# 6
# Análise dos Atributos de Voz em Reconhecimento Distribuído com a Utilização do *Codec* de Voz ITU-T G.723.1

Neste capítulo serão examinados os sistemas de reconhecimento da Fig. 3.11, com exceção do reconhecedor automático de voz (1) – RAV (1). O RAV (1) não é adequado a sistemas de reconhecimento distribuído, onde só são utilizados parâmetros do codificador no Sistema Remoto para efetuar o reconhecimento.

Será possível, aqui, ao contrário do que ocorreu no Capítulo 5, utilizar o RAV (4), uma vez que o uso de um codificador possibilita a reconstrução da voz no decodificador.

Optou-se por utilizar o codificar ITU-T G.723.1, que é um padrão apenas para redes IP. Esse codificador é do tipo CELP – *Code-Excited Linear Prediction*. Codificadores do tipo CELP são também utilizados em padrões de codificação de redes móveis. Como não serão feitas simulações de perdas de pacotes, e tendo em vista que se está quantizando o mesmo tipo de parâmetro (as LSFs), não se perde em generalidade ao utilizar um padrão de redes IP apenas e não utilizar um *codec* padrão de redes móveis.

O diagrama esquemático do sistema de reconhecimento de voz distribuído para os testes realizados neste Capítulo é mostrado na Fig 6.1.

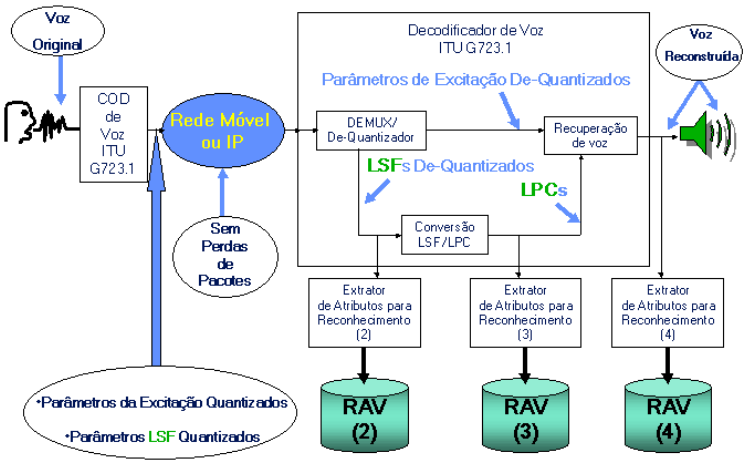

Figura 6.1 – Sistema de reconhecimento de voz distribuído a ser utilizado para o teste com uso do *codec* ITU-T G.723.1

Na Seção 6.1 deste capítulo será feita a apresentação do codec ITU-T G.723.1 e suas principais características para o cenário dos testes de reconhecimento desta dissertação. Na Seção 6.2 são apresentados e discutidos os resultados de reconhecimento com o uso do codificador padrão. Finalmente, a Seção 6.3 contém as principais conclusões dos testes com o *codec* padrão.

## 6.1. Características do *Codec* ITU-T G.723.1

O codec ITU-T G.723.1 permite a codificação de voz a taxas de 6,3 kb/s ou 5,3 kb/s [46]. 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 kb/s, o codificador utiliza para a excitação o MP-MLQ (*Multi-pulse Maximum Likelihood Quantization*), enquanto que na taxa de 5,3 kb/s é empregado o ACELP (*Algebraic Code-Excited Linear Prediction*).

A implementação aqui utilizada [20] não fornece a capacidade de codificar voz na taxa de 5,3 kb/s. Logo, também não implementa o detector de atividade de voz (VAD – *voice activity detector*).

Como o objetivo aqui não é analisar o codificador ITU-T G.723.1 serão apresentadas apenas as características do codificador que impactem direta ou indiretamente sobre o projeto dos extratores de atributos e dos reconhecedores a serem utilizados. 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), seguido de amostragem a taxa de 8 kHz e conversão para um PCM 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. Cada quadro sofre uma filtragem passa-alta a fim de remover a componente DC do sinal 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 serão utilizados apenas para obter a excitação do sistema.

O diagrama esquemático do codificador é apresentado na Fig. 6.2, onde pode-se observar seus blocos básicos, bem como sua complexidade estrutural a qual implica também em um grande consumo de recursos do terminal do usuário.

