5
Análise dos Atributos de Voz em Reconhecimento Distribuído

Neste Capítulo será realizada a análise de alguns parâmetros do sistema de reconhecimento, a fim de restringir os cenários de teste quando da utilização do sistema com um codificador padrão. Para esta análise não será utilizado o sistema completo, como apresentado na Fig. 3.11. Nesse sistema, retirou-se do codificador o bloco que extrai a excitação e o bloco que quantiza as LSFs e a excitação. Em decorrência disto, retirou-se do decodificador o bloco de de-quantização e o bloco de recuperação de voz. Com estas modificações tem-se o sistema da Fig. 5.1 a ser analisado neste capítulo.

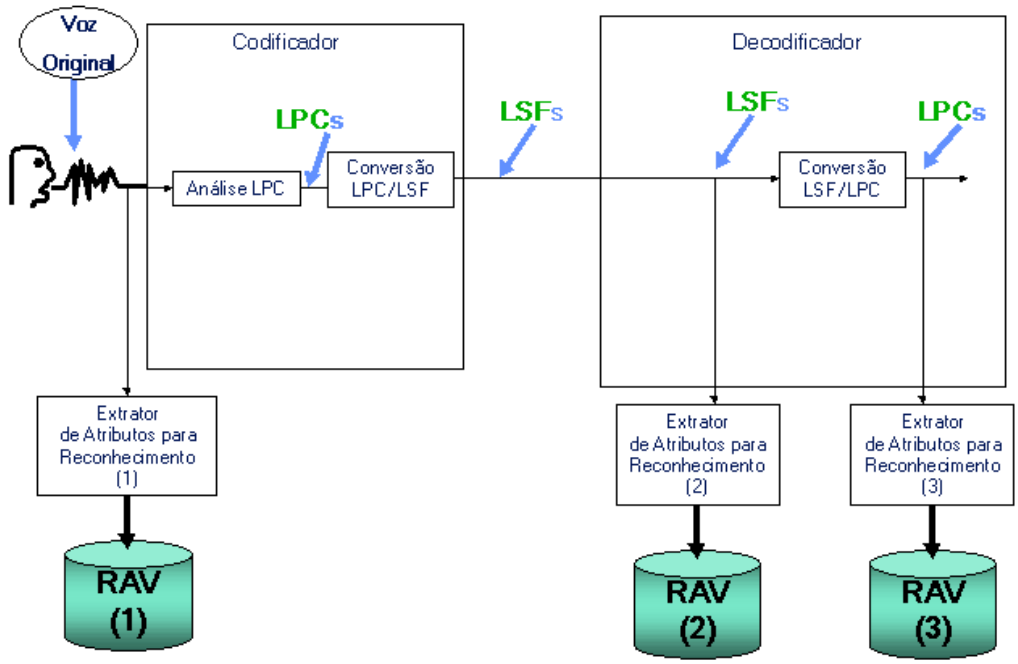

Figura 5.1 – Sistema para análise sem quantização

As modificações aqui consideradas fizeram com que o novo sistema não fosse capaz de reconstruir voz a partir dos dados obtidos pelo codificador. Sendo assim, o reconhecedor que utilizava voz reconstruída – RAV (4) – da Fig. 3.11 perdeu sua funcionalidade e foi também retirado do sistema. O novo sistema, portanto, é capaz de efetuar reconhecimento apenas a partir de atributos extraídos de voz original, de LSF e de LPC.

Na análise apresentada neste capítulo, os quadros terão duração de 25 ms e o espaçamento entre seus centros será de 10 ou 20 ms, dependendo da taxa que se deseja gerar, os parâmetros LPC / LSF e os atributos de reconhecimento.

O espaçamento entre os centros dos quadros de 10 ms foi escolhido devido ao fato do reconhecimento de voz ser realizado comumente com este espaçamento de quadro. Já o espaçamento de 20 ms foi escolhido por ser o espaçamento encontrado em alguns codificadores de voz para redes IP e ambiente celular. Assim, todos os atributos de reconhecimento do sistema da Fig. 5.1 serão obtidos utilizando esses dois espaçamentos entre centros de quadro.

