that when the MFCC feature is used, the reconstructed speech presents a high performance degradation compared to the original speech. Moreover, the MFCC of reconstructed speech is worse than that obtained with any of the LSF- and LPC-derived recognition features (MPCC, MPCEP, MLPCC) in all experiments.

Conclusion: Speech coders for mobile telephony and IP networks (e.g. the ITU-T G.723.1) generate their parameters at a rate usually lower than those required for speech recognition. Therefore, in distributed speech recognition it is paramount to interpolate them. In this Letter, we have investigated the use of a digital filter interpolation in a scenario of large vocabulary distributed continuous speech recognition in Brazilian Portuguese. We have shown that, when compared to linear interpolation, the digital filter interpolation of the decoded LSFs remarkably improved the performance of all the recognition features obtained from the bitstream of the ITU-T G.723.1 codec. The average word recognition rate gain was approximately 4% in all situations where recognition features were derived from the decoded LSFs or the LPC parameters.

Acknowledgments: The database developed for use in this work was only possible because of much effort by Salt and Light (a non-governmental organisation in Brazil, that records books for blind people) and the support given by CNPQ (a governmental organisation that supports research in Brazil).

10 June 2008
Electronics Letters online no: 20089650
doi: 10.1049/el:20089650

V.F.S. de Alencar and A. Alcaim (*Centre for Telecommunications Studies of the Catholic University (CETUC) PUC-RIO, Rio de Janeiro 22453-900, Brazil*)

E-mail: vladimir@cetuc.puc-rio.br

References

1 ITU-T Recommendation G.723.1, 'Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 Kbit/s,' March 1996
2 Choi, H.S., Kim, H.K., and Lee, H.S.: 'Speech recognition using quantized LSP parameters and their transformations in digital communication', *Speech Commun.*, 2000, **30**, pp. 223–233
3 Alencar, V.F.S., and Alcaim, A.: 'Transformations of LPC and LSF parameters to speech recognition features'. Proceedings of ICAPR, Bath, UK, August 2005
4 Alencar, V.F.S., and Alcaim, A.: 'Features interpolation domain for distributed speech recognition and performance for ITU-T G.723.1 CODEC'. Proc. of ICSLP, Antwerp, Belgium, August 2007
5 Cirigliano, R.J.R., *et al.*: 'Um Conjunto de 1000 Frases Foneticamente Balanceadas para o Português Brasileiro Obtido Utilizando a Abordagem de Algoritmos Genéticos'. XXII Simpósio Brasileiro de Telecomunicações – SBrT, 2005
6 'Corpus de Extractus de Textos Eletrônicos Nilc/Folha de São Paulo (Ceten-Folha)', (http://acdc.linguateca.pt/cetenfolha/), 14 November 2005
7 Ohshima, Y.: 'Environmental robustness in speech recognition using physiologically-motivated signal processing', Ph.D. Thesis, Carnegie Mellon University, Pittsburgh, Pennsylvania, December 1993
8 Kleijn, W.B., and Paliwal, K.K.: 'Speech coding and synthesis' (Elsevier, Amsterdam, The Netherlands, 1995)
9 Kim, H.K., and Cox, R.V.: 'A bitstream-based front-end for wireless speech recognition on IS-136 communications system', *IEEE Trans. Speech Audio Process.*, 2001, **9**, pp. 558–568
10 Oetken, G., Parks, T.W., and Schüssler, H.W.: 'New results in design of digital interpolators', *IEEE Trans. Accoust. Speech Signal Process.*, 1975, **ASSP-23**, pp. 301–309

On the Performance of ITU-T G.723.1 and AMR-NB Codecs for Large Vocabulary Distributed Speech Recognition in Brazilian Portuguese

Vladimir Fabregas Surigué de Alencar and Abraham Alcaim
CETUC – PUC/RIO – 22453-900, Rio de Janeiro – Brazil
vladimir@cetuc.puc-rio.br, alcaim@cetuc.puc-rio.br

Abstract

In this paper, we present the accuracy for large vocabulary distributed continuous speech recognition systems over ITU-T G.723.1 and AMR-NB speech codecs. Experiments were conducted using LPC and LSF-derived speech recognition features, CDHMM acoustic models, triphone units and trigram language models for the Brazilian Portuguese.

1. Introduction

The last few years have witnessed a considerable change in the way speech is carried over digital communications networks. On the other hand, the rapid growth of both mobile and IP networks motivated the integration of Automatic Speech Recognition (ASR) technologies into these networks. Due to the high complexity of ASR, distributed recognition systems, where acoustic parameters are extracted at the user terminal and recognition is performed at the remote server, are used. In this paper, we have focused on a user terminal where speech is encoded by the ITU-T G.723.1 [1] or AMR-NB [2] codec. The ITU-T G723.1 is one of the most commonly used codecs for VoIP transmission, due to its high compression rates (5.3 or 6.3 kbit/sec) and to the quality of the decoded speech. The AMR-NB is the standard codec for the Global System for Mobile Communications (GSM).

Both ITU-T G.723.1 and AMR-NB utilize LPC (Linear Predictive Coding) algorithms based on a speech production model. In this model, an excitation signal is applied to an all-pole filter (characterized by the LPC parameters), that represents the spectral envelope information of the speech signals. Usually, the LPC parameters are transformed to LSF (Line Spectral Frequencies), due to attractive properties of the latter to the quantization and interpolation procedures. It is also known that extracting speech recognition features from the parameters of a speech coder provides better recognition performance than obtaining the features from the decoded/reconstructed signal [3]. However, the parameters of a speech coder are not the most adequate ones for the remote recognition system. For this reason, different codec parameter transformations have to be considered in order to improve recognition accuracy.

Another important remark is that for satisfactory operation of the ASR system, the recognition features have to be obtained at a high rate (typically 100 Hz). However, speech coders for mobile telephony and IP networks generate their parameters at lower rates (e.g., 33 or 50 Hz). In a recent study on the efficiency of recognition features for distributed speech recognition [4], it was shown that reduction of rate significantly degrades the performance of the recognizer. Usually the linear interpolation in the LSF domain [5] is used to solve this problem. However, in this paper we will use a digital filter interpolation technique that results in a better speech recognition performance [6] when compared with linear interpolation.

This paper is organized as follows. Section 2 and 3 give a brief overview of the recognition features and codecs used in this paper. Section 4 describes the LSFs interpolation procedure, Section 5 the database and experiments and Section 6 the simulation results. Section 7 concludes the paper.

2. Recognition Features

The recognition features can be extracted directly from the LPC parameters, without the need to reconstruct the speech signal. In speech decoders, these parameters are obtained in a stage before speech reconstruction. This means that recognition features extracted in this stage are less complex than the ones obtained from the reconstructed speech, since they avoid the need of speech recovery. Moreover, it is important to remark that generating features from the reconstructed speech at the decoder yields worse recognition performance than directly extracting them from the codec parameters. Recognition features that can be obtained from the

LPC parameters are the LPCC (LPC Cepstrum) and MLPCC (Mel-Frequency LPCC) [7].