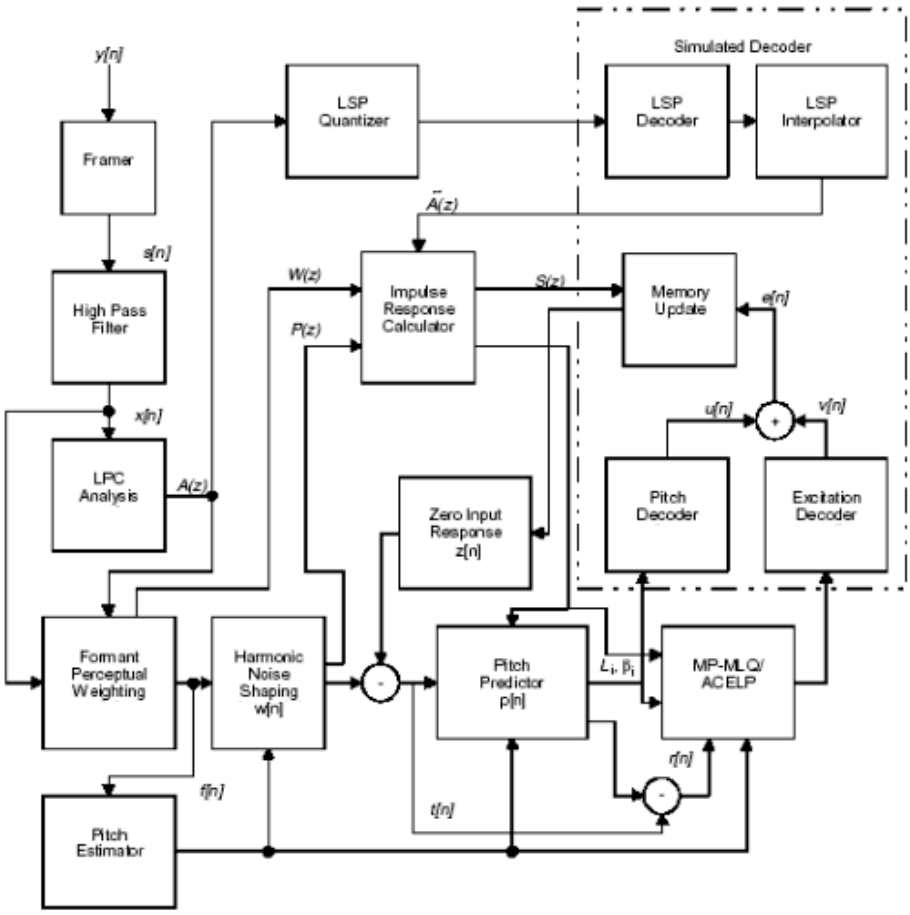


Figura 6.2 – Diagrama de blocos do codificador de voz

Outra informação bastante importante sobre o codificador são as máscaras de alocação de bits utilizadas por ambas as taxas, apresentadas nas Tabs. 6.1 e 6.2, 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.

**Bit allocation of the 6.3 kbit/s coding algorithm**

| Parameters coded | Subframe 0 | Subframe 1 | Subframe 2 | Subframe 3 | Total |
|---|---|---|---|---|---|
| LPC indices | | | | | 24 |
| Adaptive codebook lags | 7 | 2 | 7 | 2 | 18 |
| All the gains combined | 12 | 12 | 12 | 12 | 48 |
| Pulse positions | 20 | 18 | 20 | 18 | 73 (Note) |
| Pulse signs | 6 | 5 | 6 | 5 | 22 |
| Grid index | 1 | 1 | 1 | 1 | 4 |
| Total: | | | | | 189 |
| NOTE – By using the fact that the number of codewords in the fixed codebook is not a power of 2, 3 additional bits are saved by combining the 4 MSB of each pulse position index into a single 13-bit word. | | | | | |

