ATRIBUTOS EFICIENTES EM RECONHECIMENTO AUTOMÁTICO DE VOZ DISTRIBUÍDO

Vladimir F. S. de Alencar*
vladimir@cetuc.puc-rio.br

Abraham Alcaim*
alcaim@cetuc.puc-rio.br

*CETUC-PUC-RJ, Rio de Janeiro, RJ, Brasil

RESUMO

Este artigo descreve e examina atributos eficientes para sistemas de reconhecimento automático de voz (RAV) distribuídos. Os atributos são extraídos a partir de transformações dos parâmetros dos codificadores utilizados em redes móveis celulares e em redes IP que são as principais aplicações visadas em reconhecimento de voz distribuído. Em particular, são analisados atributos obtidos de parâmetros LPC e LSF, em intervalos de 10 ms e 20 ms, utilizando um reconhecedor independente do locutor baseado em HMM. É examinado também o melhor domínio para se realizar a interpolação dos parâmetros em reconhecimento automático de voz distribuído.

PALAVRAS-CHAVE: Reconhecimento de Voz Distribuído, LSF, HMM.

ABSTRACT

This paper describes and examines efficient features for distributed automatic speech recognition (ASR) systems. The features are extracted by means of transformations of codecs parameters. These codecs are the ones used in cellular mobile networks and IP networks, which are the main applications in distributed speech recognition. In particular, features obtained from LPC and LSF parameters, in intervals of 10 ms and 20 ms, are analyzed in an HMM-based speaker independent recognizer. Moreover, we examine the best domain to perform features interpolation in distributed automatic speech recognition systems.

KEYWORDS: Distributed Speech Recognition, LSF, HMM.

Artigo submetido em 30/05/2007
1a. Revisão em 18/08/2007
2a. Revisão em 18/10/2007
Aceito sob recomendação do Editor Associado Prof. Ivan Nunes Da Silva

1 INTRODUÇÃO

Com o crescimento gigantesco da *Internet* e dos sistemas de comunicações móveis celulares, as aplicações de processamento de voz nessas redes têm despertado grande interesse. Um problema particularmente importante nessa área consiste no reconhecimento automático de voz (RAV) em um sistema servidor, baseado nos parâmetros acústicos calculados e quantizados no terminal do usuário. Sistemas desse tipo, denominados Reconhecimento de Voz Distribuídos, são uma opção muito atraente por causa da alta complexidade e grande quantidade de memória requeridas em sistemas RAV.

Os esquemas de codificação de voz usados em sistemas de comunicações móveis e redes IP operam a baixas taxas de bits e utilizam, em geral, codificação preditiva linear ou LPC (*Linear Predictive Coding*), com base em um modelo de produção da fala. Nesse modelo, um sinal de excitação é aplicado a um filtro tudo-pólo (caracterizado por parâmetros LPC), que representa a informação da envoltória espectral do sinal de voz. Usualmente os parâmetros LPC são transformados para LSF (*Line Spectral Frequencies*), devido às propriedades atraentes destes últimos para os processos de quantização e interpolação.

No caso de sistemas de RAV distribuídos é preferível utilizar diretamente os parâmetros do codec do que extraí-los a partir do sinal decodificado (Choi, 2000). Como em

geral estes parâmetros não são os mais indicados como atributos de voz para o sistema de reconhecimento remoto, é importante que sejam examinadas diferentes transformações dos parâmetros do codec, que permitam uma melhor performance do reconhecedor. É também importante que seja investigado o melhor domínio de interpolação dos parâmetros em reconhecimento de voz distribuído. Os resultados dessa investigação são apresentados como uma contribuição adicional deste artigo.

Nas Seções 2 e 3 deste artigo são apresentados os atributos obtidos a partir dos parâmetros LPC e LSF, respectivamente. Resultados experimentais são fornecidos e analisados na Seção 4. Finalmente, a Seção 5 resume as conclusões do trabalho.

