Codificador de Voz Pessoal

Raissa Bezerra Rocha*, Gláucio Bezerra Rocha e **Marcelo Sampaio de Alencar
*Aluna de Mestrado do Programa de Pós-Graduação em Engenharia Elétrica (UFCG/COPELE)
**Professor do Departamento de Engenharia Elétrica (UFCG)

***Resumo*— Este artigo apresenta o desenvolvimento de um sistema de transmissão que usa um Codificador de Voz Pessoal. Desenvolvido para ser utilizado principalmente em sistemas móveis celulares, o codificador permite a transmissão do sinal de voz com baixa taxa de *bits*. Os sinais de voz são codificados com uma taxa de, no máximo, 150 *bits*/s. Para avaliar o desempenho do codificador, foi realizado o teste subjetivo ACR, e os avaliadores classificam a qualidade da maioria dos sinais de voz como razoável ou bom.**

***Palavras-Chave*— Codificação de voz, codificação fonética, taxa de *bits*, reconhecimento de fonemas.**

***Abstract*— This article presents the development of a transmission system that uses a Personal Voice Encoder. Developed for use mainly in mobile cellular systems, the encoder allows the transmission of voice signals with low bit rate. The voice signals are encoded with a bitrate of up to 150 bits/s. To evaluate the performance of the encoder, a subjetive ACR test was performed, and the evaluators classify the quality of most of the voice signals as either reasonable or good.**

***Keywords*— Vocoder, phonetic coding, bit rate, phoneme recognition.**

I. Introdução

Os sistemas de comunicações móveis celulares tiveram início na década de 1970, com a primeira geração de telefonia celular (1G), cujo precursor foi o sistema americano AMPS (*Advanced Mobile Phone System*).

Nas últimas décadas, os sistemas de telefonia móvel evoluíram de maneira expressiva com o objetivo de suprir a crescente demanda por novos serviços, maiores taxas de transmissão e aumento da capacidade da rede.

O serviço mais importante oferecido ao usuário de uma rede móvel celular é a transmissão da voz. Entretanto, devido à capacidade de transmissão restrita do canal, é preciso minimizar o número de *bits* que devem ser transmitidos, tornando-se necessário pesquisar o desenvolvimento de técnicas que busquem diminuir a taxa de *bits* utilizada para a representação do sinal digital, levando em consideração os níveis requiridos de qualidade do sinal, complexidade de implementação e retardo de comunicação.

A compressão de sinal tem o objetivo de reduzir o número de *bits* necessário para representar adequadamente os sinais de voz, imagem, vídeo ou áudio. Desta forma, em sistemas cuja capacidade de armazenamento e largura de banda são limitados, como no caso de sistemas de telefonia móvel, a compressão de sinais torna-se necessária, uma vez que define o número de *bits* que representa cada segundo de fala do sinal a ser transmitido, parâmetro fundamental em aplicações envolvendo transmissão e armazenamento de sinais.

Os algoritmos de codificação de voz podem ser divididos em duas categorias principais: codificadores de forma de onda, que são caracterizados por seguirem o sinal amostra a amostra, ou utilizando um vetor de amostras (quantização vetorial) e os *vocoders*, que descrevem o sinal de fala de um modo paramétrico, conseguindo uma diminuição na taxa de *bits*, mas também na qualidade do sinal sintetizado. Há também a codificação híbrida, que combina a qualidade dos codificadores de forma de onda com a eficiência dos codificadores paramétricos [2].

Entretanto, existem estratégias para codificação da voz a baixa taxa de *bits*, como os *vocoders* segmentais, caracterizados por particionar o sinal de voz em segmentos, como sílabas, fonemas, difones, entre outros, por meio de técnicas de reconhecimento de fala e são denominados de codificadores fonéticos. Para alcançar uma baixa taxa de transmissão, esse codificador não realiza a codificação do sinal de voz propriamente dito, mas apenas dos parâmetros que caracterizam cada segmento, com os índices e informações como energia, duração e frequência fundamental dos segmentos, que são as características prosódicas do sinal de voz.