The Line Spectral Frequencies (LSFs) are often used in speech coders due to their high coding efficiency and attractive interpolation properties [8]. Extracting recognition features from the LSFs avoids a speech decoding operation, as well as a conversion of LSF to LPC. A distributed speech recognition system that adopts this strategy becomes computationally more efficient than any other one based either on speech reconstruction or on LPC parameter transformations. The recognition features which can be obtained from the LSFs are the PCC (Pseudo-Cepstral Coefficients) [9], MPCC (Mel-Frequency PCC) [9], PCEP (Pseudo-Cepstrum) [3] and MPCEP (Mel-Frequency PCEP) [3]. It is worth to mention that these features, which are directly obtained from LSFs, correspond to approximations of the LPCC and MLPCC features obtained from LPC parameters. Note that the use of these approximations avoid the need to recover LPC parameters to obtain the recognition features.

In this paper, we will consider only the MEL scale features (MLPCC, MPCC and MPCEP), since they provide a much better performance than the ones achieved with the linear scale features (LPCC, PCC and PCEP) [4]. The MFCC (Mel-Frequency Cepstral Coefficients) features [10]-[11] will be also obtained from voice reconstructed with the ITU-T G.723.1 codec and AMR-NB codecs. The difference between the rates of ITU-T G723.1 and AMR-NB affects the performance obtained with the recognition parameters and will be observed from the simulation results.

The detailed equations describing the MLPCC, MPCC and MPCEP features are given in [4].

3. The ITU-T G.723.1 and AMR-NB Codecs

The ITU-T G.723.1 codec operates at the following bit-rates: 5.3 and 6.3 kbits/s. At the 6.3 kbits/s the ITU-T G.723.1 employes a Multipulse Maximum Likelihood Quantization (MP-MLQ) [12]. At the 5.3 kbits/s the ITU-T G.723.1 uses an Algebraic-Code-Excited Linear-Prediction (ACELP) [13]. It transmits two types of information, to be used at the receiver for synthesizing the speech signal: LSF (Line Spectral Frequencies), that represents the frequency response of the synthesis filter, and the excitation signal to the synthesis filter.

The coder is based on the principles of linear prediction analysis-by-synthesis coding and attempts to minimize a perceptually weighted error signal. The encoder operates on blocks (frames) of 240 samples each. That is equal to 30 msec at an 8 kHz sampling rate. Each block is first high pass filtered to remove the DC component and then divided into four subframes of 60 samples each. For every subframe, a 10th order Linear Prediction Coder (LPC) filter is computed using the unprocessed input signal. The LPC filter for the last subframe is converted to LSF and quantized using a Predictive Split Vector Quantizer (PSVQ). The LSF parameters are encoded with 24 bits/frame for both coded rates.

The ITU-T G.723.1 encoder is dedicated to compress the voice signals with bandwidth up to 4 kHz efficiently and to deliver an encoded data stream with a very low binary rate and a good quality of transmitted speech – typical applications being encoding of the vocal signal for video conferences via GSTN (General Switch Telecommunication Network) and Voice over IP.

The AMR-NB codec operates at the following bit-rates: 4.75, 5.15, 5.9, 6.7, 7.4, 7.95, 10.2 and 12.2 kbits/s. The AMR-NB is a ACELP type codec.

The coder operates on speech frames of 20 ms corresponding to 160 samples at the sampling frequency of 8000 sample/s. LP analysis is performed twice per frame for the 12.2 kbits/s mode and once for the other modes. For the 12.2 kbits/s mode, the two sets of LP parameters are converted to LSF and jointly quantized using Split Matrix Quantization (SMQ) with 38 bits/frame. For the other modes, the single set of LP parameters is converted to Linear Spectral Frequencies (LSF) and vector quantized using Split Vector Quantization (SVQ). At the 10.2, 7.4, 6.7 and 5.9 kbits/s the LSFs are quantized with 26 bits/frame. At the operating rate 7.95 the LSF are encoded with 27 bits/frame. At the 5.15 and 4.75 kbits/s the LSFs are quantized with 23 bits/frame.

The different bit-rates of this codec are commonly referred to as modes. The original idea has been that the modes could automatically be altered to ensure the best possible bandwidth sharing between the speech coder and the channel coder. In case of poor channel conditions, to maximize the error protection, the low bit-rate modes are selected, and in the presence of a good channel, the higher bit-rates are used. In this way, the perceptual speech quality can be kept at the highest possible level.

The standardisation in 1999 [14] of AMR-NB as the speech codec of GSM represented a big improvement of voice quality for this mobile network. The AMR-NB codec was also adopted in 1999 by 3GPP as the default speech codec for the WCDMA 3G system. The AMR codec was jointly developed by Ericsson, Nokia and Siemens.

4. LSFs Interpolation

The Linear Interpolation is a technique usually employed in a variety of applications to obtain a signal at a higher rate. In particular, it has been used in distributed speech recognition systems to

interpolate the decoded LSFs from the bitstream of the IS-641 speech coder in [15] and from the bitstream of the ITU-T G.723.1 codec in [5]. In the later case the LSFs are obtained at the higher rate of 100Hz from the ITU-T G.723.1 codec which operates at 33 Hz (one set of parameters every 30 ms) [5] [15].

In [6] a new interpolation technique was proposed that outperforms the Linear Interpolation and for this reason will be the one to be used in this paper. This new interpolation technique is designed using an up-sampler followed by a lowpass digital filter $H(z)$. The up-sampler with factor $r > 1$ (where r is the interpolation factor that in the case of ITU-T G.723.1 is 3 and in the case of AMR-NB is 2), inserts $r-1$ equidistant zero-valued samples between two consecutive samples. The lowpass filter $H(z)$ eliminates the insertion of images (in this case two images) of the original spectrum compressed by a factor r. It should be noted that in this interpolation application, one important requirement is to ensure that the sequence of input samples (in our case, the decoded LSFs) are not changed at the output. The input signal is assumed to be finite energy and band limited to the frequency range $0 \le \omega \le \alpha$, where α must be equal or smaller than 0.5 (in our case, α is assumed to be equal to 0.5). The filter $H(z)$ is a symmetric FIR filter that determines the missing samples by minimizing the mean square errors using the orthogonality principle [16]. The length of the FIR filter used in this letter is $2rL+1$, where L is an integer that determines the length of the filter (in this paper L was made equal 4). In order to reduce the complexity of directly computing the coefficients of this filter of length $2rL+1$ (equal 25 for ITU-T G.723.1 and 17 for AMR-NB), we have used the techniques proposed in [16] to design optimum interpolators with lower complexity and less hardware consuming. The procedure corresponds to replacing the design of a filter of length $2rL+1$ (equal 25 for ITU-T G.723.1 and 17 for AMR-NB) by r (equal 3 for ITU-T G.723.1 and 2 for AMR-NB) subfilters of length $2L+1$ (equal 9 for ITU-T G.723.1 and 9 for AMR-NB).

5. Description of the Database and Experiments

The database used in this paper was designed on the basis of a phonetically balanced set of 1000 sentences for the Brazilian Portuguese [17]. The speech database is composed of 50 male speakers and 50 female speakers, each one repeating once all the 1000 sentences (3528 words). The database was recorded in a studio at a 16 kHz sample rate and 16 bits per sample with a bandwith from 50 – 7000 Hz. This database was filtered and down sampled to match the ITU-T G723.1 [1] and AMR-NB [2] requirements.

The experiments, were carried out in a speaker and text independent mode, and characterize a scenario that best approximates a practical use of distributed speech recognition system. We have used a distribution of 56.25% of the database for training (75 speakers, each uttering 750 sentences), 6.25% for testing (25 different speakers, each uttering 250 different sentences) and 37.5% of the database was not used. To guarantee the statistical confidence of the results, cross-validation was employed in all experiments. Performance was measured in terms of Average Word Recognition Rates ($\overline{WRR}$), Standard Deviations (σ) and 95% Confidence Intervals.

