3
Reconhecimento de Voz Distribuído

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 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. 3.1.

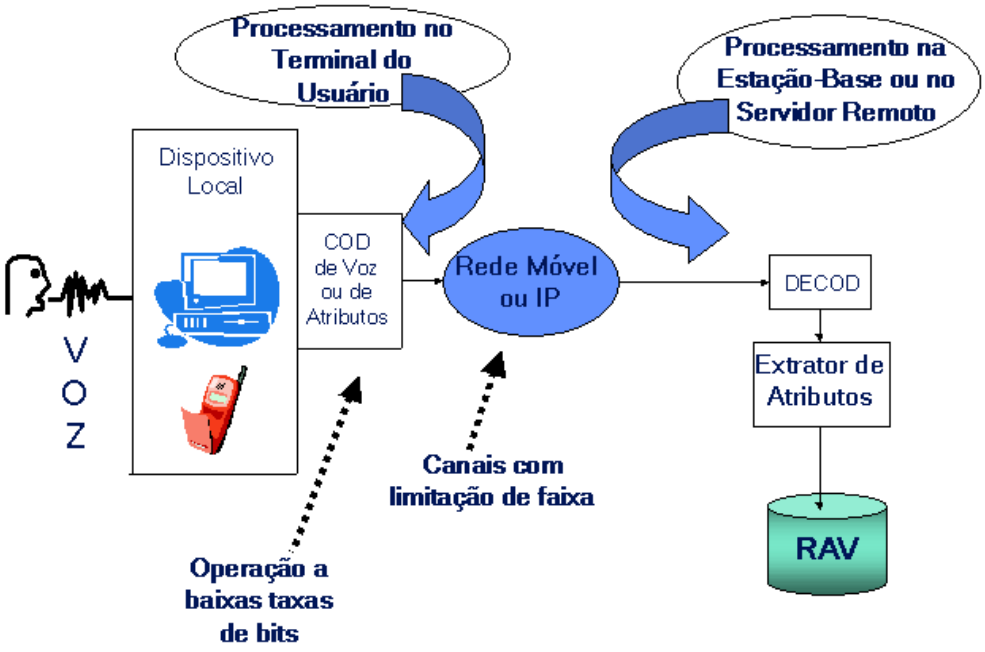

Figura 3.1 Sistemas de Reconhecimento Distribuído – Diagrama Básico

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 [14].

PUC-Rio - Certificação Digital Nº 0310467/CA

3.1.
Atributos mais Utilizados em Reconhecimento de Voz Distribuído

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.

1. Reconhecimento utilizando os parâmetros do codificador de voz

Um esquema deste tipo é ilustrado na Fig. 3.2, onde pode ser observado a sua adequação às situações em que se queira realizar o reconhecimento e a recuperação da voz do locutor.

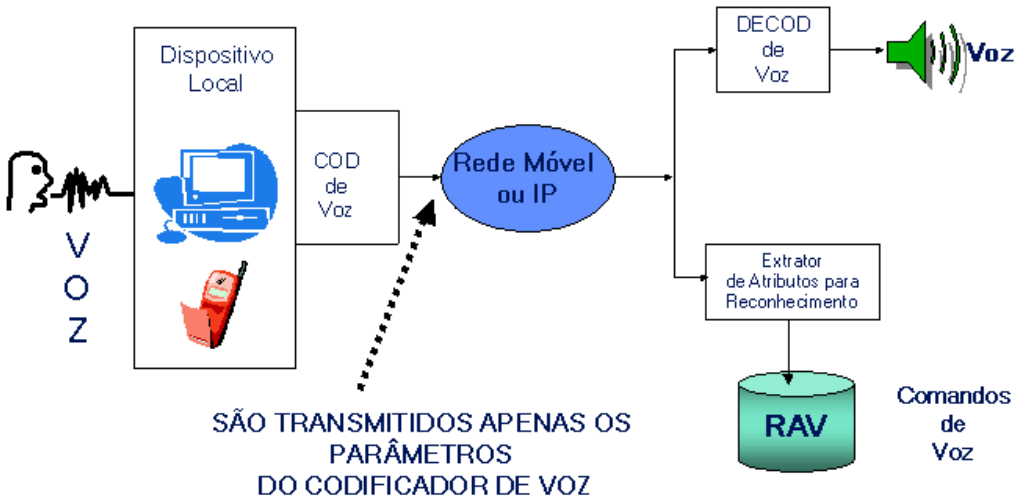

Figura 3.2 – Sistema de reconhecimento de voz distribuído baseado nos parâmetros de voz do codificador

