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.

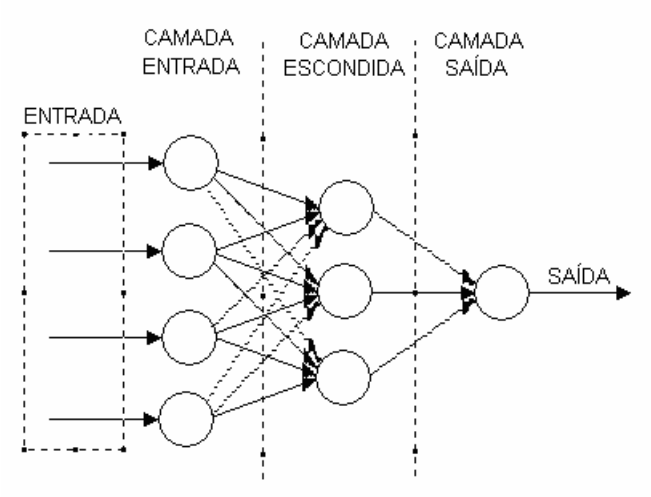

Fig. 1. Topologia das Redes Neurais.

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.