The feature extractors generate one set of 10 parameters plus its first and second derivatives, representing a total of 30 recognition features. The Acoustic Model uses three states continuous observation HMMs (Hidden Markov Models) with a mixture of twenty Gaussians per state for phone modeling. Because silence is stationary, one state was used with the same number of Gaussians. Inter- and intraword triphones are used as acoustic units. A Trigram language model for the Brazilian Portuguese was trained with a lexicon of 60,080 words with perplexity of 307 obtained from 240,000 sentences extracted from a large text corpus of Ceten-Folha [18].

6. Simulation Results

Performance results are given in two tables, Table.1 shows the recognition results for the ITU-T G723.1 and Table.2 shows the recognition results for the AMR-NB. It should be remarked that in each test case, the model parameters are trained with the same type of features (same type of interpolation), i.e., training and testing are matched in this sense.

It is also important to remember that the AMR-NB operating at 12.2 kbits/s generate LSFs at 100 Hz avoiding the need to interpolate de LSF for this mode of the codec (for the other speech codec rates LSF are generate at 50 Hz and need to be interpolated to achieve the 100 Hz).

We have also obtained the recognition performance of the MFCC feature extracted from the Reconstructed Speech and Original Speech. Comparing the results of MFCC in theses two situations (Original vs Reconstructed), we can observe that the Reconstructed speech has high performance degradation (between 11% and 15%) as compared to the Original Speech, and is worse than the ones obtained with any of the recognition features MPCC, MPCEP and MLPCC in all

experiments sets. This shows that the MFCC is very sensitive to the encoding noise. The best results are obtained with the MPCEP recognition feature.

Table 1. Recognition accuracy in ITU-T G.723.1

Features	$\overline{WRR}$	σ	Confidence Interval
MFCC - Original Speech (8kHz,16bits)	76.82%	1.63%	[76.11% ; 77.54%]
6.3 kbits/s			
MFCC - Reconstructed	63.94%	1.68%	[63.20% ; 64.68%]
5.3 kbits/s			
MFCC - Reconstructed	62.21%	1.74%	[61.45% ; 62.97%]
6.3 and 5.3 kbits/s			
MPCC - Interp. 33 Hz to 100 Hz	70.21%	1.59%	[69.51% ; 70.91%]
MPCEP - Interp. 33 Hz to 100 Hz	71.32%	1.57%	[70.63% ; 72.01%]
MLPCC - Interp. 33 Hz to 100 Hz	70.85%	1.60%	[70.15% ; 71.55%]

Table 2. Recognition accuracy in AMR-NB

Features	$\overline{WRR}$	σ	Confidence Interval
MFCC - Original Speech (8kHz,13bits)	76.41%	1.65%	[75.69% ; 77.13%]
12.2 kbits/s			
MFCC - Reconstructed	65.23%	1.67%	[64.50% ; 65.96%]
MPCC - No Interpolation 100 Hz	72.97%	1.55%	[72.29% ; 73.65%]
MPCEP - No Interpolation 100 Hz	74.10%	1.53%	[73.42% ; 74.78%]
MLPCC - No Interpolation 100 Hz	73.62%	1.56%	[72.94% ; 74.30%]
10.2 kbits/s			
MFCC - Reconstructed	65.01%	1.70%	[64.26% ; 65.76%]
7.95 kbits/s			
MFCC - Reconstructed	64.21%	1.70%	[63.46% ; 64.96%]
MPCC - Interp. 50 Hz to 100 Hz	71.75%	1.60%	[71.05% ; 72.45%]
MPCEP - Interp. 50 Hz to 100 Hz	72.89%	1.58%	[72.20% ; 73.58%]
MLPCC - Interp. 50 Hz to 100 Hz	72.41%	1.61%	[71.70% ; 73.12%]
7.4 kbits/s			
MFCC - Reconstructed	63.97%	1.72%	[63.22% ; 64.72%]
6.7 kbits/s			
MFCC - Reconstructed	62.71%	1.72%	[61.96% ; 63.46%]
5.9 kbits/s			
MFCC - Reconstructed	62.33%	1.74%	[61.57% ; 63.09%]
10.2, 7.4, 6.7 and 5.9 kbits/s			
MPCC - Interp. 50 Hz to 100 Hz	71.74%	1.61%	[71.03% ; 72.44%]
MPCEP - Interp. 50 Hz to 100 Hz	72.87%	1.58%	[72.18% ; 73.56%]
MLPCC - Interp. 50 Hz to 100 Hz	72.41%	1.63%	[71.70% ; 73.12%]
5.15 kbits/s			
MFCC - Reconstructed	62.02%	1.79%	[61.24% ; 62.80%]
4.75 kbits/s			
MFCC - Reconstructed	61.94%	1.81%	[61.15% ; 62.73%]
5.15 and 4.75 kbits/s			
MPCC - Interp. 50 Hz to 100 Hz	71.37%	1.65%	[70.64% ; 72.10%]
MPCEP - Interp. 50 Hz to 100 Hz	72.53%	1.61%	[71.82% ; 73.24%]
MLPCC - Interp. 50 Hz to 100 Hz	72.11%	1.68%	[71.37% ; 72.85%]

Now comparing the ITU-T G723.1 at the rates 6.3 and 5.3 kbits/s with the AMR-NB codec at similar rates (6.7 and 5.9 kbits/s) the latter can provide a 1.55% of $\overline{WRR}$ improvement. It is also very significant to note that the 95% confidence interval in these two cases do not overlap.It is also important to remark that the AMR-NB at lower rates (5.15 and 4.75 kbits/s) outperform the ITU-T G.723.1 codec operating at 6.3 and 5.3 kbits/s. Moreover, their 95% confidence interval again do not overlap.

Finally, should be noted that the LSFs in the AMR-NB are encoded at a higher bit rate than the one used by the ITU-T G.723.1.

7. Concluding Remarks

In this paper, we have carried out several important experiments with Large Vocabulary Distributed Continuous Speech Recognition in Brazilian Portuguese. We have also shown that the MFCC feature, which is obtained from the reconstructed speech, is highly sensitive to the encoding noise. The performance drops between 11% and 15%. The features obtained from the LSF or LPC parameters can provide much better recognition accuracy than the one obtained from the reconstructed speech (MFCC). The MPCEP feature is the best parameter to be used in an LVDCSR system employing the ITU-T G.723.1 or AMR-NB codec. It yields the highest $\overline{WRR}$ with the lowest complexity. In addition, the AMR-NB operating at a lower bit rate overperforms the ITU-T G.723.1 codec without overlapping their confidence interval.

8. References

[1] ITU-T Recommendation G.723.1, "Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 Kbit/s," March 1996

[2] 3GPP TS 26.071 V6.0.0, "Mandatory speech CODEC speech processing functions, AMR speech CODEC - General description," December, 2004.

[3] H. S. Choi, H. K. Kim, and H. S. Lee, "Speech Recognition Using Quantized LSP Parameters and their Transformations in Digital Communication", vol. 30, pp. 223-233, Speech Communication, 2000

[4] V. F. S. Alencar and A. Alcaim, "Transformations of LPC and LSF Parameters to Speech Recognition Features", Proceedings of the ICAPR, Bath, UK, August 2005

[5] V. F. S. Alencar and A. Alcaim, "Features Interpolation Domain for Distributed Speech Recognition and Performance for ITU-T G.723.1 CODEC", Proceedings of the ICSLP, Antwerp, BE, August 2007

[6] V. F. S. Alencar and A. Alcaim, "Digital Filter Interpolation of Decoded LSFs for Distributed Continuous Speech Recognition", Electronics Letters, vol.44, issue:17, pp.1039-1040, August 2008.