Neste sistema existe uma ampla variedade de atributos de reconhecimento que podem ser obtidos a partir dos parâmetros do codificador de voz ou, até mesmo, do processo de recuperação da voz pelo decodificador, o que simplifica bastante o extrator de atributos, como será visto mais a diante.

PUC-Rio - Certificação Digital Nº 0310467/CA

Para que se descreva os atributos que podem ser obtidos dos parâmetros do codificador de voz, é necessário detalhar um pouco mais a Fig 3.2, apresentando os parâmetros que o codificador transmite e quais serão utilizados para a obtenção dos atributos de reconhecimento. Isso é ilustrado na Fig 3.3.

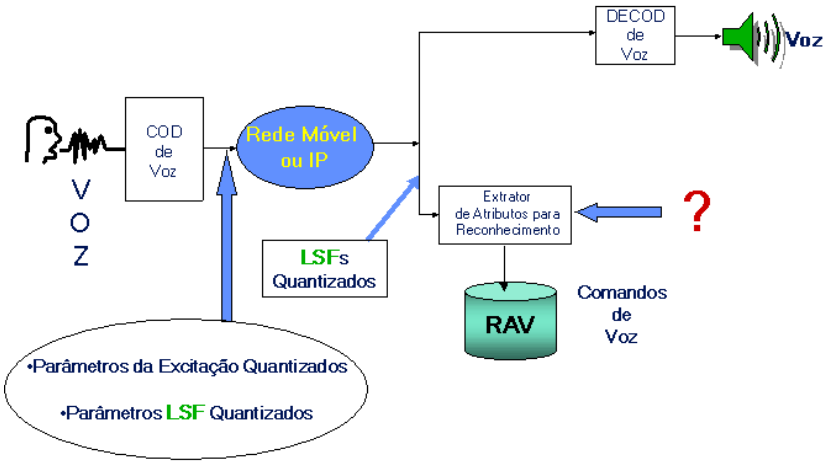

Figura 3.3 – Parâmetros do codificador de voz

O codificador de voz transmite as LSFs e os parâmetros da excitação quantizados, como apresentado na Fig 3.3. Estes parâmetros da voz trafegam pela rede e chegam ao receptor, onde deseja-se realizar o reconhecimento.

Dos parâmetros LSF quantizados o decodificador de voz do receptor obtém os parâmetros LPC de onde pode-se extrair 2 atributos de reconhecimento: LPCC (*LPC Cepstral Coefficients*) e MLPCC (*Mel LPC Cepstral Coefficients*), o que é representado esquematicamente na Fig. 3.4.