Todos os extratores de atributos estarão gerando sempre um conjunto de 10 parâmetros com suas respectivas derivadas, totalizando 20 atributos de reconhecimento. Este conjunto de parâmetros de reconhecimento foi escolhido tendo como base os trabalhos de Milner [35, 36], Coqueiro [37] , Lilly [38], Kim [15] e Choi [33].

A derivada Δ tem como objetivo capturar as variações dinâmicas do espectro do sinal de voz [39, 40, 41]. A forma de se calcular os coeficientes Δ utilizada [42] é por

$$\Delta c_k(n) = \frac{\sum_{m=-N_d}^{N_d} m c_k(n+m)}{\sum_{m=-N_d}^{N_d} m^2} \tag{5.1}$$

onde $c_k(n)$ é o k-ésimo atributo no instante n e N_d é a distância em relação ao instante n para a qual se quer calcular a diferença. O valor mais comum para N_d é 2, o qual será utilizado nesta dissertação.

Algumas características do HMM devem ser estabelecidas para que se possa efetuar o treinamento e, posteriormente, os testes de reconhecimento. Estas características são o número de estados e o número de gaussianas que compõe as misturas de cada estado. Para determinar estes parâmetros, esta dissertação se baseou no trabalho de Gallardo-Antolín [43], que propunha a utilização de três gaussianas por estado, para compor a mistura e o número de estados a ser determinado pelo número de fones em cada palavra a ser reconhecida. Verificou-se, para a base a ser trabalhada nesta dissertação, que o número de fones das palavras variava de 1 até 5. Porém, para reduzir a dificuldade de implementação do sistema de reconhecimento, optou-se por utilizar, para todos os modelos de palavras, o mesmo número de estados e gaussianas por estado. Sendo assim, determinou-se que nos RAVs desta dissertação seriam utilizados, três gaussianas por estado e o número de estados assumiria os valores 3, 4 e 5 – para se verificar qual o valor teria melhor compromisso entre complexidade e desempenho para a base de locuções apresentada no Capítulo 2. Esta base de locuções 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.

Outra questão importante diz respeito a que percentuais da base de locuções se deve utilizar para teste e treinamento. Na literatura de reconhecimento encontrou-se duas distribuições principais:

- 1ª distribuição – 50% de Treinamento e 50% de Teste [37, 44, 45]
- 2ª distribuição – 70% de Treinamento e 30% de Teste [15, 38, 39]

Optou-se por testar as duas distribuições com dois parâmetros diferentes (MFCC de voz original – RAV(1) – e MPCC de LSF – RAV (2)) em 10 ms e 20ms e verificar como essas distribuições afetariam o desempenho de reconhecimento. O resultado do teste é apresentado nas Tabs. 5.1 e 5.2.

Número de estados	3 estados		4 estados		5 estados	
Tipo do Atributo de Reconh.	MFCC	MPCC	MFCC	MPCC	MFCC	MPCC
50% Trein. / 50% de Teste	90,4%	88,9%	98,7%	97,0%	99,1%	97,2%
70% Trein. / 30% de Teste	90,6%	89,0%	99,0%	97,2%	99,4%	97,5%

Tabela 5.1 – Resultado do teste de reconhecimento de voz para os atributos obtidos a cada 10 ms em porcentagem de acertos

Número de estados	3 estados		4 estados		5 estados	
Tipo do Atributo de Reconh.	MFCC	MPCC	MFCC	MPCC	MFCC	MPCC
50% Trein. / 50% de Teste	84,3%	82,6%	94,2%	92,2%	94,6%	92,6%
70% Trein. / 30% de Teste	84,5%	82,9%	94,5%	92,7%	95,0%	93,1%

Tabela 5.2 – Resultado do teste de reconhecimento de voz para os atributos obtidos a cada 20 ms em porcentagem de acertos

Verifica-se que o uso de uma distribuição de 70% de treinamento e 30% de teste fornece sempre um melhor resultado. Logo, será adotada esta distribuição como padrão para toda a dissertação. Conclusões sobre o número de estados serão apresentados na seção seguinte.