Tabela 6.1 – Tabela de alocação de bits para o codificador operando a 6,3 kb/s

**Bit allocation of the 5.3 kbit/s coding algorithm**

| Parameters coded | Subframe 0 | Subframe 1 | Subframe 2 | Subframe 3 | Total |
|---|---|---|---|---|---|
| LPC indices | | | | | 24 |
| Adaptive codebook lags | 7 | 2 | 7 | 2 | 18 |
| All the gains combined | 12 | 12 | 12 | 12 | 48 |
| Pulse positions | 12 | 12 | 12 | 12 | 48 |
| Pulse signs | 4 | 4 | 4 | 4 | 16 |
| Grid index | 1 | 1 | 1 | 1 | 4 |
| Total: | | | | | 158 |

Tabela 6.2 – Tabela de alocação de bits para o codificador operando a 5,3 kb/s

Vale ressaltar que como a taxa de transmissão de LSF e a precisão com que a mesma é transmitida (número de bits por quadro) são comuns tanto ao codificador funcionando a 5,3 ou 6,3 kb/s, os resultados para os atributos de reconhecimento que dependam apenas das LSFs quantizadas poderão ser considerados como se o codificador implementado pudesse operar em ambas as taxas.

Já no caso da voz reconstruída, deverá haver uma maior degradação do desempenho dos atributos obtidos através da mesma, quando a taxa de 5,3 kb/s estiver disponível no sistema. Entretanto, este fator não é de grande preocupação, pois com a finalidade de reconhecimento, pode-se inibir o funcionamento do codificador na taxa de 5,3 kb/s, a fim de garantir um maior desempenho do sistema de reconhecimento que esteja baseado em voz reconstruída.

A estrutura do decodificador apresentada pela norma ITU-T G.723.1 é aqui ilustrada na Fig. 6.3, tendo sua apresentação justificada pelo fato de demostrar

grande semelhança com a estrutura esquemática do decodificador apresentada na Fig. 6.1.

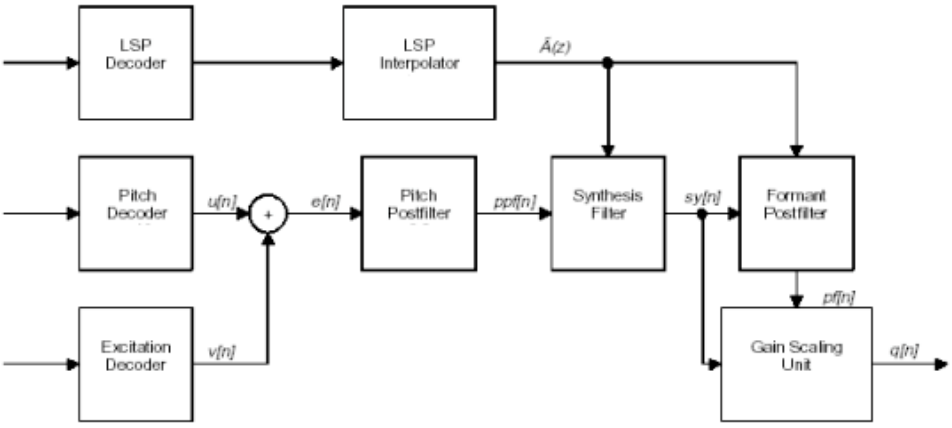


Figura 6.3 – Diagrama de bloco do decodificador de voz

Logo após o bloco do *LSP Decoder* da Fig. 6.3 está o ponto de inserção do extrator de atributos (2), como também após o bloco *LSP Interpolator* da mesma figura se encontra o ponto de inserção do extrator de atributos (3) e depois do bloco *Gain Scaling Unit* obtém-se a voz reconstruída e, assim, o extrator de atributos (4).

## 6.2. Desempenho do Sistema de Reconhecimento com o Uso do *Codec* ITU-T G.723.1

Como foi verificado no Capítulo 5 que a interpolação no domínio das LSFs representa sempre o melhor ganho de reconhecimento, todos os atributos baseados em LSF (MPCC e MPCEP) ou LPC (MLPCC), estarão sendo obtidos na taxa de 10 ms através da interpolação das LSFs de 30 ms para 10 ms, o que implica na interpolação linear de fator 3 apresentada em (3.3).

Como o MFCC é obtido diretamente de voz recontruída, o mesmo não necessita de nenhuma técnica de interpolação, sendo obtido diretamente na taxa de 10 ms.