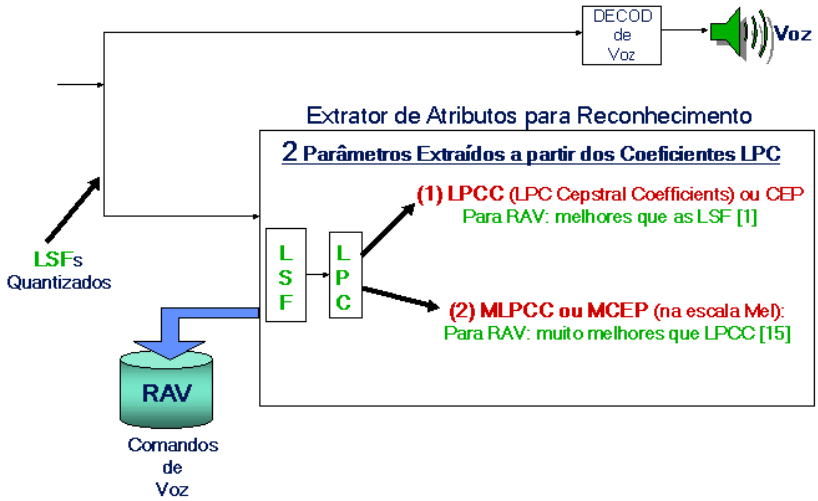

Figura 3.4 – Atributos extraídos dos coeficientes LPC

PUC-Rio - Certificação Digital Nº 0310467/CA

Diretamente dos parâmetros LSFs quantizados também pode-se obter atributos de reconhecimento, sendo eles em número de quatro: PCC (*Pseudo-Cepstral Coefficients*), MPCC (*Mel Pseudo-Cepstral Coefficients*), PCEP (*Pseudo-Cepstrum*) e MPCEP (*Mel Pseudo-Cepstrum*), o que é representado esquematicamente nas Fig 3.5 e 3.6.