2 ATRIBUTOS DE RECONHECIMENTO OBTIDOS DE TRANSFORMAÇÕES DE PARÂMETROS LPCS

Esta seção trata dos atributos de reconhecimento que podem ser extraídos diretamente dos parâmetros LPC (*Linear Predictive Coefficients*), sem a necessidade de reconstrução do sinal de voz. Esta abordagem se deve ao fato de que, dentro dos decodificadores de voz utilizados para telefonia celular e voz sobre IP, os parâmetros LPC já serem produzidos naturalmente no seu processo de recuperação de voz, em um estágio anterior à obtenção da voz reconstruída. Sendo assim, atributos de reconhecimento de voz, obtidos neste estágio, são mais leves computacionalmente do que os obtidos de voz reconstruída, pois evitam a necessidade de recuperação da mesma. Além disso, como comentado anteriormente, gerar atributos a partir da voz reconstruída por um decodificador é menos recomendável do que obtê-los diretamente dos parâmetros do codec.

Os atributos de reconhecimento que podem ser obtidos dos parâmetros LPC são os LPCC (LPC *Cepstrum*) e MLPCC (*Mel-Frequency* LPCC) - (Ohshima, 1993). Os atributos LPCC são obtidos a partir dos parâmetros LPC por uma fórmula recursiva a ser deduzida a seguir. Já os MLPCC são obtidos dos LPCCs através de uma filtragem passa-tudo de primeira ordem que também será apresentada a seguir.

2.1 LPC Cepstrum (LPCC)

O processo de obtenção dos atributos LPCC a partir dos coeficientes LPC será formulado no domínio da freqüência, a partir do logaritmo complexo da função de transferência do sistema LPC. Esse procedimento é análogo ao cálculo do *cepstrum* no domínio da Transformada Discreta de Fourier a partir do sinal de voz (Ohshima, 1993).

Primeiramente, se constrói a função de transferência do sistema LPC de ordem p (p inteiro positivo), que é dada por

$$H(z) = \frac{G}{A(z)} = \frac{G}{1 - \sum_{i=1}^{p} a_i z^{-i}} \tag{1}$$

onde a_i é o i-ésimo parâmetro LPC, G é o fator de ganho e o denominador é o filtro inverso A(z).

Calculando a derivada do polinômio complexo $\ln(H(z))$, em relação a $\rho = z^{-1}$, obtém-se

$$\frac{\partial}{\partial \rho} \ln(H(\rho)) = \frac{\partial}{\partial \rho} [\ln(G) - \ln(A(\rho))] = \frac{\sum_{l=1}^{p} l a_l \rho^{l-1}}{1 - \sum_{i=1}^{p} a_i \rho^i} \tag{2}$$

Como $H(z)$ é a função de transferência do sistema LPC obtido no transmissor (que é um sistema causal, pois para p positivo, esta forma se remete a um filtro IIR causal conforme descrito na referência (Mitra, 1998)), onde são utilizados métodos para garantir sua estabilidade, a mesma deverá ter todos os seus pólos dentro do círculo unitário. Isso significa que $\ln(H(z))$ é unilateral, o que leva a escrever

$$C(z) = \sum_{i=0}^{+\infty} c_i z^{-i} \tag{3}$$

onde c_i é o i-ésimo atributo LPCC e $C(z)$ é logaritmo complexo da função de transferência do sistema LPC.