Como métodos de codificação do sinal de fala utilizados em sitemas celulares, podem ser citados: o codificador RPE-LTP, utilizado no padrão GSM, que possui taxa de transmissão de 13 k*bits*/s e MOS (*Mean Opinion Score*) de 3,8, os codificadores VSELP/ACELP que proporciona taxa de 7,95 k*bits*/s e MOS de 3,8, o QCELP com taxa de até 9,6 k*bits*/s e MOS de 3,45 padronizado para telefonia celular digital CDMA e o codificador AMR-WB, com taxa de transmissão de até 23,85 k*bits*/s e MOS 4,14 utilizado na tecnologia WCDMA [22], [20].

O Codificador de Voz Pessoal (CVP) é um caso particular dos *vocoders* segmentais, em que os segmentos são fonéticos. Ele tem como característica a codificação do sinal de voz a uma baixa taxa de *bits* com o objetivo de ser utilizado principalmente em comunicações móveis celulares.

Além desta seção introdutória, este artigo está dividido em mais quatro seções. A Seção II descreve os fundamentos do codificador proposto, bem como as epatas realizadas em seu desenvolvimento. A Seção III apresenta o método de avaliação subjetiva utilizado para analisar o desempenho do codificador. A apresentação e análise dos resultados estão na Seção IV e, por fim, as conclusões e os trabalhos futuros são descritos na Seção V.

II. Descrição do codificador

Uma das características dos usuários de telefonia móvel é armazenarem vários números de telefones celulares em seus

Raissa Bezerra Rocha, Gláucio Bezerra Rocha e Marcelo Sampaio de Alencar¸ Universidade Federal de Campina Grande (UFCG), Instituto de Estudos Avançados em Comunicações (Iecom), Campina Grande - PB, Brasil, E-mails: {raissa, malencar}@iecom.org.br, glauciobr@gmail.com.

aparelhos. No entanto, a sua comunicação com outros usuários no sistema móvel é feita com maior frequência com familiares e amigos mais próximos. Nessas comunicações, o codificador proposto surge como um sistema alternativo, e os usuário podem optar por utilizá-lo para realizarem uma comunicação com baixo custo da ligação, caso as companhias telefônicas façam a cobrança por taxa de transmissão.

O uso deste codificador também possibilita um aumento na capacidade do canal de transmissão, uma vez que com uma menor taxa de transmissão, a largura de banda requerida por cada usuário é menor, sendo possível multiplexar mais usuários em um mesmo canal de comunicação.

O CVP é do tipo fonético, visto que, entre as técnicas de codificação, é a que se obtém a menor taxa de *bits*. Deste modo, tem a característica de utilizar um sistema de reconhecimento de fala com o obejtivo de segmentar o sinal de voz em segmentos fonéticos. Para alcançar uma redução na taxa de transmissão, esse codificador, em vez de codificar amostras do sinal de voz, quantiza parâmetros correspondentes a cada segmento de fala, como índices e informações sobre energia e duração.

Sabendo que no português brasileiro há apenas trinta e oito fonemas, no projeto do codificador de voz, é proposto a codificação destes segmentos por meio da atribuição de índices pré-estabelecidos, sendo assim possível codificá-los com, no máximo, seis *bits*, bem como suas informações de energia e duração. No entanto, o codificador tem a característica de fornecer uma taxa de *bits* variável, de acordo com a quantidade de fonemas pronunciada por segundo por cada usuário.

O codificador proposto tem a característica de ser pessoal. Isso parte do princípio de que inicialmente para que seja possível o seu uso, é necessário formar um banco de unidades acústicas específico para cada usuário, mediante a pronúncia de frases pré-estabelecidas. Ao receber uma soliticação de chamada, os aparelhos celulares são programados para identificar qual usuário cadastrado na agenda telefônica está solicitando uma comunicação. Ao realizar esta identificação, o receptor do codificador faz uma busca do banco de unidades daquele determinado usuário. Ao identificar o banco de unidades, o sistema realiza a síntese com os segmentos específicos daquele usuário contidos em seu banco de unidades.