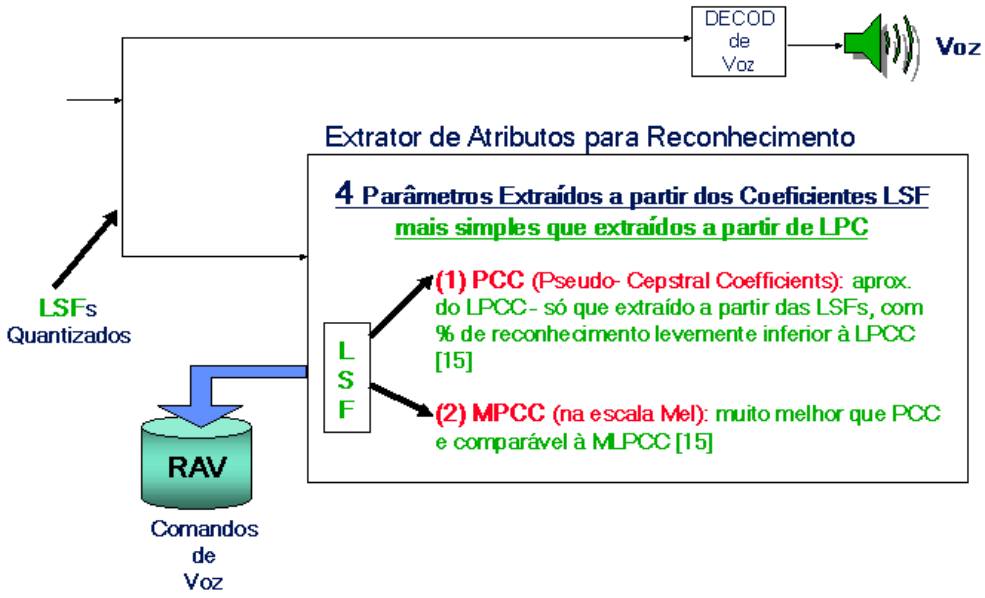

Figura 3.5 – Atributos PCC e MPCC obtidos de parâmetros LSF quantizados

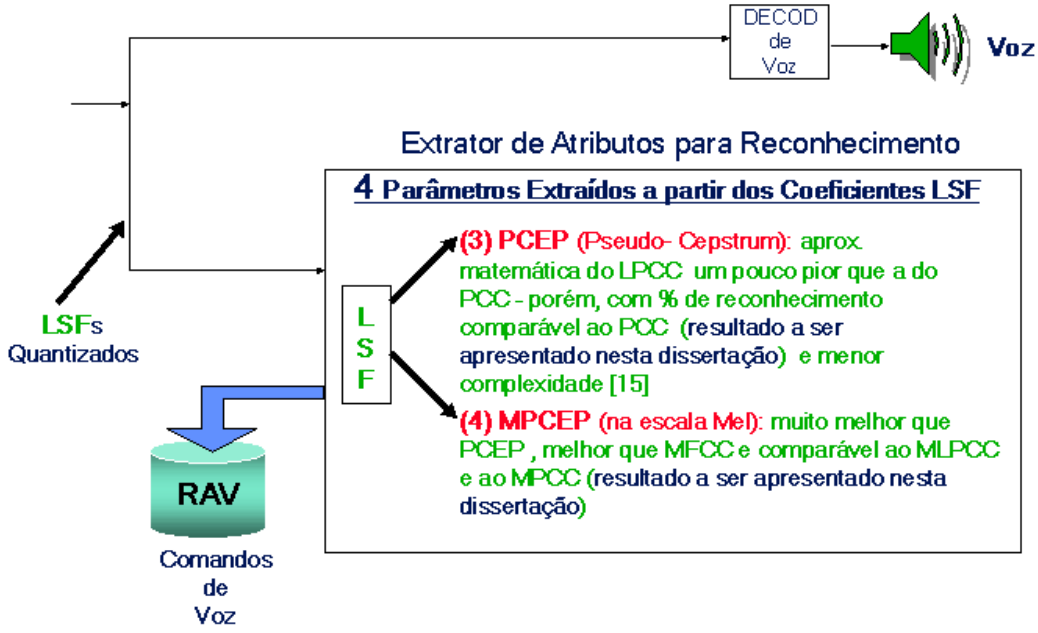

Figura 3.6 – Atributos PCEP e MPCEP obtidos de parâmetros LSF quantizados

PUC-Rio - Certificação Digital Nº 0310467/CA

2. Reconhecimento de voz a partir de voz decodificada

Uma ilustração deste sistema é apresentada na Fig. 3.7, 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 [16].

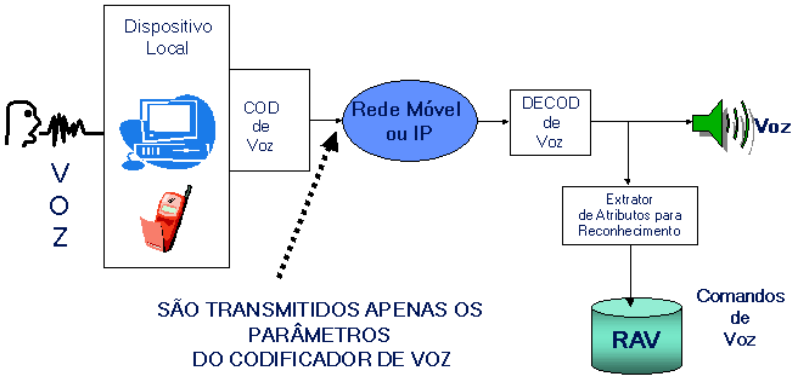

Figura 3.7 – Sistema de reconhecimento de voz distribuído baseado em voz decodificada

Da voz reconstruída podem ser obtidos vários atributos de reconhecimento, dentre os quais: CC (*Cepstral Coefficients*), MFCC (*Mel-Frequency Cepstral Coefficients*), PLP-Cepstrum (*Perceptual Linear Predictive-Cepstrum*) e ZCPA (*Zero Crossings with Peak Amplitudes*), o que é apresentado esquematicamente na Fig 3.8.