Derivando $C(z)$ em relação a ρ e igualando a (2), obtém-se a equação

$$\sum_{j=1}^{+\infty} j c_j \rho^{j-1} = \frac{\sum_{l=1}^{p} l a_l \rho^{l-1}}{1 - \sum_{i=1}^{p} a_i \rho^i} \tag{4}$$

que pode ser reescrita na forma

$$\left(\sum_{j=1}^{+\infty} j c_j \rho^{j-1} \right) \left(1 - \sum_{i=1}^{p} a_i \rho^i \right) = \sum_{l=1}^{p} l a_l \rho^{l-1} \tag{5}$$

Comparando os coeficientes das séries de ρ em ambos os lados, pode-se obter a equação recursiva que permite o cálculo dos atributos LPCC, onde o parâmetro c_0 é

determinado pelo termo constante da definição original de$H(z)$. Essa equação é dada por

$$c_i = \begin{cases} \ln(G) & i = 0 \\ a_1 & i = 1 \\ a_i + \sum_{j=1}^{i-1} \frac{i-j}{i} c_{i-j} a_j & 1 < i \leq p \\ \sum_{j=1}^{p} \frac{i-j}{i} c_{i-j} a_j & i > p \end{cases} \quad (6)$$

2.2 *Mel-Frequency* LPC Cepstrum (MLPCC)

O processo de obtenção do atributo MLPCC passa pela transformação do eixo de freqüência real para o eixo de freqüência na escala mel dos atributos LPCC (Choi, 2000). Para ser realizada esta transformação, utiliza-se um banco de n filtros passa-tudo de primeira ordem iguais entre si que permitem efetuar a transformação do eixo de freqüência real para o eixo de freqüência na escala mel - onde n é o número de atributos LPCC obtidos através de (6) - (Oppenheim, 1972). Todos os filtros deste banco terão sua função de transferência $\psi(z)$ passa-tudo de primeira ordem (Mitra, 1998) dada pela expressão

$$\psi(z) = \frac{z^{-1} - a^*}{1 - az^{-1}} \quad (7)$$

onde a é o coeficiente deste filtro passa-tudo e a^* é o conjugado de a. Cada coeficiente cepestral c_i deve passar por um filtro deste banco de filtros, sendo a saída de cada filtro o atributo MLPCC c_i^m.

Como o objetivo de cada filtro é realizar a aproximação da escala mel de freqüências, tem-se que analisar o que a função de transferência em (7) está realizando com os eixos das freqüências. Para isto, será considerado a real, o que facilitará a implementação do filtro (Wölfel, 2003).

Para que seja feita esta análise, deve-se reescrever ψ, em função de $e^{j\Omega}$, como

$$\psi\left(e^{j\Omega}\right) = e^{-j\theta(\Omega)} \quad (8)$$

pois isto permite analisar o que está sendo feito com os eixos de freqüência, onde Ω é a freqüência real e

$$\theta(\Omega) = \arctan\left[\frac{\left(1-a^2\right) sen\Omega}{\left(1+a^2\right)\cos\Omega - 2a}\right] \quad (9)$$

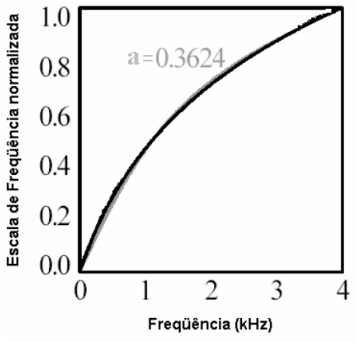

Figura 1: Aproximação da escala mel pela transformação bilinear, escala mel em preto e aproximação em cinza.

é a freqüência na escala mel expressa em função da freqüência real Ω.

Ao se ajustar a curva de $\theta(\Omega)$ à curva da escala mel, para a freqüência de amostragem de 8 kHz, por meio da variação do termo a, obtém-se a curva da Figura 1 (Wölfel, 2003), onde $\Omega = 2\pi f/f_s$, f_s é igual a 8kHz e o eixo vertical está normalizado em relação a π.

O valor de a real que melhor aproxima a curva da escala mel é 0,3624 (Wölfel, 2003), como mostrado na Figura 1.

3 ATRIBUTOS DE RECONHECIMENTO OBTIDOS DE TRANSFORMAÇÕES DE PARÂMETROS LSFS

As *Line Spectral Frequencies* (LSFs) são usualmente utilizadas para codificação de voz devido à sua grande eficiência de codificação e suas propriedades atraentes para interpolação (Kleijn, 1995).