A implementação do codificador está divida no desenvolvimento do emissor e receptor. O emissor é constituído por um segmentador e reconhecedor fonético que converte o sinal acústico em uma sequência de segmentos fonéticos. A informação transmitida ao receptor consiste na a sequência dos índices fonéticos além das informações de caráter prosódicos, como energia e duração de cada fonema reconhecido.

O receptor realiza a síntese por concatenação de segmentos acústicos para a formação de palavras e é construído em duas etapas. A primeira consiste na segmentação em fonemas das frases pré-selecionadas para compor o banco de unidades. A segunda etapa está relacionada à síntese por concatenação dos segmentos acústicos armazenados no banco de unidades juntamente com adaptações prosódicas de acordo com as informações recebidas do emissor do codificador.

A. *Emissor*

O emissor, ilustrado na Figura 1, é constituído das seguintes etapas: segmentação e reconhecimento de fonemas, atribuição de índices aos fonemas, estimação da energia e duração de cada fonema e codificação das informações por meio de um codificador de *Huffman*. A seguir, tem-se a descrição de cada uma delas.

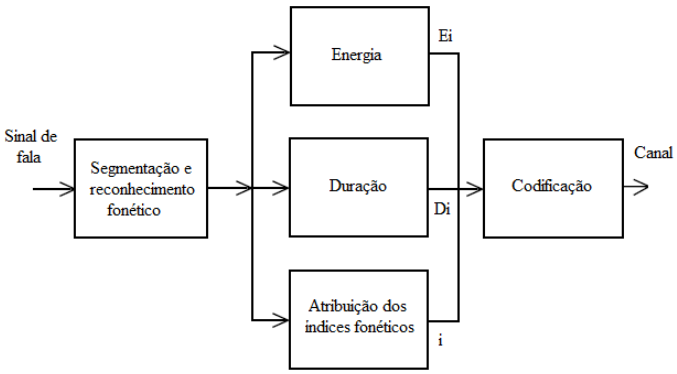

Fig. 1. Diagrama de blocos do emissor do codificador.

1) Segmentação e Reconhecimento Fonético: A segmentação e reconhecimento de fonemas é o primeiro e o mais importante passo para a implementação do codificador proposto. Seu resultado consiste no tempo inicial e final de cada fonema, necessário para se obter as informações prosódicas e sintetizar o sinal com um bom desempenho. É realizada com a utilização de técnicas de reconhecimento de sinais de fala, obtendo-se uma sequência de fonemas reconhecidos cujos índices formam uma das saídas do emissor do codificador.

Todas as etapas da segmentação e reconhecimento de fonemas descritas nessa seção são realizadas utilizando o *software* HTK (*Hidden Markov Models Toolkit*), disponível em [4].

A primeira etapa a ser realizada para a segmentação fonética consiste na pré-ênfase do sinal de voz. Isso porque uma das características do sinal de voz é de conter a maior parte da sua energia concentrada nas baixas frequências. Entretanto, as altas frequências também são responsáveis pela produção de sons, especificamente dos sons surdos. Desta forma, é necessário pré-enfatizar o sinal de voz com o objetivo de enfatizar as frequências mais altas e tornar o espectro do sinal de voz mais plano.

Inicialmente, para se obter o modelo do sistema glotal, passa-se o sinal de voz por um filtro de primeira ordem, passa-alto, L(z) do tipo:

$$L(z) = 1 - a_p z^{-1} \quad (1)$$

O parâmetro a_p é denominado fator de pré-ênfase. Neste trabalho foi utilizado $a_p = 0,95$. Assim, a pré-ênfase é realizada por meio da fórmula usual [6] [9]

$$s_p(n) = s(n) - 0,95s(n-1)$$

Após a etapa da pré-ênfase, inicia a etapa da segmentação do sinal para análise a curtos intervalos. É possível segmentar o sinal de voz em janelas ou quadros de duração definida, desde que esteja dentro do intervalo em que o sinal de voz é considerado quase estacionário, ou seja, geralmente entre

10 e 30 ms [13]. A segmentação é levada a efeito com superposição de 50% entre os quadros, visando reduzir os efeitos da descontinuidade entre segmentos. Neste trabalho, foi utilizada uma janela de *Hamming* com 25 ms e deslocamento da janela em análise de 10 ms [9].