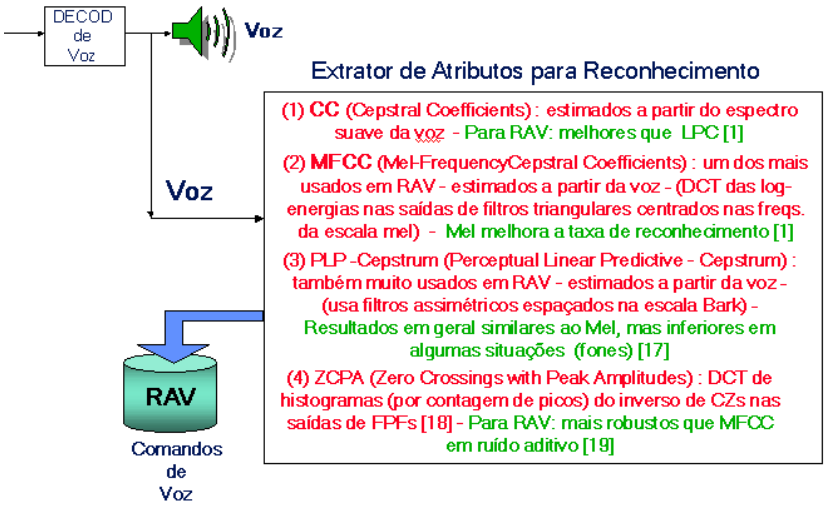

Figura 3.8 – Atributos obtidos de voz reconstruída

PUC-Rio - Certificação Digital Nº 0310467/CA

3. Reconhecimento de voz a partir da codificação dos atributos para reconhecimento

Uma ilustração deste sistema é apresentada na Fig. 3.9, onde se pode observar que o mesmo é bastante adequado para situações onde deseja-se realizar apenas o reconhecimento, devido à impossibilidade de recuperar voz.

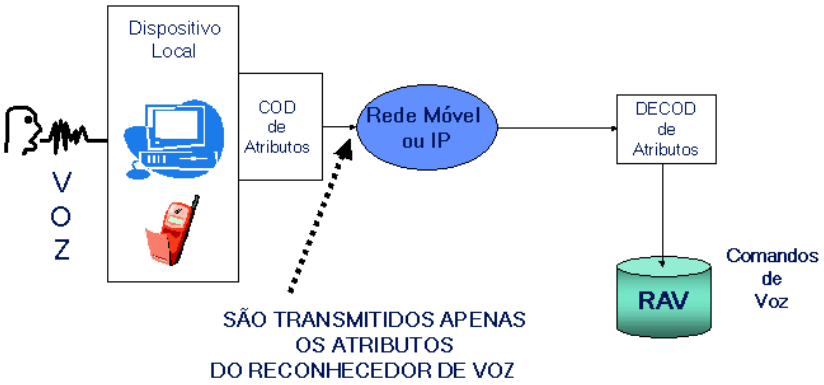

Figura 3.9 – Sistema de reconhecimento de voz distribuído com codificação dos atributos de reconhecimento no *front-end* local

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. Um exemplo deste sistema é mostrado na Fig. 3.10.

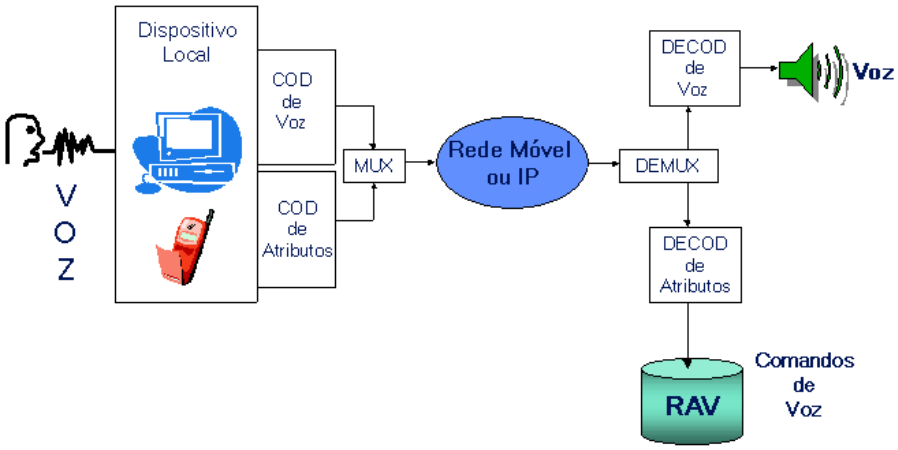

Figura 3.10 - Sistema de reconhecimento de voz distribuído com codificação dos atributos de reconhecimento e codificação de voz no *front-end* local

PUC-Rio - Certificação Digital Nº 0310467/CA

Os atributos de reconhecimento gerados pelo codificador de atributos do *front-end* local são os mesmos obtidos de voz recontruída 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 [16].