A obtenção de atributos de reconhecimento, a partir das LSFs evita tanto a necessidade de utilização de um decodificador de voz, como a transformação para LPC no receptor para a realização do reconhecimento. O sistema de reconhecimento de voz distribuído que evita tal utilização se torna mais leve computacionalmente que quaisquer outros baseados em parâmetros que dependam da reconstrução da voz ou dos parâmetros LPC. Os atributos de reconhecimento que podem ser obtidos desta forma são os PCC (*Pseudo-Cepstral Coefficients*) - (Kim, 2000), PCEP (*Pseudo-Cepstrum*) - (Choi, 2000), MPCC (*Mel-Frequency* PCC) - (Kim, 2000) e MPCEP (*Mel-Frequency* PCEP) - (Choi, 2000).

Cabe ressaltar que estes atributos, obtidos diretamente de LSF, são aproximações dos atributos LPCC e MLPCC, anteriormente apresentados. Estas aproximações têm como finalidade evitar a necessidade de recuperação dos

parâmetros LPC, reduzindo a complexidade computacional do sistema.

3.1 *Pseudo-Cepstral Coefficients* (PCC)

O atributo PCC é obtido diretamente de LSF. Porém, sua dedução utiliza o atributo LPCC a partir de LPC, com manipulações matemáticas e aproximações que permitem obtê-lo diretamente de LSF sem necessitar dos parâmetros LPC. Esses procedimentos serão apresentados em seguida.

Um filtro inverso de ordem p estável, onde todas as raízes se encontram dentro do círculo unitário, é definido por

$$A_p(z) = \sum_{i=0}^{p} b_i z^{-i} \tag{10}$$

onde $b_0 = 1$ e b_i é o i-ésimo coeficiente de predição linear (LPC) sendo que $b_i = -a_i$ quando $i \geq 1$.

As LSFs de ordem p são definidas como sendo as raízes complexas dos polinômios $P(z)$ e $Q(z)$, os quais são expressos por

$$P(z) = A_p(z) + z^{-(p+1)} A_p(z^{-1}) \tag{11}$$

$$Q(z) = A_p(z) - z^{-(p+1)} A_p(z^{-1}) \tag{12}$$

Para obter a relação entre LPCC e LSF é preciso inicialmente realizar a multiplicação de (11) e (12), resultando em

$$P(z)Q(z) = A_p^2(z) \left[1 - \left\{ \frac{z^{-(p+1)} A_p(z^{-1})}{A_p(z)} \right\}^2 \right]$$

$$= (1 - z^{-2}) \prod_{i=1}^{p} \left(1 - e^{jw_i} z^{-1}\right) \left(1 - e^{-jw_i} z^{-1}\right) \tag{13}$$

onde w_i é o i-ésimo parâmetro LSF. Definindo

$$R(z) = \frac{z^{-(p+1)} A_p(z^{-1})}{A_p(z)} \tag{14}$$

e aplicando o logaritmo nos dois lados de (13) chega-se a

$$2 \log A_p(z) + \log\left(1 - R^2(z)\right) = \log\left(1 - z^{-2}\right)$$
$$+ \sum_{i=1}^{p} \left(\log\left(1 - e^{jw_i} z^{-1}\right) + \log\left(1 - e^{-jw_i} z^{-1}\right) \right) \tag{15}$$

