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.

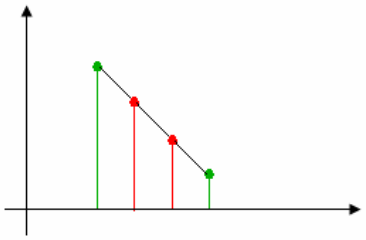

Figura 5.1 – Representação gráfica da interpolação Linear de fator 3

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.

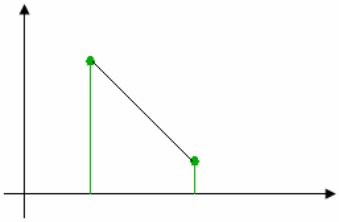

Figura 5.2 – Representação gráfica do Sinal Original

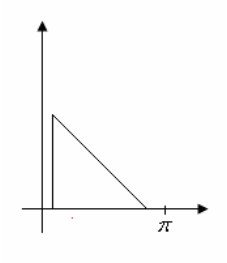

Figura 5.3 – Representação gráfica do Espectro em Freqüência do Sinal Original

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.

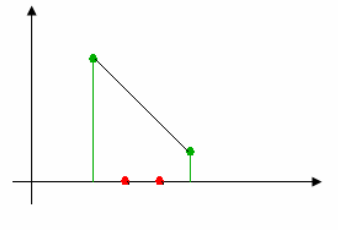

Figura 5.4 – Representação gráfica do Sinal sobre-amostrado de fator 3

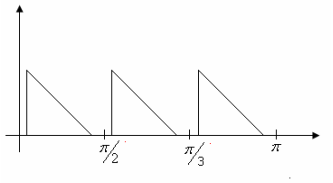

Figura 5.5 – Representação gráfica do Espectro em Freqüência do Sinal sobre-amostrado de fator 3

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.

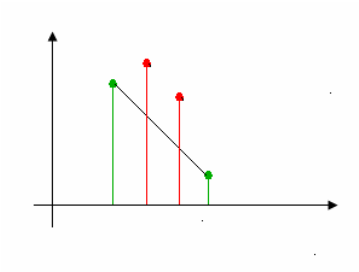

Figura 5.6 – Representação gráfica do Sinal sobre-amostrado de fator 3 depois de filtragem passa-baixa

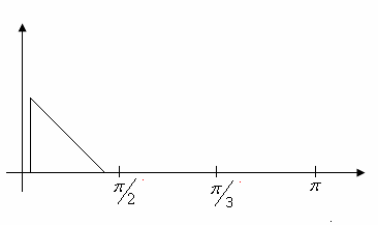

Figura 5.7 – Representação gráfica do Espectro em Freqüência do Sinal sobre-amostrado de fator 3 depois de filtragem passa-baixa

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.