Em seguida, busca-se extrair as informações mais relevantes do sinal de voz. Para cada segmento de fala janelado foram extraídos os coeficientes MFCC (*Mel Frequency Cepstral Coeficients*) que têm a característica de representar o sinal de voz baseado no comportamento do ouvido humano [7]. Os coeficientes Mel-Cepstrais são obtidos aplicando inicialmente, em cada janela do sinal de voz, a Transformada Rápida de Fourier (FFT – *Fast Fourier Transform*). Após a obtenção do espectro, o sinal é passado por um conjunto de filtros triangulares na escala Mel, em que é possível verificar a redução da contribuição das frequências mais elevadas, característica do ouvido humano. Em seguida é obtido o logaritmo da energia e, como última etapa para obtenção dos coeficientes MFCC, aplica-se uma IFFT [21].

Além disso, adiciona-se aos coeficientes MFCCs as suas derivadas de primeira e segunda ordem adaptando a modelagem acústica que assume que os vetores acústicos estão descorrelacionados dos seus vetores vizinhos, as características dos órgãos do aparato vocal humano que garantem que há continuidade entre sucessivas estimativas espectrais. Desta forma, é obtido, para cada janela, um vetor com 39 coeficientes.

Depois de serem realizadas as etapas descritas, que consistem no processamento do sinal de voz, dá-se início ao desenvolvimento do modelo acústico.

O modelo acústico tem o objetivo de construir, por meio das características extraídas do sinal de voz, um modelo matemático que represente cada tipo de segmento fundamental da fala, que pode ser em nível de palavras, de sentenças, ou mesmo nível fonético, como no caso deste trabalho.

O codificador utiliza modelos de HMM (*Hidden Markov Models*) [1] para dividir o sinal em segmentos fonéticos. Como o trabalho busca reconhecer cada fonema da fala contínua, são criados modelos acústicos para cada fonema da língua, buscando a partir do vetor que representa o sinal sonoro, inferir qual sequência de fonemas gera aquele vetor.

Um HMM consiste em um modelo estatístico baseado na teoria dos processos de Markov, utilizado para modelar processos estocásticos, diferenciando-se pelo fato dos seus estados não serem conhecidos, mas apenas o sinal emitido em cada um dos estados. Deste modo, é definido como um par de processos estocáticos (X,Y), em que X representa uma cadeia de Markov de primeira ordem e não é diretamente observável, enquanto Y é uma sequência de variáveis aleatórias que assumem valores no espaço de parâmetros acústicos (observações).

Assim, um HMM é caracterizado por um conjunto de M estados conectados por transições. A cada instante de tempo t existe uma mudança de estado do sistema para estados diferentes ou para o mesmo estado, e um símbolo é emitido com uma determinada densidade de probabilidade de saída. A sequência de símbolos emitidos é chamada de sequência de observações, que representa a saída do HMM.

O treinamento dos HMMs consiste em ajustar os parâmetros do modelo para satisfazer algum critério de otimização. Ou seja, dado um modelo λ e uma sequência de observações $O = \{O_1, O_1, ..., O_T\}$, ajustar os parâmetros do modelo $\{A, B, \pi\}$ de modo a representar com maior eficiência o sinal que está sendo modelado maximizando $P\{O|\lambda\}$.