Fazendo, agora, a expansão em série de Fourier em ambos os lados de (15), obtém-se (Kim, 2000)

$$-2 \sum_{n=1}^{\infty} c_n e^{-jwn} + \sum_{n=1}^{\infty} R_n e^{-jwn} =$$
$$- \sum_{n=1}^{\infty} \frac{1}{n} \left(1 + (-1)^n\right) e^{-jwn}$$
$$- \sum_{n=1}^{\infty} \frac{1}{n} \sum_{i=1}^{p} \left(e^{jnw_i} + e^{-jnw_i}\right) e^{-jwn} \tag{16}$$

onde c_n é o n-ésimo atributo LPCC que satisfaz a relação (Ohshima, 1993)

$$\log A_p(e^{jw}) = - \sum_{n=1}^{\infty} c_n e^{-jwn} \tag{17}$$

e R_n é a transformada inversa de Fourier de $\log(1 - R^2(z))$. Pode-se mostrar que a expansão dada pela equação (16) converge para (15), segundo os critérios de convergência das Séries de Fourier (Kim, 2000). De (16) pode-se obter, com algumas manipulações matemáticas,

$$c_n = \frac{1}{2n} \left(1 + (-1)^n\right) + \frac{1}{n} \sum_{i=1}^{p} \cos\, nw_i + \frac{R_n}{2} \tag{18}$$

Observando-se a equação (18), percebe-se que ainda existe o termo $R_n/2$ que depende dos parâmetros LPC e que os demais só dependem das LSFs. Sendo assim, será desconsiderado este termo, dando origem à expressão do atributo PCC, definido por

$$\hat{c}_n = \frac{1}{2n} \left(1 + (-1)^n\right) + \frac{1}{n} \sum_{i=1}^{p} \cos\, nw_i \tag{19}$$

É razoável esperar que desprezar o fator $R_n/2$ não venha a prejudicar o desempenho, pois este fator será zero, ou assumirá valores muito pequenos, para a maioria dos casos (Kim, 2000).

3.2 *Pseudo-Cepstrum* (PCEP)

Com base na dedução matemática dos atributos PCC, se torna bastante trivial a obtenção dos PCEP. Esses atributos são obtidos a partir dos PCC, eliminando-se o termo $\frac{1}{2n}\left(1 + (-1)^n\right)$ que não depende da voz, ou seja, não

depende dos parâmetros LSF. A expressão do n-ésimo PCEP é dada por

$$\hat{d}_n = \frac{1}{n}\sum_{i=1}^{p} \cos\, nw_i \qquad (20)$$

É razoável esperar um bom desempenho espectral dos PCEP, pois os mesmos fornecem uma envoltória espectral bastante parecida com a do *cepstrum* obtido diretamente de voz (Choi, 2000). O PCEP possui a vantagem de apresentar ainda uma carga computacional mais baixa do que o atributo PCC obtido anteriormente.

3.3 *Mel-Frequency* PCC (MPCC)

Para obter os atributo MPCC a partir dos PCC basta manipular as LSFs a serem utilizadas em (19), onde w_i é substituído por w_i^m, definido pela transformação

$$w_i^m = w_i + 2\tan^{-1}\left(\frac{0,45\, sin\, w_i}{1 - 0,45\, \cos\, w_i}\right) \qquad (21)$$

Essa equação consiste em uma forma de se transformar os eixos de freqüência de um determinado conjunto de parâmetros nos eixos de freqüência da escala mel (Gurgen, 1990). Com esta alteração de eixo obtém-se os atributos MPCC, dados pela expressão

$$\hat{c}_n^m = \frac{1}{2n}\left(1 + (-1)^n\right) + \frac{1}{n}\sum_{i=1}^{p} \cos\, nw_i^m \qquad (22)$$

onde $\hat{c}_n^m$ é o n-ésimo MPCC.

3.4 *Mel-Frequency* PCEP (MPCEP)

Para se chegar aos atributos MPCEP, basta repetir o procedimento descrito para os MPCC, o que resulta na seguinte expressão

$$\hat{d}_n^m = \frac{1}{n}\sum_{i=1}^{p} \cos\, nw_i^m \qquad (23)$$

onde $\hat{d}_n^m$ é o n-ésimo MPCEP.