[7] Y. Ohshima, "Environmental Robustness in Speech Recognition using Physiologically-Motivated Signal Processing," PH. D. Thesis, Carnegie Mellon University, Pittsburgh, Pennsylvanya, December 1993

[8] W. B. Kleijn and K. K. Paliwal, Speech Coding and Synthesis, Amsterdam, The Netherlands: Elsevier, 1995

[9] H. K. Kim, S. H. Choi and H. S., Lee, “On Approximating Line Spectral Frequencies to LPC Cepstral Coefficients,” IEEE Trans. Speech and Audio Processing, vol. 8, pp. 195 – 199, March 2000.

[10] S. Young, et al., The HTK Book (for HTK Version 3.4), (http://htk.eng.cam.ac.uk/), December 2006.

[11] S. B. Davies and P. Mermelstein, “Comparasion of Parametric Representations for Mono syllabic Word Recognition in Continuously Spoken Sentences,” vol.28, pp.357-366, IEEE Trans. ASSP, August 1980.

[12] B.S. Atal, and J.R. Remde, “A new model of LPC excitation for producing natural-sounding speech at low bit rates,” Proceedings of ICASSP, pp. 614–617, 1982.

[13] R. Salami, C. Laflamme, J. P. Adoul, and D. Massaloux, “A toll quality 8 kb/s speech codec for the personal communications system (PCS),” IEEE Trans. Vehicular Technol., vol. 43, pp. 808–816, Aug. 1994.

[14] K. Järvinen, “Standardisation of the Adaptive Multi-rate Codec,” European Signal Processing Conference (EUSIPCO), Tampere, Finland, 4–8 Sept. 2000.

[15] H. K. Kim and R. V. Cox, “A Bitstream-Based Front-End for Wireless Speech Recognition on IS-136 Communications System,” IEEE Trans. On Speech and Audio Processing, vol. 9, pp. 558-568, July 2001.

[16] G. Oetken, T. W. Parks and H. W. Schüssler, “New results in design of digital interpolators,” IEEE Trans. On Accoustics, Speech & Signal Processing, ASSP-23:301-309, June 1975

[17] R. J. R. Cirigliano, et al., “Um Conjunto de 1000 Frases Foneticamente Balanceadas para o Português Brasileiro obtido utilizando a abordagem de Algoritmos Genéticos”, XXII Simpósio Brasileiro de Telecomunicações – SBrT, 2005

[18] “Corpus de Extractus de Textos Eletrônicos Nilc/ Folha de São Paulo (Ceten-Folha)”, (http://acdc.linguateca.pt/cetenfolha/), 14 November 2005

Acknowledgment

The database developed to be used in this work was only possible due a big effort of Salt and Light (a non governmental organization in Brazil, that records books for blind people) and the support given by CNPQ (a governmental organization that supports research in Brazil).

Recuperação de Pacotes Perdidos em Sistemas de Reconhecimento de Voz Distribuído usando Redes Neurais

Vladimir F. S. de Alencar e Abraham Alcaim

***Resumo*—Este artigo propõe uma nova técnica de reconstrução de pacotes perdidos em rajadas em sistemas de reconhecimento de voz distribuído com amplo vocabulário utilizando os codificadores de voz ITU-T G.723.1 e AMR-NB. A nova técnica, que é baseada em Redes Neurais, explora o conhecimento do sinal sem inserir um atraso significante. Experimentos foram conduzidos utilizando o atributo de reconhecimento de voz derivado de LSF (MPCEP), modelos acústicos CDHMM (Continuous Density HMM), unidades trifone e modelos de linguagem trigrama para o Português Brasileiro. Resultados de simulação mostraram que a técnica proposta supera o desempenho de reconhecimento quando comparada com as técnicas de Inserção de Zeros e a Interpolação Linear.**

***Palavras-Chave*— Redes Neurais, Reconhecimento de Voz Distribuído, ITU-T G.723.1, AMR-NB, LSF, LPC, HMM.**

***Abstract*—In this Paper, we propose a novel technique to reconstruct burst-like lost packets in large vocabulary distributed continuous speech recognition systems operating with ITU-T G.723.1 and AMR-NB speech codecs. The new technique, which is based on Neural Networks, takes advantage of the knowledge of the signal without inserting any significant delay. Experiments were conducted using an LSF-derived speech recognition feature (MPCEP), CDHMM (Continuous Density HMM) acoustic models, triphone units and trigram language models for the Brazilian Portuguese. Simulation results show that the proposed technique improves recognition performance as compared to Zero Insertion and Linear Interpolation schemes.**

***Index Terms*— Neural Networks, Distributed Speech Recognition, ITU-T G.723.1, AMR-NB, LSF, LPC, HMM.**

I. INTRODUÇÃO

O desenvolvimento tecnológico do mundo atual tem estimulado a demanda cada vez maior por máquinas inteligentes. Dentro desse panorama, a área de reconhecimento automático de voz (RAV) é uma das que tem 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 interesses cada vez maiores. 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.

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). Por esse motivo, para o desenvolvimento de sistemas voltados a estas redes é necessário conhecer os codificadores de voz utilizados nas mesmas.

Neste artigo, nos baseamos em um terminal usuário onde a voz fosse codificada pelo codec ITU-T G.723.1 [1] ou AMR-NB [2]. O ITU-T G.723.1 é um dos codecs mais amplamente utilizados para a transmissão de voz sobre IP (VoIP), devido a suas taxas elevadas da compressão (5,3 ou 6,3 kbit/s) e à qualidade da voz decodificada. O AMR-NB é o codec padrão para o Sistema Global para as Comunicações Móveis (GSM).

Os Codificadores ITU-T G.723.1 e AMR-NB utilizam os algoritmos LPC (Linear Predictive Coding) baseados em um modelo da produção da voz. Neste modelo, um sinal da excitação é aplicado a um filtro só de pólos (caracterizado pelos parâmetros LPC), o qual representa a informação espectral do envelope do sinal de voz. Geralmente, os parâmetros do LPC são transformados em LSF (Linear Spectral Frequencies), devido às propriedades atrativas do último para os procedimentos de quantização e de interpolação. Sabe-se também que extrair os atributos de reconhecimento dos parâmetros de um codificador de voz fornece um desempenho melhor de reconhecimento do que se obtendo os atributos do sinal decodificado/reconstruído [3]. Entretanto, os parâmetros dos codificadores de voz não são os mais adequados para o sistema de reconhecimento remoto. Por esta razão, diferentes transformações dos parâmetros dos codecs foram consideradas a fim melhorar o desempenho de

Vladimir F. S. de Alencar e Abraham Alcaim, CETUC, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, Brasil, E-mails: vladimir@cetuc.puc-rio.br, alcaim@cetuc.puc-rio.br.

reconhecimento. Neste artigo, nós consideraremos somente o atributo MPCEP (Mel Frequency Pseudo Cepstrum), pois o mesmo demonstrou em [4] que fornece um desempenho melhor, com uma complexidade menor, do que outros atributos de reconhecimento obtidos dos parâmetros do codec.