O método mais conhecido e utilizado para o treinamento dos HMMs é o algoritmo de *Baum-Welch*. Este método consiste em um conjunto de equações recursivas, empregando o critério da maximização da verossimilhança, em que o processo de treinamento é repetido enquanto a verossimilhança na interação atual é maior do que a verossimilhança da iteração anterior. Para definir o conjunto de equações de re-estimação dos parâmetros do modelo por meio do algoritmo de *Baum-Welch* é necessário definir dois outros algoritmos, *forward* e *backward*.

O *corpus* utilizado para o treinamento dos HMMs foi obtido de [3], composto por 500 frases e 25 locutores com 20 frases cada, sendo 18 homens e 7 mulheres.

O protótipo definido para os modelos HMM consiste em cinco estados, dos quais três estados emissores de símbolos com uma determinada densidade de probabilidade de saída e dois não emissores, utilizando a estrutura *left-right*, como ilustrado na Figura 2. O sistema construído utiliza 38 fonemas, além do modelo que representa o silêncio e um que representa a pausa entre palavras.

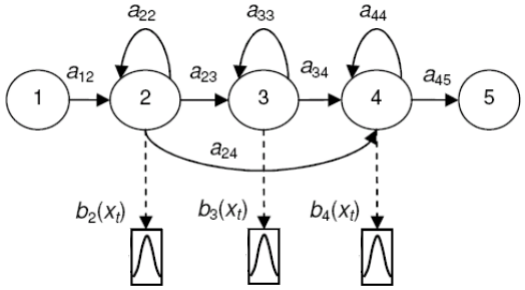

Fig. 2. Modelo de fonema baseado em HMM [18].

Um modelo acústico que utiliza um HMM por fonema (monofone) supõe que um fonema pode ser seguido por qualquer outro. Entretando, uma das características do trato vocal é que seus articuladores não se movem de uma posição para outra imediatamente na maioria das transições de fonemas. Diante disso, para modelar a fala contínua é importante considerar os efeitos contextuais causados pelas diferentes maneiras que alguns fonemas podem ser pronunciados. Neste contexto, o ideal é modelar e treinar cada um dos diferentes contextos de um mesmo fone com um HMM diferente de forma a obter uma boa discriminação entre eles. Assim, em vez de usar modelos de monofones, o ideal é usar no mínimo modelos de trifones.

A migração da utilização dos modelos de monofones para trifones provoca um grande aumento no número de modelos. Para solucionar o problema de falta de dados para o treinamento dos HMMs, é comum o uso da união de misturas, ou seja, compartilhar os componentes das misturas de gaussianas entre os estados dos HMMs. Isso porque muitos modelos de trifones possuem características acústicas semelhantes, sendo possível o compartilhamento das distribuições de probabilidade em seus estados. Neste trabalho, a união dos estados foi realizada por meio da construção de uma árvore binária de decisão para cada fone [12], [11].

Para que seja possível o reconhecimento de fonemas utilizando as etapas de extração de parâmetros e modelo acústico, é necessário que o codificador calcule a verossimilhança de qualquer sequência de vetores X, dada uma sequência de fonemas W, ou seja, maximizando $P\{X|W\}$. Neste trabalho, a decodificação foi realizada com o algoritmo de Viterbi. A Figura 3 ilustra o processo utilizado na segmentação e reconhecimento de fonemas.

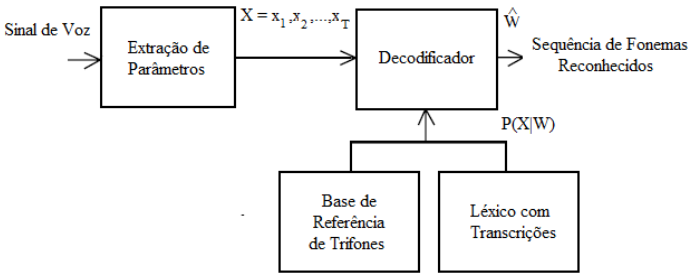

Fig. 3. Diagrama em blocos de um sistema de reconhecimento de fonemas baseado em modelos estatísticos.