4 RESULTADOS EXPERIMENTAIS

As simulações realizadas neste trabalho têm como uma de suas finalidades principal determinar quais atributos

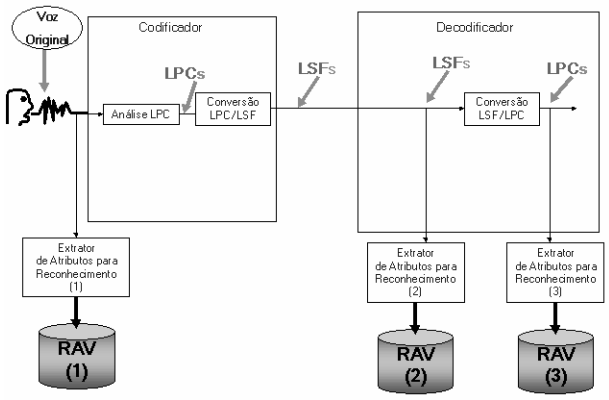

Figura 2: Sistema para análise sem quantização.

de reconhecimento possuem melhor compromisso entre desempenho de reconhecimento e carga computacional, quando se visa aplicá-los em sistemas de reconhecimento de voz distribuídos. A Figura 2 ilustra os parâmetros e sistemas que estarão sendo analisados nesta seção. Note-se que o efeito de quantização não está sendo considerado neste trabalho.

Conforme mostrado na Figura 2 estarão sendo examinados os seguintes extratores de atributos:

- Extrator de atributos (1) – fornece os atributos MFCC (*Mel-Frequency Cepestral Coefficients*) (Young, 2002 e Davies, 1980) a partir de voz original em 10 ms e em 20 ms
- Extrator de atributos (2) – fornece os atributos PCC, PCEP, MPCC e MPCEP a partir das LSFs em 10 ms e em 20 ms
- Extrator de atributos (3) – fornece os atributos LPCC e MLPCC a partir dos parâmetros LPC em 10 ms e em 20 ms

Cabe ressaltar que o MFCC é gerado a partir de voz original e só está sendo obtido para se ter uma referência de desempenho de reconhecimento para os outros atributos. O MFCC é usualmente empregado em sistemas de reconhecimento de voz que não operam em redes de comunicações. Note-se que este atributo não poderá ser utilizado em redes de comunicações onde não há transmissão de informação adicional para o sistema de reconhecimento, além do que o codificador padrão já transmite.

Todos os extratores de atributos estarão gerando sempre um conjunto de 10 parâmetros com suas respectivas derivadas, totalizando 20 atributos de reconhecimento.

A derivada Δ tem como objetivo capturar as variações dinâmicas do espectro do sinal de voz (Milner, 1996, Hanson, 1990 e Milner, 2002) . A forma de se calcular os coeficientes Δ (Young, 2002) é dada por

$$\Delta c_k(n) = \frac{\sum_{m=-N_d}^{N_d} m c_k(n+m)}{\sum_{m=-N_d}^{N_d} m^2} \quad (24)$$

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 neste artigo.

Nas simulações realizadas, os quadros de voz têm duração de 25 ms e o espaçamento entre seus centros é de 10 ou 20 ms, dependendo da taxa que se deseja gerar os parâmetros LPC / LSF e, consequentemente, 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 para fornecer um bom desempenho. Já o espaçamento de 20 ms foi escolhido por ser o espaçamento encontrado em alguns codificadores de voz para redes IP e ambiente celular que poderão operar em sistemas de reconhecimento de voz distribuído. Assim, todos os atributos de reconhecimento do sistema da Figura 2 serão obtidos utilizando esses dois espaçamentos entre centros de quadro.

O sistema de RAV aqui considerado é um reconhecedor de palavras isoladas independente de locutor, onde a 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. Uma distribuição de 70% de treinamento e 30% de teste da base de locuções estará sendo utilizada.

Os sistemas de reconhecimento utilizam HMMs de cinco estados com mistura de três gaussianas por estado. Todos foram implementados com o uso do HTK (*HMM Toolkit*) - (Young, 2002).