Uma outra observação importante é que para o funcionamento satisfatório dos sistemas RAV, os atributos de reconhecimento têm que ser obtidos em uma taxa elevada (tipicamente 100 Hz). Entretanto, os codificadores de voz para a telefonia móvel e redes IP geram seus parâmetros em taxas mais baixas (por exemplo, 33 ou 50 Hz). A Interpolação Linear no domínio das LSF [5] é usada geralmente para resolver este problema. Entretanto, neste artigo, nós usaremos uma técnica de interpolação com filtro digital (também no domínio das LSFs) que apresenta um desempenho melhor no reconhecimento da voz [6] quando comparada com a interpolação linear.

O problema de perda de pacotes em rajadas nas redes IP e redes móveis é um dos fatores mais importantes a serem considerados na análise de sistemas de reconhecimento de voz distribuídos. Perdas de pacotes em rajadas causam uma redução drástica do desempenho do reconhecimento de voz. Neste artigo, nós apresentamos uma técnica nova para a reconstrução dos pacotes perdidos baseada em Redes Neurais e comparamos seu desempenho de reconhecimento com os aqueles obtidos com as técnicas de inserção de zeros e interpolação linear.

Na seção II deste artigo nós fornecemos uma revisão breve dos codecs ITU-T G.723.1 e AMR-NB. Na seção III descrevemos o procedimento de Interpolação das LSFs. Na seção IV tratamos das perdas de pacotes em rajadas nas redes IP e Móveis Celulares. Na seção V, apresentamos a reconstrução de pacotes perdidos usando Inserção de Zeros e Interpolação Linear. Na seção VI, nós propomos uma nova técnica baseada em Redes Neurais a fim reconstruir os pacotes perdidos. As condições experimentais são apresentadas na seção VII. Na Seção VIII, analisamos os resultados de simulação. Finalmente, a seção IX apresenta as conclusões.

II. CODECS ITU-T G.723.1 E AMR-NB

O codec ITU-T G.723.1 permite a codificação de voz a taxas de 6,3 kbit/s ou 5,3 kbit/s [1]. A taxa mais elevada fornece uma voz de melhor qualidade, porém a taxa mais baixa também fornece uma boa qualidade de voz. A diferença entre essas taxas resulta do tipo de excitação a ser utilizada e transmitida para o decodificador. Na taxa de 6,3 kbit/s, o codificador utiliza para a excitação o MP-MLQ (Multi-pulse Maximum Likelihood Quantization), enquanto que na taxa de 5,3 kbit/s é empregado o ACELP (Algebraic Code-Excited Linear Prediction). O codificador opera sobre quadros de 240 amostras cada, o que equivale a 30 ms a uma taxa de amostragem de 8 kHz. Os 10 parâmetros LSF são codificados por um Predictive Split Vector Quantizer em 24 bits/quadro para ambas as taxas de codificação.

O codec de AMR-NB opera-se nas seguintes taxas de bits: 4,75, 5,15, 5,9, 6,7, 7,4, 7,95, 10,2 e 12,2 kbit/s. O AMR-NB é um codificador do tipo ACELP [2]. Opera sobre quadros de voz de 20 ms que correspondem a 160 amostras na freqüência de amostragem de 8 kHz. A análise LP é executada duas vezes por quadro para a taxa do codificador de 12,2 kbit/s e uma vez para as outras taxas. Para a taxa de 12,2 kbit/s, os dois conjuntos de parâmetros LP são convertidos para dois conjuntos de 10 LSFs os quais são conjuntamente quantizados usando-se um Split Matrix Quantization (SMQ) com 38 bits/quadro. Para as outras taxas, o único conjunto de parâmetros LP é convertido para 10 LSFs e quantizado com um Split Vector Quantization. Em 10,2, 7,4, 6,7 e 5,9 as LSFs são quantizas com 26 bits/quadro e em 7,95 kbit/s as LSFs são codificadas com 27 bits/quadro. Nas taxas de 5,15 e 4,75 as LSFs são quantizadas com 23 bits/quadro. Note-se que as diferentes taxas de bits deste codec são geralmente chamadas de modos. A padronização do AMR-NB em 1999 [7] como o codec de voz do GSM representou uma melhoria grande da qualidade da voz para as redes móveis. O codec AMR-NB foi adotado também em 1999 por 3GPP como o codec de voz para o sistema de WCDMA 3G. O codec AMR foi desenvolvido conjuntamente pela Ericsson, Nokia e Siemens [7].

III. INTERPOLAÇÃO DAS LSFS

A Interpolação Linear é uma técnica empregada geralmente em sistemas de reconhecimento de voz distribuídos para interpolar as LSFs decodificadas [5], [8]. Em [6] uma nova técnica foi proposta que supera a Interpolação Linear e por esta razão será usada neste artigo. Esta nova técnica de interpolação é projetada usando um up-sampler seguido por um filtro digital passa-baixa $H(z)$. O up-sampler com fator $r > 1$ (onde r é o fator de interpolação, no caso do ITU-T G.723.1 é 3 e no caso de AMR-NB é 2) insere $r-1$ amostras zeradas equidistantes entre duas amostras consecutivas. O filtro digital passa-baixa $H(z)$ elimina a inserção das imagens (neste caso duas imagens) do espectro original comprimido por um fator r [6].

IV. PERDAS DE PACOTES EM RAJADAS

Embora o IP e as redes móveis sejam completamente diferentes, ambos sofrem de perdas de pacotes em rajadas. Em redes móveis as perdas ocorrem em momentos de forte desvanecimento do sinal, enquanto que em redes IP as perdas de pacotes ocorrerem devido aos congestionamentos. Nós adotamos que exatamente um quadro é encapsulado em um pacote.

Para considerar as características de rajadas do processo de perdas de pacotes, o mesmo foi aproximado por um modelo Markoviano de dois-estados, conhecido também como modelo de Gilbert [9]. Os dois estados referem-se aos eventos "pacote recebido" e "pacote perdido", respectivamente, p denota a probabilidade da transição do estado "pacote recebido" para o de "pacote perdido", e q a probabilidade da transição do estado "pacote perdido" para o estado "pacote

recebido". A taxa de perda de pacotes (PLR - packet Lost Rate), sabido também como a probabilidade incondicional de perda (ulp - unconditional loss probability) é dado por: $PLR = p/(p+q)$. O comprimento da rajada (*plg* - packet loss gap) conhecido também como o comprimento médio da rajada (B) é dado por $B = 1/(1-clp)$, onde clp (conditional loss probability) é a probabilidade condicional de perda de pacotes, isto é, a probabilidade da transição do estado "pacote perdido" para "pacote perdido" (isto é. $clp = 1-q$). O modelo de perda de pacotes foi simulado neste artigo com as condições de rede usadas em [9] e apresentados na Tabela I.

TABELA I
SIMULAÇÃO DAS CONDIÇÕES DE REDE USANDO O MODELO DE GILBERT.