3.2. Estrutura do Sistema de Reconhecimento Distribuído Objeto da Dissertação

Tendo sido apresentados os blocos básicos necessários a um sistema de reconhecimento, bem como as abordagens mais utilizadas para implementação de um sistema de reconhecimento distribuído, faz-se necessário apresentar as características específicas do sistema a ser trabalhado nesta dissertação. Esse sistema é mostrado na Fig. 3.11 e será a base de todas as considerações a serem tratadas nesta Seção e nos Capítulos 5 e 6.

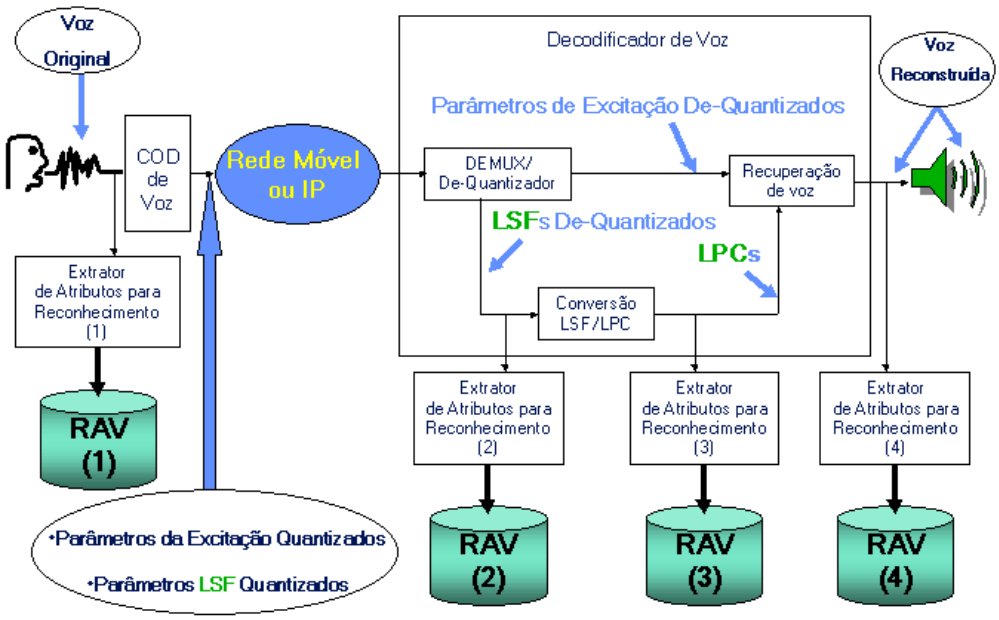

Figura 3.11 – Sistema de reconhecimento distribuído objeto da dissertação

Como pode ser observado na Fig. 3.11, existem vários pontos do sistema de voz sobre Redes IP/Redes de Telefonia Celular onde estarão sendo feitas as

PUC-Rio - Certificação Digital Nº 0310467/CA

obtenções de atributos de reconhecimento e testados os seus respectivos desempenhos.

Os extratores de atributos são os blocos responsáveis pela obtenção dos atributos de reconhecimento, podendo ser resumidos por:

- **Extrator de atributos (1) –** tem como base a voz original para a obtenção dos atributos de reconhecimento. A MFCC de voz original será o único atributo obtido por este extrator, e terá a finalidade de servir como referência para os resultados a serem obtidos com os outros atributos, pois é o que apresenta melhor desempenho [15], o qual se estará tentando atingir. A MFCC será utilizada pelo RAV (1) – Reconhecimento automático de voz (1).
- **Extrator de atributos (2)** – tem como base as LSFs de-quantizadas obtidas pelo processo de de-quantização das LSFs recebidas pelo decodificador. Os atributos PCC, PCEP, MPCC e MPCEP, são obtidos a partir das LSFs de-quantizadas e são utilizados no reconhecimento de voz do RAV (2) – Reconhecimento automático de voz (2).
- **Extrator de atributos (3)** – tem como base os parâmetros LPC, que são obtidos da conversão LSF / LPC a partir das LSFs de-quantizadas. Os atributos LPCC e MLPCC, são obtidos a partir dos parâmetros LPC e são utilizados para o reconhecimento de voz do RAV (3) – Reconhecimento automático de voz (3).
- **Extrator de atributos (4)** – tem como base a voz reconstruída a partir dos parâmetros de excitação de-quantizados e dos parâmetros LPC. A MFCC de voz reconstruída será o único atributo obtido por este extrator de atributos, e terá a finalidade de servir como referência para os resultados a serem obtidos com os outros atributos. Será o atributo utilizado pelo RAV (4) – Reconhecimento automático de voz (4).