Na Tabela 1 são apresentados os resultados dos testes de reconhecimento quando os atributos são extraídos a cada 10 ms e 20 ms. Pode-se observar, como esperado, que a geração dos atributos a cada 20 ms fornece um desempenho bem inferior quando comparado com o obtido com atributos extraídos a cada 10 ms. Verifica-se, também, dessa tabela , que os atributos na escala mel (MLPCC, MPCEP e MPCC) sempre fornecem melhor desempenho que os atributos na escala de freqüência real (LPCC, PCEP e PCC). Observa-se, ainda, que os atributos de reconhecimento de voz para ambiente distribuído (MLPCC, MPCEP e MPCC), possuem resultados similares e são os que mais se aproximam do MFCC obtido de voz original. È importante ressaltar que, obviamente, o MFCC nunca poderá ser obtido o decodificador a partir de voz original.

Tabela 1: RESULTADOS DOS TESTES DE RECONHECIMENTO (PORCENTAGEM DE ACERTO)

	Atributos obtidos a cada 10 ms	Atributos obtidos a cada 20 ms
LPCC	95,80%	90,80%
PCC	94,60%	90,20%
PCEP	95,00%	90,40%
MLPCC	98,30%	93,80%
MPCC	97,50%	93,10%
MPCEP	98,20%	93,70%
MFCC	99,40%	95,00%

É importante lembrar que os atributos MPCEP e MPCC no decodificador completo são obtidos diretamente das LSFs, 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 que fornecem serviço de reconhecimento e que não tenham como finalidade reconstruir a voz. Pode-se observar na Tabela 1 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 0,8%.

Uma observação também interessante que se pode tirar da Tabela 1 é que os atributos MPCEP sempre possuem desempenho melhor do que os MPCC, apesar dos MPCEP representarem uma aproximação mais grosseira que os MPCC para os atributos MLPCC.

Se forem apreciados, agora, apenas os resultados de MPCEP e MLPCC da Tabela 1 pode-se observar que a diferença de desempenho é de apenas 0,1%. Esse resultado é bastante interessante quando considerado em conjunto com a complexidade computacional, pois o MPCEP representa uma economia de processamento bastante grande.

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.

Neste sentido, buscar-se-á determinar qual o melhor domínio para se aplicar a interpolação linear dos atributos MLPCC, MPCC e MPCEP apresentados na Tabela 1. 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.

Para efetuar a interpolação linear dos atributos de 20 ms para 10 ms será utilizado um fator de interpolação de valor 2, o que significa multiplicar a taxa de obtenção dos atributos por 2.

A Tabela 2 mostra os resultados obtidos para a interpolação dos atributos MLPCC, MPCC e MPCEP nos diversos domínios, bem como os valores apresentados na Tabela 1 para os atributos escolhidos, permitindo assim uma fácil apreciação e comparação dos resultados.

Comparando-se primeiramente o desempenho do uso da interpolação linear no domínio dos próprios atributos com os atributos não interpolados na Tabela 2, 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.

Tabela 2: RESULTADOS DOS TESTES DE RECONHECIMENTO DE VOZ COM OU SEM INTERPOLAÇÃO (PORCENTAGEM DE ACERTO)

	MLPCC	MPCC	MPCEP
Atrib. a 10 ms sem interp.	98,3%	97,5%	98,2%
Atrib. a 10 ms com interp. de LSF	96,0%	95,7%	96,0%
Atrib. a 10 ms com interp. de LPC	93,8%	não	não
Atrib. a 10 ms com interp. dos próprios parâm.	93,9%	93,8%	94,4%
Atrib. a 20 ms sem interp.	93,8%	93,1%	93,7%

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 Tabela 2, 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 (2^a linha da Tabela 2) e dos atributos obtidos a cada 10ms sem interpolação (1^a linha da Tabela 2), verifica-se que ainda existe uma boa margem para melhoria de desempenho.