PLR(%)	clp	B	p	q
0	-	-	0	0
10	0.15	1.18	0.10	0.85
20	0.30	1.43	0.20	0.70
30	0.35	1.54	0.30	0.65
40	0.50	2.00	0.30	0.50

V. RECONSTRUÇÃO USANDO INSERÇÃO DE ZEROS E INTERPOLAÇÃO LINEAR

Existem algumas aproximações para melhorar o desempenho do sistema de reconhecimento de voz, na presença de imperfeições do canal tais como apagamentos dos quadros. Uma solução simples é inserção dos zeros na posição dos pacotes perdidos. Uma outra aproximação é interpolação linear, entre pacotes recebidos com sucesso (em nosso caso, quadros). O destino recebe, por exemplo, o primeiro conjunto de LSFs quantizadas. Entretanto, devido às imperfeições do canal, não é recebido o segundo conjunto. Na chegada do terceiro conjunto, o receptor pode aproximar o segundo pela interpolação linear do primeiro conjunto com o terceiro. Certamente, a interpolação de mais de um conjunto é praticável em troca de um incremento indesejável de atraso [9]. Para aplicações de redes IP, se n quadros consecutivos de duração t cada um, é perdido, o atraso devido à interpolação é $D_i = nt + RTT/2$, onde RTT (Round-Trip Time) é o tempo para um pacote ir da fonte ao destino e então de volta à fonte. Anote que valores típicos para RTT varia de 10 a 700 ms e de acordo com [9], atrasos aceitáveis para aplicações de VoIP não devem exceder 800 ms [9].

É importante notar que a primeira técnica (Inserção Zero) ignora as características do sinal. Consequentemente, não explora o conhecimento do sinal para melhorar o desempenho do reconhecimento. Por outro lado, o uso da segunda técnica (Interpolação Linear) implica geralmente em longo atraso nos pacotes reconstruídos.

VI. RECONSTRUÇÃO USANDO REDES NEURAIS

Pelas razões expostas na seção anterior, nós propusemos neste artigo, uma nova técnica baseada em Redes Neurais para reconstrução dos pacotes perdidos (com a vantagem de usar o conhecimento do comportamento do sinal) e evitar o retardo significativo para a reconstrução do sinal. O atraso da técnica proposta é somente o tempo das Redes Neurais para computar a saída. Este cálculo está baseado nos quadros de LSFs recebidos antes do pacote perdido ou das LSFs interpoladas obtidas antes do pacote perdido que se deseja recuperar.

Na Figura 1 é apresentada a topologia das Redes Neurais escolhida baseado em resultados de simulações obtidas em uma série de estudos preliminares. A camada escondida é composta de 3 neurônios cuja função selecionada para o neurônio foi a tangente hiperbólica. A função linear foi selecionada para o neurônio da camada da saída. Foram utilizadas 10 Redes Neurais com esta topologia, cada uma para uma das 10 LSFs de cada quadro. As 4 entradas de cada Rede Neural são os valores das LSFs em $T-4$, $T-3$, $T-2$ e $T-1$ onde T é o instante em que um quadro é perdido. A saída é a LSF reconstruída em T. Este valor da LSF será usado no sistema de reconhecimento de voz e como uma entrada da rede neural se a LSF de $T+1$ for perdida também. Cada uma das 10 Redes Neurais são treinadas inicialmente com a mesma base de dados usada no treinamento do HMMs (Hidden Markov Models). É interessante observar que quando são recebidos 5 quadros sucessivamente com sucesso, são usados os primeiros 4 pacotes como entradas das Redes Neurais e o quinto pacote como sua saída. Este procedimento tem como única finalidade re-treinar (re-estimar) as Redes Neurais.

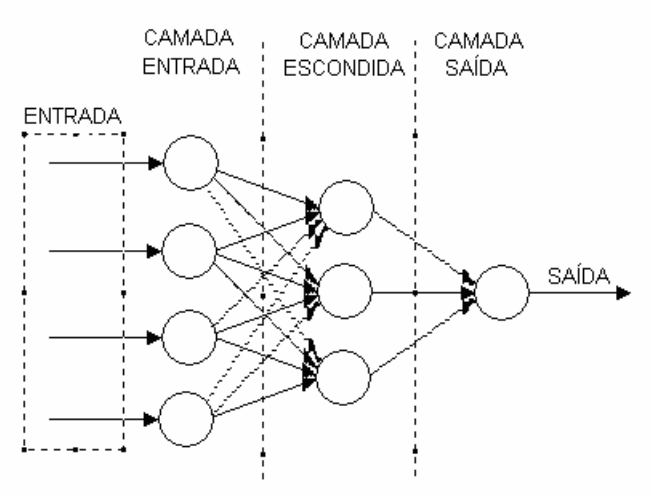

Fig. 1. Topologia das Redes Neurais.

VII. CONDIÇÕES EXPERIMENTAIS

A base de dados usada neste artigo é composta de 50 locutores masculinos e 50 femininos, onde cada locutor fala 1000 sentenças (3.528 palavras) no português Brasileiro. A base de dados foi gravada em um estúdio em uma frequência de amostragem de 16 kHz e em 16 bits por amostra com uma largura de faixa de 50 - 7000 Hertz. Esta base de dados foi filtrada e sub-amostrada para ser compatível com as entradas especificadas pelo ITU-T G.723.1 [1] e pelo AMR-NB [2]. As simulações foram realizadas em um cenário independente do locutor e do texto, e caracterizam o cenário que melhor se aproxima do uso prático do sistema de reconhecimento de voz distribuído. Foi utilizada uma distribuição de 56.25% da base de dados para o treinamento (75 locutores, cada um falando 750 sentenças), 6.25% para testar (25 locutores diferentes, cada um falando 250 sentenças diferentes) e 37.5% da base de dados não foi utilizada. Para garantir a confiança estatística dos resultados, foi empregada a validação cruzada em todas as simulações. O desempenho foi medido nos termos das taxas médias de reconhecimento de palavra ($\overline{WRR}$), desvio padrão (σ) e intervalos de confiança de 95%.

Os extratores de atributos geram um conjunto de 10 atributos mais suas primeiras e segundas derivadas, representando um total de 30 atributos de reconhecimento. Note que os 10 atributos correspondem aos 10 MPCEP convertidos das LSFs quantizadas pelos dois codecs em taxas diferentes. A diferença entre as taxas do ITU-T G.723.1 e AMR-NB afeta significativamente o desempenho obtido com os atributos de reconhecimento, o que será observado nos resultados de simulação. O modelo acústico usa HMMs contínuas de três estados (Hidden Markov Models) com uma mistura de vinte Gaussianas por estado para modelar o fone. Considerando o silêncio estacionário, foi usado um estado com o mesmo número de Gaussians para representá-lo. Os trifones Inter- e Intra-palavra são usados como unidades acústicas. O modelo de linguagem Trigrama para o português Brasileiro foi treinado com um léxico de 60.080 palavras com perplexidade de 307 obtidas de 240.000 sentenças extraídas de um corpus grande de textos do Ceten-Folha [10].

VIII. ANÁLISE DOS RESULTADOS DE SIMULAÇÃO

Os resultados do desempenho são apresentados em cinco tabelas, onde em cada tabela são mostrados o desempenho de reconhecimento para o MPCEP obtido das LSF em diversas taxas dos codificadores ITU-T G.723.1 e AMR-NB para diferentes condições de rede. A tabela II mostra os resultados do reconhecimento para uma rede ideal sem perda dos pacotes. As tabelas III, IV, V e VI mostram os desempenhos de reconhecimento para redes reais com taxas da perda de pacotes PLR e comprimento médio das rajadas B dados por $PLR = 0,\ 10,\ 20,\ 30$ e 40% e $B = 0,\ 1{,}18,\ 1{,}43,\ 1{,}54$ e $2{,}00$, respectivamente. Deve-se observar que em cada caso de teste, os parâmetros do modelo são treinados com o mesmo tipo de atributos (o mesmo tipo de reconstrução), isto é, treinamento e teste estão casados neste sentido. É também importante lembrar que o AMR-NB quando operando em 12,2 kbit/s gera LSFs em 100 Hz, o que evita a necessidade de interpolação das LSFs para esta taxa do codec (para as outras taxas do codec AMR-NB, as LSFs são geradas em 50 Hz e necessitam ser interpoladas para atingir os 100 Hz).