Para o sistema de reconhecimento será usado o *HTK Toolkit*, que é uma ferramenta amplamente utilizada para implementar sistemas de reconhecimento automático de voz (RAV) baseados em HMM.

PUC-Rio - Certificação Digital Nº 0310467/CA

O *HTK* é composto por quatro blocos, que são:

- Preparação dos Dados – Bloco do *HTK* que permite a obtenção de alguns atributos de reconhecimento. Este bloco não será utilizado nesta dissertação, pois os atributos de reconhecimento foram implementados em MATLAB®, levando em conta as peculiaridades do sistema que se quer implementar, o qual será detalhado mais adiante.
- Treino – Bloco do *HTK* que permite criar as HMMs referentes a cada palavra e treiná-las com as locuções de treinamento.
- Teste – Bloco do *HTK* que permite testar a capacidade de reconhecimento das HMMs, treinadas no bloco de Treino do HTK, com as locuções de teste.
- Análise dos resultados – Bloco do *HTK* que obtém as estatísticas de desempenho do reconhecimento das locuções de teste, como percentual de acertos e número absoluto de acertos.

Como já citado anteriormente, o MATLAB® será utilizado para implementar os diversos extratores de atributos necessários à implementação do sistema de reconhecimento. Além disso, o MATLAB® será utilizado para permitir o funcionamento dos blocos de codificação e decodificação de voz do padrão ITU-T G.723.1, que foram implementados nesta ferramenta por Peter Kabal [20].

Os blocos de extratores de atributos serão detalhados nos Capítulos 5 e 6, onde são utilizados, e os resultados de reconhecimento para cada um deles também será avaliado. Antes, porém, faz-se necessário apresentar um método que será amplamente utilizado nos próximos Capítulos, principalmente nos blocos de extração de atributos. Este método é a interpolação linear.

A interpolação linear é um dos métodos comumente utilizados para estimar valores entre pares de valores adjacentes de seqüências discretas no tempo. 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

PUC-Rio - Certificação Digital Nº 0310467/CA

$$x_u[n] = \begin{cases} x[n/L], & n = 0, L, 2L, 3L, \ldots \\ 0, & para \quad n \neq 0, L, 2L, 3L, \ldots \end{cases} \tag{3.1}$$

onde $L > 1$ é o fator de sobre-amostragem que se quer utilizar e $L - 1$ 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 [21].

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

$$y[n] = x_u[n] + \frac{1}{2}(x_u[n-1] + x_u[n+1]) \tag{3.2}$$

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

$$y[n] = x_u[n] + \frac{1}{3}(x_u[n-1] + x_u[n+2]) + \frac{2}{3}(x_u[n-2] + x_u[n+1]) \tag{3.3}$$

Só foram apresentadas as expressões para o sinal interpolado pelos fatores 2 e 3, pois como será visto mais adiante, só serão considerados aumentos de taxa de um determinado parâmetro por estes mesmos fatores.

3.3. Conclusão

Foi apresentada nesse capítulo a definição de sistema de reconhecimento distribuído e as abordagens utilizadas para implementá-lo. Detalhou-se também a estrutura do sistema de reconhecimento que é a base dos experimentos e análises desta dissertação.

No capítulo seguinte, será feita a apresentação teórica dos atributos de reconhecimento que serão utilizados para a implementação do sistema de reconhecimento apresentado na Seção 3.2.

PUC-Rio - Certificação Digital Nº 0310467/CA