Logo, no sistema da Fig. 6.3, são obtidos:

- Extrator de atributos (2) – obtém dos parâmetros LSFs os atributos MPCC e MPCEP em 10 ms a partir da interpolação das LSFs;
- Extrator de atributos (3) – obtém dos parâmetros LPC os atributos MLPCC em 10 ms a partir da interpolação das LSFs;
- Extrator de atributos (4) – obtém, a partir de voz reconstruída, os atributos MFCC em 10 ms.

Novamente estará sendo feito o uso de HMMs com 5 estados e 3 gaussianas por estado, treinadas com 70% da base de locuções e testadas com os 30% restantes da base. Esta base de locução foi apresentada no Capítulo 2 e é composta por 50 locutores do sexo masculino e 50 locutores do sexo feminino, onde cada locutor realizou três repetições dos dígitos 0,1,2,3,4,5,6,7,8,9 e a palavra meia, totalizando 3300 locuções. Os resultados obtidos são apresentados na Tab 6.3.

| x | MPCC | MPCEP | MLPCC | MFCC |
|---|---|---|---|---|
| Porcentagem de acertos | 91,8% | 92,3% | 91,9% | 88,1% |

Tabela 6.3 – Resutados dos testes de reconhecimento com o *codec* ITU-T G.723.1

Na Tab. 6.3, pode-se observar que a MFCC obtida de voz reconstruída é a que apresenta o pior desempenho de todos os parâmetros de reconhecimento para este codificador. Cabe ressaltar, porém, que a MFCC não é um parâmetro de reconhecimento robusto ao ruído, onde o ruído presente é o ruído de quantização do codificador. Isto explica o baixo desempenho quando utilizada com voz reconstruída. Outro fato que deve ser destacado é que mesmo treinando um sistema de reconhecimento com a voz já contaminada pelo ruído e efetuando o teste com o mesmo tipo de ruído, este sistema de reconhecimento terá desempenho inferior ao do mesmo treinado e testado com voz sem ruído [17].

A partir da Tab. 6.3 pode-se verificar, também, que mesmo com o uso do codificador, o parâmetro MPCEP, apesar de ser o de mais simples obtenção, foi o que teve melhor desempenho no teste de reconhecimento.

Pode-se concluir, ainda, que se o sistema não tem como objetivo reconstruir voz a partir do sinal transmitido e o único objetivo do receptor é utilizar os parâmetros do codificador para efetuar o reconhecimento da voz, os atributos

MPCEP são os mais adequados por serem os mais leves computacionalmente e por terem o melhor desempenho segundo a Tab. 6.3.

## 6.3. Conclusão

Neste capítulo foram obtidos resultados de reconhecimento com a utilização do codificador de voz padrão ITU-T G.723.1, que na codificação das LSFs tem sua estrutura similar aos métodos utilizados em redes móveis. Assim, como não foram também consideradas peculiaridades de perdas de pacotes de cada tipo de rede, o comportamento dos resultados aqui apresentados deve ser semelhante para outros codificadores.

O fato do resultado de reconhecimento da MFCC de voz reconstruída (Tab. 6.3) ser bastante inferior ao da MFCC de voz original (Tab. 5.3) na mesma taxa de obtenção, mostrou que a codificação prejudica muito o desempenho deste atributo, tornado-o pior que os demais atributos examinados. Isso significa que a MFCC é inadequada para o cenário de reconhecimento de voz distribuída.

Note-se que os parâmetros MPCC e MPCEP representam uma aproximação mais grosseira do MLPCC. Apesar disso, verificou-se que no cenário estudado neste Capítulo, os atributos MLPCC, obtidos a partir de parâmetros LPC, não apresentam melhoria na taxa de reconhecimento em relação aos atributos MPCEP e MPCC, obtidos a partir de LSF. Por outro lado, os atributos MPCEP e MPCC são mais leves computacionalmente para o sistema de reconhecimento que o MLPCC.

Observou-se, também, no cenário estudado neste capítulo, que dentre os atributos obtidos de LSF (MPCEP e MPCC), o MPCEP possui um desempenho 0,5% maior que o MPCC, apesar do mesmo ser uma aproximação mais grosseira do atributo MLPCC quando comparado com o MPCC. Conclui-se, então, que o atributo que apresenta melhor compromisso entre complexidade e desempenho de reconhecimento é o MPCEP.

No próximo Capítulo serão apresentadas as conclusões finais desta dissertação, bem como as sugestões para trabalhos futuros.