7
Conclusões e Sugestões para Trabalhos Futuros

A grande motivação para o presente trabalho foi a constatação de que os atributos de reconhecimento, usualmente utilizados para o reconhecimento automático de voz, assim como sua taxa de obtenção, não são adequados ao ambiente de processamento distribuído.

7.1.
Conclusões

Com o intuito de reproduzir o ambiente de reconhecimento distribuído mais comum, onde as pessoas através do telefone são levadas a optar entre itens do serviço que se quer utilizar, escolheu-se trabalhar com o reconhecimento de palavras isoladas (dígitos) independentes do locutor. Para isso se utilizou uma base de locuções 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.

No ambiente de redes IP e sistemas móveis celulares são utilizados para a codificação de voz codificadores paramétricos padrões, normalmente baseados em parâmetros LSF. Optou-se por montar, então, um sistema de reconhecimento distribuído baseado nos parâmetros destes codificadores.

Primeiramente se considerou um sistema mais simples do que o que inclui um codificador padrão. A idéia era apenas examinar em dois períodos de geração diferentes (10 ms e 20 ms), os atributos de reconhecimento:

- obtidos de parâmetros LPC – que são os atributos LPCC e MLPCC;
- obtidos de parâmetros LSF – que são os atributos PCEP, PCC, MPCC e MPCEP; e
- obtidos de voz original – que é o atributo MFCC.

O objetivo de se obter a MFCC de voz original era se ter uma referência de comparação para os demais atributos de reconhecimento que se queria testar.

Efetuaram-se testes para determinar quais porcentagens da base de locuções deveriam ser utilizados para treinamento e teste. Das duas opções existentes na literatura verificou-se que era melhor utilizar a distribuição de 70% para treinamento e 30 % para teste.

Simultaneamente foram realizados testes com diferentes configurações de HMM, que levaram a concluir que a utilização de HMMs de cinco estados e três gaussianas por estado era a configuração de melhor relação entre desempenho e custo computacional.

Concluiu-se, ainda, que os atributos na escala Mel (MPCC, MLPCC, MPCEP) apresentavam sempre melhor desempenho que os atributos na escala real (PCC, LPCC, PCEP). Por esse motivo, os atributos MPCC, MLPCC, MPCEP foram mantidos para os demais testes abandonando-se os demais atributos (PCC, LPCC, PCEP).

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.

Note-se que os parâmetros MPCC e MPCEP representam aproximações do MLPCC. Apesar disso, a perda máxima de desempenho sofrida com a simplificação realizada para a obtenção dos MPCC e MCEP em relação ao MLPCC é de 0,7%, quando esta perda existir.

Uma observação também interessante que se pode tirar 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.

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 de 4% no percentual de acerto de reconhecimento).

O passo seguinte foi, então, buscar um bom domínio e aplicar uma técnica de interpolação para aproveitar este potencial. Com isso fica claro que o objetivo de se obter os atributos em duas diferentes taxas era verificar se o uso da técnica de interpolação linear, nos diversos domínio, seria capaz de conseguir alguma

melhoria de desempenho dos atributos mais eficientes e quanto o desempenho final se aproximaria dos atributos já obtidos na taxa mais alta (10 ms).

Na comparação do desempenho entre o uso ou não da interpolação linear no domínio dos próprios atributos, verificou-se que praticamente não houve ganho com a utilização da interpolação. O parâmetro que conseguiu um melhor desempenho 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 forneceu melhoria de reconhecimento para o único parâmetro que podia ser interpolado neste domínio, que era o MLPCC.

Comparou-se, então, 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. Com essa interpolação obteve-se um ganho de aproximadamente 2,2%, 2,6% e 2,3% para os parâmetros MLPCC, MPCC e MPCEP, respectivamente. Esses ganhos possibilitaram que os desempenhos se aproximassem bem mais dos resultados obtidos com os atributos gerados a cada 10 ms. Porém, quando foram apreciadas em conjunto a interpolação no domínio das LSFs e os atributos obtidos a cada 10ms sem interpolação, verificou-se que ainda existe uma boa margem para melhoria de desempenho.

Concluiu-se, então, para a interpolação linear, que era interessante utilizá-la pelo fato de sempre oferecer ganho de desempenho. Porém, verificou-se que era mais adequado efetuá-la no domínio das LSFs.

A partir destas conclusões, passou-se para a utilização de um sistema mais completo, onde o mesmo já dispunha de um *codec* de voz padrão, ITU-T G.723.1, visando testar os atributos (MPCC, MLPCC, MPCEP) em um cenário mais próximo do real. Decidiu-se também que neste cenário seria obtida a MFCC de voz reconstruída, para que se pudesse ter uma referência de comparação.

Neste sistema com o *codec* foram utilizados:

- 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 foi 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.

Concluiu-se 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 neste cenário o ruído presente é o ruído de quantização do codificador. Isto explica o baixo desempenho quando a MFCC é 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]. Isso justifica a observação anterior.

Verificou-se, ainda, que mesmo com o uso do codificador, o parâmetro MPCEP, apesar de ser o de mais simples obtenção, foi o que propiciou melhor desempenho no teste de reconhecimento.

Finalmente, é possível concluir 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.

7.2. Sugestões para Trabalhos Futuros

Uma primeira sugestão interessante seria analisar o desempenho do sistema utilizando o *codec* ITU-T G.723.1 e os atributos aqui propostos, bem como outros, na presença de perda de pacotes. A realização do estudo destes atributos seria realizada para diversas taxas de perda de pacotes, respeitando a distribuição estatística característica para cada ambiente nas suas diversas situações de utilização, bem como para os diversos codificadores existentes para estes ambientes.

Fica também como sugestão para trabalhos futuros, a implementação dos atributos apresentados nesta dissertação em outros codificadores de voz que

tenham como base parâmetros LPC e LSF. Destacam-se os codificadores utilizados em aplicações militares que possuem codificadores especialmente projetados e que podem ter comportamentos peculiares, apesar de serem baseados nos mesmos parâmetros (devido às baixíssimas taxas de codificação e aspectos de segurança neles inseridos).

Uma experiência interessante a ser realizada consiste em verificar o comportamento de atributos robustos ao ruído, como por exemplo o ZCPA, quando obtidos de voz reconstruída por um decodificador padrão e compará-los com os melhores atributos aqui obtidos.