5.1. Obtenção dos Atributos de Voz em 10ms e 20ms e o Desempenho do Reconhecimento para Cada Taxa

Os testes a serem realizados nesta Seção têm como finalidade principal determinar quais os atributos de reconhecimento possuem melhor compromisso entre desempenho de reconhecimento e carga computacional para sua obtenção.

No sistema da Fig. 5.1 serão obtidos:

- Extrator de atributos (1) – obtém atributos de voz original, efetuará a extração da MFCC de voz original em 10 ms e em 20 ms
- Extrator de atributos (2) – obtém dos parâmetros LSFs os atributos PCC, PCEP, MPCC e MPCEP em 10 ms e em 20 ms
- Extrator de atributos (3) – obtém dos parâmetros LPC os atributos LPCC e MLPCC em 10 ms e em 20 ms

Cabe ressaltar que o MFCC de voz original só está sendo obtido para se ter uma referência de desempenho de reconhecimento para os outros atributos. Note-se que este atributo não poderá ser utilizado no sistema tratado neste trabalho, onde não há transmissão de informação adicional de voz para o sistema de reconhecimento, além do que o codificador padrão já transmite.

Obviamente não serão obtidos atributos MFCC a partir de voz reconstruída porque o sistema considerado neste Capítulo não codifica excitação, não permitindo assim a reconstrução da voz. A opção de não codificar a excitação foi tomada devido ao fato dos atributos PCC, PCEP, LPCC, MPCC, MPCEP e MLPCC em teste não precisarem da mesma para a sua obtenção.

Nas Tabs. 5.3 e 5.4 são apresentados os resultados dos testes de reconhecimento quando os atributos são extraídos a cada 10 ms e 20 ms, respectivamente.

X	LPCC	PCC	PCEP	MLPCC	MPCC	MPCEP	MFCC
3 estados	86,0%	85,4%	85,7%	89,0%	89,0%	89,0%	90,6%
4 estados	95,0%	94,4%	94,8%	97,9%	97,2%	97,9%	99,0%
5 estados	95,8%	94,6%	95,0%	98,3%	97,5%	98,2%	99,4%

Tabela 5.3 – Resultado do teste de reconhecimento de voz para os atributos obtidos a cada 10 ms em porcentagem de acertos

X	LPCC	PCC	PCEP	MLPCC	MPCC	MPCEP	MFCC
3 estados	79,9%	79,5%	79,7%	83,0%	82,9%	82,9%	84,5%
4 estados	90,1%	89,6%	89,9%	92,9%	92,7%	92,9%	94,5%
5 estados	90,8%	90,2%	90,4%	93,8%	93,1%	93,7%	95,0%

Tabela 5.4 – Resultado do teste de reconhecimento de voz para os atributos obtidos a cada 20 ms em porcentagem de acertos

Pode-se observar, das Tab. 5.3 e 5.4, que para todos os atributos, o uso de HMMs com cinco estados apresenta melhor desempenho do que o obtido com HMMs de 3 ou 4 estados. Verifica-se, ainda, que os atributos na escala mel (MLPCC, MPCEP e MPCC), sempre possuem melhor desempenho que os atributos na escala de freqüência real (LPCC, PCEP e PCC). A partir desses resultados decidiu-se adotar como padrão para o restante dos testes a serem realizados nesta dissertação, as HMMs de cinco estados, e, ainda, manter os testes apenas com os parâmetros na escala de freqüência Mel (MLPCC, MPCEP e MPCC).

Nas Tabs. 5.3 e 5.4 pode-se verificar, também, que os atributos de reconhecimento de voz para ambiente distribuído (MLPCC, MPCEP e MPCC), possuem resultados bastante bons quando comparados com a MFCC obtida de voz original.