TABELA II
DESEMPENHO DE RECONHECIMENTO PARA REDES SEM PERDAS DE PACOTES.

Atributos	$\overline{WRR}$	σ	Intervalo de Confiança
MPCEP - ITU-T G723,1 (5,3 e 6,3 kbit/s)	71,32%	1,57%	[70,63% ; 72,01%]
MPCEP - AMR-NB (12,2 kbit/s)	74,10%	1,53%	[73,42% ; 74,78%]
MPCEP - AMR-NB (10,2, 7,40, 6,70 e 5,90 kbit/s)	72,87%	1,58%	[72,18% ; 73,56%]
MPCEP - AMR-NB (7,95 kbit/s)	72,89%	1,58%	[72,20% ; 73,58%]
MPCEP - AMR-NB (5,15 e 4,75 kbit/s)	72,53%	1,61%	[71,82% ; 73,24%]

TABELA III
DESEMPENHO DE RECONHECIMENTO PARA REDES COM PLR=10% E B=1,18.

Atributos	$\overline{WRR}$	σ	Intervalo de Confiança
Inserção de Zeros			
MPCEP - ITU-T G723,1 (5,3 e 6,3 kbit/s)	66,21%	1,59%	[65,51% ; 66,91%]
MPCEP - AMR-NB (12,2 kbit/s)	70,01%	1,54%	[69,34% ; 70,69%]
MPCEP - AMR-NB (10,2, 7,40, 6,70 e 5,90 kbit/s)	68,12%	1,60%	[67,42% ; 68,82%]
MPCEP - AMR-NB (7,95 kbit/s)	68,15%	1,59%	[67,45% ; 68,85%]
MPCEP - AMR-NB (5,15 e 4,75 kbit/s)	67,09%	1,64%	[66,37% ; 67,81%]
Interpolação Linear			
MPCEP - ITU-T G723,1 (5,3 e 6,3 kbit/s)	67,52%	1,59%	[66,82% ; 68,22%]
MPCEP - AMR-NB (12,2 kbit/s)	71,45%	1,54%	[70,78% ; 72,13%]
MPCEP - AMR-NB (10,2, 7,40, 6,70 e 5,90 kbit/s)	69,53%	1,59%	[68,83% ; 70,23%]
MPCEP - AMR-NB (7,95 kbit/s)	69,55%	1,59%	[68,85% ; 70,25%]
MPCEP - AMR-NB (5,15 e 4,75 kbit/s)	68,51%	1,63%	[67,80% ; 69,23%]
Redes Neurais			
MPCEP - ITU-T G723,1 (5,3 e 6,3 kbit/s)	67,54%	1,58%	[66,85% ; 68,23%]
MPCEP - AMR-NB (12,2 kbit/s)	71,49%	1,54%	[70,82% ; 72,17%]
MPCEP - AMR-NB (10,2, 7,40, 6,70 e 5,90 kbit/s)	69,54%	1,59%	[68,84% ; 70,24%]
MPCEP - AMR-NB (7,95 kbit/s)	69,58%	1,59%	[68,88% ; 70,28%]
MPCEP - AMR-NB (5,15 e 4,75 kbit/s)	68,52%	1,62%	[67,81% ; 69,23%]

TABELA IV
DESEMPENHO DE RECONHECIMENTO PARA REDES COM PLR=20% E B=1,43.

Atributos	$\overline{WRR}$	σ	Intervalo de Confiança
Inserção de Zeros			
MPCEP - ITU-T G723,1 (5,3 e 6,3 kbit/s)	61,57%	1,64%	[60,85% ; 62,29%]
MPCEP - AMR-NB (12,2 kbit/s)	65,82%	1,58%	[65,13% ; 66,51%]
MPCEP - AMR-NB (10,2, 7,40, 6,70 e 5,90 kbit/s)	63,71%	1,65%	[62,99% ; 64,43%]
MPCEP - AMR-NB (7,95 kbit/s)	63,77%	1,64%	[63,05% ; 64,49%]
MPCEP - AMR-NB (5,15 e 4,75 kbit/s)	62,22%	1,70%	[61,48% ; 62,97%]
Interpolação Linear			
MPCEP - ITU-T G723,1 (5,3 e 6,3 kbit/s)	62,63%	1,63%	[61,92% ; 63,35%]
MPCEP - AMR-NB (12,2 kbit/s)	66,84%	1,58%	[66,15% ; 67,53%]
MPCEP - AMR-NB (10,2, 7,40, 6,70 e 5,90 kbit/s)	64,72%	1,65%	[64,00% ; 65,44%]
MPCEP - AMR-NB (7,95 kbit/s)	64,78%	1,63%	[64,07% ; 65,50%]
MPCEP - AMR-NB (5,15 e 4,75 kbit/s)	63,21%	1,70%	[62,47% ; 63,96%]
Redes Neurais			
MPCEP - ITU-T G723,1 (5,3 e 6,3 kbit/s)	63,12%	1,61%	[62,42% ; 63,83%]
MPCEP - AMR-NB (12,2 kbit/s)	67,37%	1,56%	[66,69% ; 68,06%]
MPCEP - AMR-NB (10,2, 7,40, 6,70 e 5,90 kbit/s)	65,28%	1,64%	[64,56% ; 66,00%]
MPCEP - AMR-NB (7,95 kbit/s)	65,34%	1,61%	[64,64% ; 66,05%]
MPCEP - AMR-NB (5,15 e 4,75 kbit/s)	63,71%	1,69%	[62,97% ; 64,45%]

TABELA V
DESEMPENHO DE RECONHECIMENTO PARA REDES COM PLR=30% E B=1,54.

Atributos	$\overline{WRR}$	σ	Intervalo de Confiança
Inserção de Zeros			
MPCEP - ITU-T G723,1 (5,3 e 6,3 kbit/s)	57,79%	1,68%	[57,06% ; 58,53%]
MPCEP - AMR-NB (12,2 kbit/s)	62,01%	1,63%	[61,30% ; 62,73%]
MPCEP - AMR-NB (10,2, 7,40, 6,70 e 5,90 kbit/s)	59,64%	1,69%	[58,90% ; 60,38%]
MPCEP - AMR-NB (7,95 kbit/s)	59,72%	1,68%	[58,99% ; 60,46%]
MPCEP - AMR-NB (5,15 e 4,75 kbit/s)	58,10%	1,75%	[57,33% ; 58,87%]
Interpolação Linear			
MPCEP - ITU-T G723,1 (5,3 e 6,3 kbit/s)	58,57%	1,68%	[57,84% ; 59,31%]
MPCEP - AMR-NB (12,2 kbit/s)	62,81%	1,63%	[62,10% ; 63,53%]
MPCEP - AMR-NB (10,2, 7,40, 6,70 e 5,90 kbit/s)	60,27%	1,69%	[59,53% ; 61,01%]
MPCEP - AMR-NB (7,95 kbit/s)	60,37%	1,68%	[59,64% ; 61,11%]
MPCEP - AMR-NB (5,15 e 4,75 kbit/s)	58,68%	1,75%	[57,91% ; 59,45%]
Redes Neurais			
MPCEP - ITU-T G723,1 (5,3 e 6,3 kbit/s)	59,93%	1,66%	[59,20% ; 60,66%]
MPCEP - AMR-NB (12,2 kbit/s)	64,49%	1,60%	[63,79% ; 65,19%]
MPCEP - AMR-NB (10,2, 7,40, 6,70 e 5,90 kbit/s)	63,91%	1,67%	[63,18% ; 64,64%]
MPCEP - AMR-NB (7,95 kbit/s)	63,99%	1,66%	[63,26% ; 64,72%]
MPCEP - AMR-NB (5,15 e 4,75 kbit/s)	62,41%	1,73%	[61,65% ; 63,17%]

TABELA VI
DESEMPENHO DE RECONHECIMENTO PARA REDES COM PLR=40% E B=2,00.