5 CONCLUSÕES

Neste artigo foi avaliado o desempenho de diversos atributos para reconhecimento automático de voz obtidos de parâmetros LSF e parâmetros LPC, onde se verificou que o atributo MPCEP, obtido a partir de LSF, é o que apresenta melhor compromisso entre desempenho de reconhecimento e carga computacional. No que se refere à carga computacional, essa afirmação resulta de dois fatos: 1^o) O MPCEP é uma aproximação mais grosseira do MPCC (pois evita o cálculo de um dos termos da equação de obtenção do atributo MPCC), 2^o) O MPCEP tem uma carga computacional mais baixa que o MLPCC, pois ele é obtido na fase mais inicial do decodificador (a partir das LSF's), evitando a conversão LSF-LPC (usada para obtenção do MLPCC) e seu procedimento de cálculo numérico a partir de LSF é ainda mais simples do que o do MLPCC a partir de LPC.

Comparando o desempenho de reconhecimento para 10 ms e 20 ms, fica claro que existe um espaço bastante grande para ser explorado de modo a melhorar o desempenho de reconhecimento para o sistema que extrai os atributos em intervalos de 20 ms. Note que há uma diferença de aproximadamente 4% no percentual de acerto de reconhecimento.

Tentando aproveitar esta diferença de percentual de acerto foi aplicada interpolação linear em diversos domínios, o que nos levou a concluir que é interessante usar a interpolação linear dos atributos para aumentar a taxa dos mesmos. A interpolação linear aumenta o desempenho do reconhecedor. Além disso, conclui-se que 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.

REFERÊNCIAS

Choi, H. S., Kim, H. K. and Lee, H. S., (2000) "Speech Recognition Using Quantized LSP Parameters and their Transformations in Digital Communication", Speech Communication, vol. 30, pp. 223-233.

Davies, S. B. and Mermelstein, P., August 1980, "Comparasion of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," IEEE Trans. ASSP, vol.28, pp.357-366.

Gurgen, F. S., Sagayama, S. and Furui, S., November 1990, "Line spectrum frequency-based distance measures for speech recognition," Proc. ICSLP, Kobe, Japan, pp.521-524.

Hanson, B. A. and Applebaum, T. H. (1990) "Robust Speaker-Independent Word Features Using Static, Dynamic and Acceleration Features," Proc. ICASSP, pp.857-860.

Kim, H. K., Choi, S. H. and H. S., Lee, March 2000, "On Approximating Line Spectral Frequencies to LPC Cepstral Coefficients," IEEE Trans. Speech and Audio Processing, vol. 8, pp. 195 – 199.

Kleijn, W. B. and Paliwal, K. K., (1995) Speech Coding and Synthesis, Amsterdam, The Netherlands: Elsevier.

Milner, B. P. (1996) "Inclusion of Temporal Information into Features for Speech Recognition," Proc. ICSLP, pp. 256-259.

Milner, B., May 2002, "A Comparison of Front-End Configurations for Robust Speech Recognition," Proc. ICASSP, Orlando, Florida pp. 797-800.

Mitra, S. K., (1998) Digital Signal Processing: A Computer-Based Approach, McGraw-Hill International Editions.

Ohshima, Y., December 1993, "Environmental Robustness in Speech Recognition using Physiologically-Motivated Signal Processing," PH. D. Thesis, Carnegie Mellon University, Pittsburgh, Pennsylvanya.

Oppenheim, A. V. and Johnson, D. H., June 1972, "Discrete Representation of Signals," Proc. IEEE, vol. 60, pp.681- 691.

Wölfel, M., McDonough, J. and Waibel, A., (2003) "Minimum Variance Distortionless Response on a Warped Frequency Scale," Eurospeech, Geneva.

Young, S., Evermann, G., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V. and Woodland, P., December 2002, The HTK Book (for HTK Version 3.2.1).