É importante lembrar que os atributos MPCEP e MPCC no decodificador completo são obtidos diretamente das LSFs de-quantizadas, correspondendo ao primeiro estágio do decodificador, enquanto que os atributos MLPCC são obtidos do segundo estágio do decodificador, após a conversão LSF / LPC. Essas características tornam os atributos MPCEP e MPCC mais leves computacionalmente do que os MLPCC para os sistemas de reconhecimento que não tenham como finalidade reconstruir a voz. Pode-se observar na Tabs. 5.3 e 5.4 que a perda máxima de desempenho que se pode sofrer com a simplificação realizada para a obtenção dos MPCC e MPCEP em relação ao MLPCC é de 0,7%, quando esta perda existir.

Uma observação também interessante que se pode tirar das Tabs. 5.3 e 5.4 é que os atributos MPCEP sempre possuem desempenho igual ou melhor do que os MPCC, apesar dos MPCEP representarem uma aproximação mais grosseira que os MPCC para os atributos MLPCC.

Se forem apreciados então apenas os resultados de MPCEP e MLPCC das Tabs. 5.3 e 5.4 pode-se observar que quando existe diferença de desempenho é de no máximo 0,1%, o que é bastante interessante quando se analisa em conjunto com a complexidade computacional, pois se verifica que o MPCEP possui desempenho equivalente ao MLPCC e representa uma economia de processamento.

Com a finalidade de destacar os resultados mais relevantes e permitir comparações futuras, foi montada a Tab. 5.5, que apresenta os resultados de percentual de acerto de reconhecimento nos testes para os atributos MLPCC, MPCC e MPCEP, nas duas diferentes taxas.

X	MLPCC	MPCC	MPCEP
Atributos obtidos a cada 10 ms	98,3%	97,5%	98,2%
Atributos obtidos a cada 20 ms	93,8%	93,1%	93,7%

Tabela 5.5 – Resultado do teste de reconhecimento de voz para HMMs de cinco estados

Comparando o desempenho de reconhecimento para 10 ms e 20 ms, fica claro que existe um espaço bastante grande para ganho de desempenho de reconhecimento para o sistema que extrai os atributos em intervalos de 20 ms (diferença de aproximadamente 4% no percentual de acerto de reconhecimento).

O passo seguinte é, então, buscar um bom domínio e aplicar uma técnica de interpolação para aproveitar este potencial.

Nesta dissertação, buscar-se-á determinar qual o melhor domínio para se aplicar a interpolação linear destes atributos.

5.2. Interpolação de Atributos de Voz para Reconhecimento

Nesta Seção será analisada a interpolação linear dos atributos de reconhecimento apresentados na Tab. 5.5, em diferentes domínios de interpolação. Os domínios de interpolação a serem analisados são:

- o domínio dos próprios parâmetros;
- o domínio das LSFs;
- o domínio dos parâmetros LPC.

Cabe-se ressaltar que para se obter MLPCC a partir dos parâmetros LPC é necessário que se recebam as LSFs no decodificador e as transforme em parâmetros LPC. Já os atributos MPCC e MPCEP são obtidos diretamente das LSFs recebidas no decodificador. Logo os atributos MPCC e MPCEP não podem ser interpolados no domínio LPC, o que ao contrário não ocorre com o atributo MLPCC que pode sim ser interpolado neste domínio.

Para efetuar a interpolação linear dos atributos de 20 ms para 10 ms será utilizado um fator de interpolação de valor 2, como apresentado em (3.2).

A Tab. 5.6 mostra os resultados obtidos para a interpolação dos diversos atributos nos diversos domínios, bem como os valores apresentados na Tab 5.5, permitindo assim uma fácil apreciação e comparação dos resultados.

X	MLPCC	MPCC	MPCEP
Atributos obtidos a cada 10 ms sem interpolação	98,3%	97,5%	98,2%
Atributos obtidos a cada 10 ms com interpolação de LSF	96,0%	95,7%	96,0%
Atributos obtidos a cada 10 ms com interpolação de LPC	93,8%	não	não
Atributos obtidos a cada 10 ms com interpolação dos próprios parâmetros	93,9%	93,8%	94,4%
Atributos obtidos a cada 20 ms sem interpolação	93,8%	93,1%	93,7%

Tabela 5.6 – Resultado do teste de reconhecimento de voz com ou sem interpolação.

Comparando-se primeiramente o desempenho do uso da interpolação linear no domínio dos próprios atributos com os atributos não interpolados na Tab. 5.6,

verifica-se que praticamente não houve ganho com a utilização da interpolação no domínio dos próprios parâmetros. O parâmetro que conseguiu um maior ganho com essa interpolação foi o parâmetro MPCEP que teve seu desempenho aumentado de 0,7% na porcentagem de acerto de reconhecimento.

A interpolação no domínio LPC foi a que apresentou pior desempenho em todos os sentidos, pois não representou ganho de reconhecimento para o único parâmetro que podia ser interpolado neste domínio.

É interessante comparar, agora, o desempenho dos atributos obtidos a cada 10 ms através da interpolação linear no domínio das LSFs com os atributos obtidos a cada 20ms não interpolados. Da Tab. 5.6, pode-se verificar que se obtém um ganho de aproximadamente 2,2%, 2,6% e 2,3% para os parâmetros MLPCC, MPCC e MPCEP, respectivamente, quando se usa interpolação no domínio das LSFs. Com esses ganhos, esses parâmetros se aproximam bem mais dos resultados obtidos com os atributos gerados a cada 10 ms. Porém se forem apreciadas em conjunto a linha onde tem-se a interpolação no domínio das LSFs (2ª linha da Tab. 5.6) e dos atributos obtidos a cada 10ms sem interpolação (1ª linha da Tab. 5.6), verifica-se que ainda existe uma boa margem para melhoria de desempenho.

5.3. Conclusão

Neste Capítulo, foram efetuados diversos testes de reconhecimento usando o sistema sem quantização ilustrado na Fig 5.1. Os resultados apresentados neste capítulo permitirão a realização de testes mais direcionados quando da utilização do codificador padrão no Capítulo 6.

As principais conclusões obtidas neste capítulo foram:

1. Optou-se por utilizar uma distribuição de 70% da base de locuções para treinamento e 30% para teste;
2. HMMs com cinco estados e três gaussianas por estado, para esta base de dados, é a que apresenta melhor desempenho para todos os atributos;
3. Os atributos na escala mel são os que apresentam sempre melhor desempenho de reconhecimento;

4. Os atributos MLPCC são obtidos dos parâmetros LPC, recuperados da conversão LSF/LPC. Já os atributos MPCEP e MPCC são obtidos dos parâmetros LSF. Os mesmos foram aqui considerados pois podem ser obtidos a partir dos parâmetros transmitidos pelo codificador padrão e possuem desempenho comparável com o desempenho de MFCC obtida de voz original;
5. É interessante usar a interpolação linear dos atributos para aumentar a taxa dos mesmos, pois esta aumenta o desempenho do reconhecedor;
6. O melhor domínio para se efetuar a interpolação linear de qualquer atributo é o domínio das LSFs, onde se obtém resultados significativamente acima dos obtidos utilizando interpolação no domínio dos parâmetros LPC e no domínio dos próprios atributos;
7. Apesar dos atributos MPCEP serem de obtenção mais simples que os MPCC, porque os mesmos são obtidos de LSF de-quantizadas e são uma aproximação mais grosseira dos MLPCC do que os MPCC, eles possuem desempenho um pouco superior aos dos MPCC .

No próximo capítulo serão realizados testes com os atributos e técnicas que apresentaram melhor desempenho neste capítulo, porém utilizando o *codec* de voz padrão ITU-T G.723.1. Esse codificador é um codificador de voz padronizado pela ITU (International Telecommunication Union) para utilização em redes IP. No próximo capítulo também será obtido para efeito de comparação, o desempenho do atributo MFCC obtido a partir de voz reconstruída obtida pelo decodificador ITU-T G.723.1.