O *corpus* utilizado no reconhecimento é composto de 200 frases, pronunciadas por 10 oradores diferentes, sendo 7 do sexo masculino e 3 do sexo feminino.

Para a análise de resultados, utilizou-se a ferramenta do HTK, que realiza um alinhamento entre as sequências fornecidas pelo decodificador e os textos transcritos das frases, e gera uma taxa de erro de fonemas, conforme a WER (*Word Error Rate*), definida em [13], [14],

$$\text{WER} = \frac{S + I + D}{N} \tag{2}$$

em que, N é o número total de palavras na sequência de teste e S, I e D são, respectivamente, o número total de erros por substituição (*substitution*), inserção (*insertion*) e supressão (*deletion*) na sequência reconhecida.

2) Atribuição de Índices: Da etapa de segmentação e reconhecimento de fonemas obtém-se como saída uma sequência de segmentos fonéticos com seus respectivos tempos iniciais e finais.

Para cada fonema reconhecido é atribuído um índice pré-estabelecido, totalizando quarenta diferentes índices, sendo trinta e oito índices referentes aos fonemas do português brasileiro e dois índices utilizados para referenciar o silêncio e a pausa entre palavras.

3) Estimação da Energia: A energia dos sinal de voz está concentrada na região de frequências mais baixas do espectro e, para sinais quem possuem valor médio nulo, como é o caso dos sinais de fala, a energia pode ser definada como a média do quadrado dos valores das amostras. Deste modo, a energia para cada fonema pode ser obtida por

$$E = \frac{1}{Na}\sum_{n=1}^{Na} x^2(n). \tag{3}$$

em que, x representa as amostras do sinal de voz e *Na* a quantidade de amostras em cada fonema.

4) Estimação da Duração: Estimar a duração de cada fonema é fundamental para um bom desempenho do codificador. A duração é obtida na etapa do reconhecimento de fonema, que fornece o áudio segmentado com o tempo inicial e final de cada fonema.

B. Codificação de Huffman

As informações obtidas do emissor são codificadas com o código de *Huffman*, escolhido por apresentar melhor desempenho em aplicações estatísticas [15].

O método de *Huffman* codifica a partir da ordenação decrescente das frequências, construindo uma árvore estritamente binária (Árvore de *Huffman*), base para a codificação e decodificação.

A árvore binária é construída recursivamente a partir da junção dos dois símbolos de menor probabilidade, que são então somados em símbolos auxiliares e estes recolocados no conjunto de símbolos. O processo termina quando todos os símbolos foram unidos em símbolos auxiliares, com a probabilidade final unitária, formando uma árvore binária. A árvore é então percorrida, atribuindo-se valores binários de 1 ou 0 para cada aresta, e os códigos são gerados a partir desse percurso.

C. Receptor

O receptor do CVP tem a função de converter a sequência de índices fonéticos em um sinal acústico. Para isto, realiza uma síntese por concatenção que gera o sinal de fala a partir da justaposição de segmentos fonéticos pré-gravadados. Esses segmentos são obtidos por meio da segmentação de frases foneticamente balanceadas obtidas de [19].

Inicialmente o receptor realiza a busca dos fonemas armazenados por meio da sequência dos índices dos fonemas reconhecidos. Em seguida, é feita uma adaptação em cada fonema a partir das novas informações obtidas do emissor, em que é realizado um ajuste da energia por meio de um relação da energia média do segmento armazenado com a energia recebida, e duração realizando uma comparação da quantidade de amostras recebidas com a quantidade de amostras presentes no fonema armazenado.

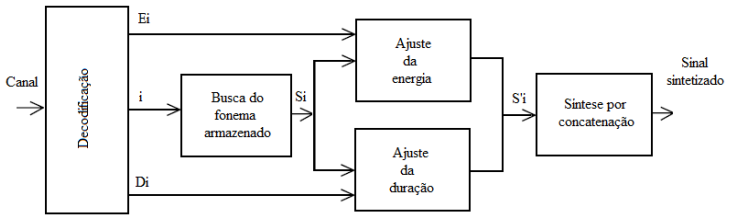

Fig. 4. Diagrama de blocos do receptor do codificador.

III. AVALIAÇÃO DO CODIFICADOR

Para avaliar o desempenho do codificador proposto, foi realizado o teste subjetivo ACR (*Absolute Category Rating*), da recomendação P.800 do ITU-T [16].

O método de avaliação subjetiva ACR consiste em uma metodologia de estímulo único, ou seja, os áudios processados pelo codificador são apresentados um por vez e, após cada apresentação, o avaliador classifica subjetivamente a qualidade do sinal processado segundo uma escala pré-determinada, mostrada na Tabela I. A pontuação final para os áudios consiste na média aritmética da pontuação obtida para cada áudio, ou seja, a MOS.

TABELA I
ESCALA DE OPNIÃO USADA NO TESTE ACR.

Qualidade da Fala	Pontuação
Excelente	5
Boa	4
Razoável	3
Pobre	2
Ruim	1

IV. RESULTADOS

Para avaliar o desempenho do codificador, foram selecionados dez áudios distintos. Cada áudio tem menos de três segundos, obtendo em média nove fonemas por segundo no conjunto dos áudios.

A taxa de *bits* resultante depende da quantidade de fonemas que cada áudio possui. Deste modo, as informações dos índices dos segmentos fonéticos, duração e energia de cada fonema são codificados com três ou quatro *bits*. A duração média de cada fonema foi de 80 ms.

O teste ACR foi realizado em um ambiente silencioso, de forma individual, com um total de doze avaliadores, não especializados na área e de idade e formação acadêmica diversas. A Tabela II mostra os resultados obtidos nos testes subjetivos e a taxa de *bits* para cada áudio.

Os resultados indicam que cinco dos áudios obtiveram notas acima de três, sendo considerados de qualidade razoável a boa. Por outro lado, três dos áudios alcançaram notas entre dois e três, resultando em áudios de qualidade pobre. Os demais áudios obtiveram notas abaixo de dois, sendo classificados como áudios de qualidade ruim.

A etapa do reconhecimento de fonemas obteve uma WER $= 20\%$.

TABELA II
RESULTADOS DOS TESTES SUBJETIVOS E TAXA DE *bits*.

Áudios	Pontuação	Taxa de *bits*(*bits*/s)
Áudio 1	3,4	112,5
Áudio 2	2,2	112,5
Áudio 3	1,5	150,0
Áudio 4	3,1	112,5
Áudio 5	2,1	150,0
Áudio 6	1,2	150,0
Áudio 7	2,0	112,5
Áudio 8	3,3	150,0
Áudio 9	3,2	150,0
Áudio 10	3,5	150,0

V. CONCLUSÕES E TRABALHOS FUTUROS

Este artigo descreve o desenvolvimento de um Codificador de Voz Pessoal. Baseado na codificação fonética, tem como principal característica a transmissão do sinal de voz com uma baixa taxa de *bits*.

Os resultados mostram que o CVP permitiu a codificação do sinal de voz com uma taxa de *bits* de, no máximo, 150 *bits*/s, sendo a maior parte dos áudios utilizados nos testes subjetivos classificados como de razoável a bom. Os demais apresentaram qualidade inferior. A qualidade dos áudios foi afetada pela alta taxa de erros de 20% obtidos no reconhecimento de fonemas.

Com o objetivo de dar continuidade ao seu desenvolvimento e torná-lo o mais eficiente possível, pretende-se como desenvolvimento futuro o aprimoramento da técnica de reconhecimento de fonemas, por meio de um melhor treinamento dos HMMs com um maior banco de voz e da inclusão de um número maior de misturas gaussianas em cada modelo HMM obtido, além de verificar o reconhecimento com a extração de outros atributos do sinal de voz, como PNCC. Será realizada a análise de diferentes técnicas de reconhecimento além da HMM, como a técnica *Type-2 Fuzzy Hidden Markov Models*, que promete melhor desempenho que a técnica HMM.

Além disso, é necessário verificar o desempenho do codificador em situações de emoções na fala e ambientes ruidosos.

AGRADECIMENTOS

Os autores agradecem o apoio da Universidade Federal de Campina Grande (UFCG), do Instituto de Estudos Avançados em Comunicações (Iecom) e do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

REFERÊNCIAS

[1] L. E. Baum and T. Petrie, *Statistical inference for probabilistic functions of finite state Markov chains*. Ann. Math. Stat, 1966.
[2] M. S. de Alencar, Telefonia Celular Digital, Editora Érica, 2007.
[3] http://www.laps.ufpa.br/falabrasil/, Visitado em 23 de outubro de 2011.
[4] http://htk.eng.cam.ac.uk/, Visitado em 9 de novembro de 2011.
[5] S. e. Young, *The HTK Book*, Microsoft Corporation, 2000.
[6] L. R. Rabiner and R. W. Schafer, *Discrete-time Processing of Speech Signals*, Prentice Hall, New Jersey, 1978.
[7] S. Davis and P. Merlmestein, *Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences.*, pp. 357-366, August, 1980.
[8] L. F. M. P. Coelho, Etiquetagem Automática de Sinais de Fala Segmentação e Classificação Fonética, Faculdade de Engenharia da Universidade do Porto, Dissertação de Mestrado, Porto, Portugal, Fevereiro, 2005.
[9] J. M. Fechine, Reconhecimento Automático de Identidade Vocal Utilizando Modelagem Híbrida: Paramétrica e Estatística, Universidade Federal da Paraíba, Tese de Doutorado, 1999.
[10] C. E. de M. Ribeiro, Codificação de Fala Baseada em Segmentos Classificados Foneticamente, Universidade Técnica de Lisboa, Tese de Doutorado, 1999.
[11] P. Woodland and S. Young, *The HTK tied-state continuous speech recognizer*, Proc. Eurospeech'93, 1993.
[12] M. Hwang and X. Huang, *Shared distribution hidden markov models for speech recognition*, IEEE Trans Speech and Audio Processing, 1993.
[13] L. R. Rabiner and B. Juang, *Fundamentals on Speech Recognition*, New Jersey, Prentice Hall, 1996.
[14] J. R. Bellegarda and D. Nahamoo, *Tied Mixture Continuous Parameter Modeling for Speech Recognition*, IEEE Transactions on Acoustics Speech and Signal Processing, vol. 38, No. I, December, 1990.
[15] G. C. da Silva e P. E. D. Pinto, Análise comparativa de métodos de compactação de dados sem perda, Universidade Estadual do Rio de Janeiro, Rio de Janeiro, Brasil.
[16] ITU-T, *ITU-T Recommendation P.800, Methods for Objective and Subjective Assessment of Quality*, August, 1996.
[17] M. S. de Alencar, Telefonia Digital, Editora Érica, 2011.
[18] R. T. Tevah, Implementação de um Sistema de Reconhecimento de Fala Contínua Com Amplo Vocabulário Para o Português Brasileiro, Dissertação de mestrado, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil, Junho 2006.
[19] A. Alcaim and J. A. Solewicz and J. A. de Moraes, Frequência de Ocorrência dos Fones e Listas de Frases Foneticamente Balanceadas no Português Falado no Rio de Janeiro, Revista da Sociedade Brasileira de Telecornunicacoes, p. 23-41, vol. 7, No. 1, Dezembro, 1992.
[20] J. L. A. Carvalho e D. Dias, Técnicas de Codificação de Voz Aplicadas em Sistemas Móveis Celulares.
[21] D. D. C. da Silva, Reconhecimento de Fala Contínua para o Português Brasileiro em Sistemas Embarcados, Universidade Federal de Campina Grande, Tese de Doutorado, Dezembro, 2011.
[22] M. de S. Freitas, A Qualidade da Voz em Sistemas de Telecomunicações, Universidade Federal Fluminense, Dissertação de Mestrado, 2009.