Atributos	$\overline{WRR}$	σ	Intervalo de Confiança
Inserção de Zeros			
MPCEP - ITU-T G723,1 (5,3 e 6,3 kbit/s)	49,20%	1,75%	[48,43% ; 49,97%]
MPCEP - AMR-NB (12,2 kbit/s)	56,40%	1,70%	[55,66% ; 57,15%]
MPCEP - AMR-NB (10,2, 7,40, 6,70 e 5,90 kbit/s)	52,99%	1,77%	[52,22% ; 53,77%]
MPCEP - AMR-NB (7,95 kbit/s)	53,13%	1,75%	[52,36% ; 53,90%]
MPCEP - AMR-NB (5,15 e 4,75 kbit/s)	51,06%	1,84%	[50,25% ; 51,87%]
Interpolação Linear			
MPCEP - ITU-T G723,1 (5,3 e 6,3 kbit/s)	49,31%	1,75%	[48,54% ; 50,08%]
MPCEP - AMR-NB (12,2 kbit/s)	56,59%	1,70%	[55,85% ; 57,34%]
MPCEP - AMR-NB (10,2, 7,40, 6,70 e 5,90 kbit/s)	53,27%	1,76%	[52,50% ; 54,04%]
MPCEP - AMR-NB (7,95 kbit/s)	53,37%	1,75%	[52,60% ; 54,14%]
MPCEP - AMR-NB (5,15 e 4,75 kbit/s)	51,32%	1,84%	[50,51% ; 52,13%]
Redes Neurais			
MPCEP - ITU-T G723,1 (5,3 e 6,3 kbit/s)	52,22%	1,71%	[51,47% ; 52,97%]
MPCEP - AMR-NB (12,2 kbit/s)	59,47%	1,65%	[58,75% ; 60,19%]
MPCEP - AMR-NB (10,2, 7,40, 6,70 e 5,90 kbit/s)	56,04%	1,72%	[55,29% ; 56,80%]
MPCEP - AMR-NB (7,95 kbit/s)	56,14%	1,70%	[55,40% ; 56,89%]
MPCEP - AMR-NB (5,15 e 4,75 kbit/s)	54,07%	1,79%	[53,29% ; 54,86%]

Dos resultados da simulação fica claro que a Inserção de Zeros é definitivamente a pior aproximação para a solução da perda de pacotes. Agora comparando a Inserção de Zeros, a Interpolação Linear e a Redes Neurais para a reconstrução de pacotes perdidos de LSFs nas tabelas III, IV, V e VI, pode-se ver que a técnica proposta que usa Redes Neurais supera as duas outras técnicas em todos os casos. Entretanto, as melhorias são somente significativas nas tabelas V e VI, correspondendo à perda de pacotes - PLR - de 30% e 40%, respectivamente, onde as condições das redes IP e das redes móveis são mais severas. No caso onde $PLR = 40\%$ (Tabela VI), o esquema novo fornece ganhos de reconhecimento de aproximadamente 3% quando comparado com a técnica da Interpolação Linear.

Agora comparando o ITU-T G.723.1 nas taxas 6,3 e 5,3 kbit/s com o codec de AMR-NB nas taxas similares (6,7 e 5,9 kbit/s), em toda as condições de rede, o AMR-NB fornece um ganho em torno de 1,50%. É também muito significativo notar que os intervalos de confiança de 95% não se sobrepõem. Além disso, é importante observar que o AMR-NB em suas taxas mais baixas (5,15 e 4,75 kbit/s) supera o codec de ITU-T G.723.1 que opera em 6,3 e 5,3 kbit/s para todos os valores de PLR. Outra vez, seus intervalos de confiança de 95% não se sobrepõem. Note que as LSFs de onde os atributos de reconhecimento são extraídos, são codificados em uma taxa de bits mais elevada pelo AMR-NB em comparação ao ITU-T G.723.1. Finalmente, está claro que as Redes Neurais são uma técnica atrativa para a reconstrução de pacotes perdidos para ambos os codificadores de voz.

IX. CONCLUSÕES

Neste artigo, nós realizamos diversas experiências importantes em Reconhecimento de Voz Contínuo Distribuído com amplo vocabulário no português Brasileiro. Nós propusemos o uso de Redes Neurais para a reconstrução de pacotes perdidos em sistemas Móveis e redes IP. Comparando com a Inserção de Zeros e a técnica de Interpolação Linear, as Redes Neurais mostraram ser o melhor método para reconstruir pacotes perdidos em sistemas de Reconhecimento de Voz Distribuído que empreguem os codecs ITU-T G.723.1 ou AMR-NB, especialmente em condições severas de perda de pacotes. Além disso, nós mostramos que o AMR-NB que opera em uma taxa de bits mais baixa supera o codec ITU-T G.723.1 nas taxas de reconhecimento, sem sobreposição dos seus intervalos de confiança em 95%, em todas as condições da rede.

REFERÊNCIAS

[1] ITU-T Recommendation G.723.1, "Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 kbit/s," Março 1996.

[2] 3GPP TS 26.071 V6.0.0, "Mandatory speech CODEC speech processing functions, AMR speech CODEC - General description," Dezembro, 2004.

[3] H. S. Choi, H. K. Kim, and H. S. Lee, "Speech Recognition Using Quantized LSP Parameters and their Transformations in Digital Communication", vol. 30, pp. 223-233, Speech Communication, 2000.

[4] V. F. S. Alencar and A. Alcaim, "Transformations of LPC and LSF Parameters to Speech Recognition Features", Proceedings of the ICAPR, Bath, UK, Agosto 2005.

[5] V. F. S. Alencar and A. Alcaim, "Features Interpolation Domain for Distributed Speech Recognition and Performance for ITU-T G.723.1 CODEC", Proceedings of the ICSLP, Antwerp, BE, Agosto 2007.

[6] V. F. S. Alencar and A. Alcaim, "Digital Filter Interpolation of Decoded LSFs for Distributed Continuous Speech Recognition", Electronics Letters, vol.44, issue:17, pp.1039-1040, Agosto 2008.

[7] K. Järvinen, "Standardisation of the Adaptive Multi-rate Codec," European Signal Processing Conference (EUSIPCO), Tampere, Finland, 4–8 Setembro 2000.

[8] H. K. Kim and R. V. Cox, "A Bitstream-Based Front-End for Wireless Speech Recognition on IS-136 Communications System," IEEE Trans. On Speech and Audio Processing, vol. 9, pp. 558-568, Jullho 2001.

[9] J. Wang and J. Gibson, "Parameter interpolation to enhance the frame erasure robustness of CELP coders in packet networks", Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, 2001.

[10] "Corpus de Extractus de Textos Eletrônicos Nilc/ Folha de São Paulo (Ceten-Folha)", 14 Novembro 2005.

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo