Efeitos da Codificação MP3 em Sistemas de Reconhecimento Automático de Locutor via GMM

Frederico Q. D'Almeida, Francisco A. O. Nascimento, Pedro A. Berger, Lúcio M. da Silva

***Resumo* – Os sistemas de Reconhecimento Automático de Locutor (RAL) são de grande importância em aplicações forenses e de investigação. Nessas aplicações, um dos desafios é a variabilidade do formato de gravação do áudio que alimenta o sistema RAL. Este artigo apresenta um estudo abrangente sobre os efeitos da codificação MP3 no desempenho de sistemas de RAL utilizando modelos de mistura de gaussianas (G*aussian Mixture Models – GMM*). São apresentados resultados de diversas simulações, com várias taxas de codificação (kb/s) e para várias freqüências de amostragem, demonstrando que, mesmo com taxas de codificação reduzidas e baixas freqüências de amostragem, é possível obter índices de acerto próximos a 100%, em grupo de 30 locutores.**

***Palavras-Chave* – Reconhecimento automático de locutor, mistura de modelos de gaussianas, codificação MP3.**

***Abstract* – Automatic Speaker Recognition (ASR) systems are very useful in forensics and in investigative applications. One of the greatest challenges for these applications is the audio codification variability. This paper presents an extensive study on the effects of MP3 coding over Gaussian Mixture Models (GMM) ASR systems. We present results of several simulations using different bit rates and sampling frequencies ratios showing that, even on low bit rate coding and low sampling frequency conditions, it's possible to achieve correct recognitions close to 100%, for a 30 speakers universe.**

***Keywords* – Automatic speaker recognition, Gaussian mixture models, MP3 codification.**

I. INTRODUÇÃO

Os modernos sistemas de reconhecimento automático de locutor (*automatic speaker recognition* – ASR) baseados em modelos de misturas de gaussianas (*gaussian mixture models* – GMM) e utilizando parâmetros mel-cepstrais têm se mostrado bastante eficientes na tarefa de identificar o autor de determinado trecho de voz [1,6]. Contudo, um fator limitante do desempenho desses sistemas é a qualidade do material disponível para comparação.

Em particular para aplicações forenses e investigativas, em decorrência da popularização dos gravadores digitais portáteis de áudio, os trechos de áudio analisados muitas vezes estão codificados no formato MP3 com baixas taxas de bits por segundo. Dessa forma, para permitir a aplicação das técnicas de ASR a situações reais da perícia criminal, é necessário validar seu desempenho nessas situações.

Recentemente, tem se iniciado o estudo do efeito da codificação MP3 no cálculo dos parâmetros mel-cepstrais [7]. Entretanto, os trabalhos nessa área são dirigidos para áudio com codificações a altas taxas (superiores a 64 kb/s), situação muito rara nas aplicações forenses.

Nesse trabalho, é avaliado o efeito da redução da taxa de codificação (bits/s) de áudio no formato MP3 no desempenho dos sistemas de ASR. Foram feitas diversas simulações, variando, além da taxa de codificação, a freqüência de amostragem do sinal e buscando identificar métodos capazes de minimizar a degradação da performance do reconhecimento causada pela codificação.

II. MODELOS DE MISTURA DE GAUSSIANAS (GMMS)

Os modelos de mistura de Gaussianas (GMMs) são uma ferramenta para a modelagem de dados particularmente útil em situações onde a distribuição dos valores das variáveis modeladas apresenta algumas concentrações (*clusters*) distintas. Esse modelo representa, essencialmente, uma distribuição formada pelo somatório ponderado de M distribuições Gaussianas e pode ser representado pela equação:

$$p(\vec{x} \mid \lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}) \tag{1}$$

em que $\vec{x}$ é um vetor de parâmetros (variáveis) D-dimensional, $b_i(\vec{x})$, i = 1, ..., M, são as M distribuições Gaussianas que compõem o modelo, também chamadas de componentes da mistura, e p_i , $i = 1, ..., M$ são os pesos de cada componente na mistura, também denominados de coeficientes de mistura.

Cada componente do GMM, $b_i(\vec{x})$, é uma Gaussiana D-dimensional de forma

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2}\left|\Sigma_i\right|^{1/2}} \exp\left\{-\frac{(\vec{x}-\vec{\mu}_i)'\Sigma_i^{-1}(\vec{x}-\vec{\mu}_i)}{2}\right\} \tag{2}$$

com valor médio $\vec{\mu}_i$ e matriz covariância Σ_i .

Os pesos das componentes da mistura são propriamente normalizados de forma que

$$\sum_{i=1}^{M} p_i = 1 \tag{3}$$

Na equação (1), λ representa a descrição completa de um GMM, incluindo suas médias, pesos e matriz de covariância.

$$\lambda = \{p_i, \vec{\mu}_i, \Sigma_i\}, i = 1, ..., M \tag{4}$$

Em sistemas de reconhecimento automático de locutor (ASR – *Automatic Speaker Recognition*), a voz de cada locutor é modelada por um GMM distinto, dando origem a um modelo λ_s, $s = 1, ..., S$; sendo S o número total de locutores modelados.

Em geral, para simplificação dos cálculos, a matriz de covariância é feita diagonal. Experimentos realizados em [4] demonstraram que essa simplificação do modelo não causa perda de desempenho na identificação.

A. Treinamento dos GMMs

Para o treinamento dos GMM é necessário, para cada locutor a ser modelado, um arquivo de áudio contendo gravações de sua voz. Esses arquivos são chamados de arquivos de treinamento e serão simbolizados por TR_s onde s indica o locutor a quem pertence a voz gravada nesse arquivo. Para cada arquivo de treinamento, são calculados diversos vetores de parâmetros, $\vec{x}_t$, para diferentes instantes de tempo, t. O conjunto desses vetores de parâmetros é representado por

$$X_s = \{\vec{x}_1, \vec{x}_2, ..., \vec{x}_T\} \tag{5}$$

Em (5), para simplificar a notação, não foi adicionado, no lado direito, o índice s indicativo do locutor.

O objetivo do treinamento do GMM é ajustar os parâmetros do modelo, λ_s, de forma a maximizar a probabilidade de ocorrência do conjunto de vetores de parâmetros X_s.

Para simplificar o problema, considera-se que cada vetor de parâmetros, $\vec{x}_t$, é independente dos demais, de forma que é possível escrever:

$$p(X_s \mid \lambda_s) = \prod_{t=1}^{T} p(\vec{x}_t \mid \lambda_s) \tag{6}$$

A equação (6) é uma função não-linear dos parâmetros do modelo λ_s, o que impede uma maximização direta. Em geral, a maximização de (6) é realizada com o algoritmo *expectation-maximization* (EM) descrito em [2].

O algoritmo EM funciona em duas etapas distintas. Na primeira etapa, chamada de etapa E (*expectation*), são calculadas as probabilidades $b_i(\vec{x}_t)$ de cada uma das componentes do GMM para cada um dos vetores de parâmetros; essas probabilidades são comumente denominadas de ativações das componentes. Na segunda etapa, chamada de etapa M (*maximization*), o modelo é atualizado da seguinte forma:

$$\begin{aligned}
\overline{p}_i &= \frac{1}{T}\sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda_s) \\
\overline{\vec{\mu}}_i &= \frac{\sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda_s)\, \vec{x}_t}{\sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda_s)} \\
\overline{\vec{\sigma}}_i^2 &= \frac{\sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda_s)\, I\, \vec{x}_t^2}{\sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda_s)} - \overline{\vec{\mu}}_i^2
\end{aligned} \tag{7}$$

sendo que a barra horizontal sobre um parâmetro indica que se trata do parâmetro já atualizado (ciclo de treinamento $n+1$). Os parâmetros sem a barra são os parâmetros antigos (ciclo de treinamento n).

Além disso, em (7), I representa a matriz identidade, σ_i representa um vetor de variâncias que compõe a diagonal de Σ_i (os demais coeficientes são zero, se for utilizada matriz diagonal) e:

$$p(i \mid \vec{x}_t, \lambda_s) = \frac{p_i b_i(\vec{x}_t)}{\sum_{k=1}^{M} p_k b_k(\vec{x}_t)} \tag{8}$$

Com essas atualizações, a cada ciclo do algoritmo EM, a probabilidade expressa em (6) é maior que no ciclo anterior.

$$p(X_s \mid \overline{\lambda}_s) \geq p(X_s \mid \lambda_s) \tag{9}$$

A finalização do treinamento é realizada, em geral, quando se atinge um número máximo de ciclos ou quando o valor de "probabilidade" calculado em (6) se estabiliza.

B. Identificação do Locutor

Para identificar o locutor a quem pertence a voz em uma arquivo de teste, TT, dentre um grupo de locutores S representados por GMMs λ_1, λ_2 , ..., λ_S é necessário determinar qual dos modelos (λ_s) apresenta a maior probabilidade *a posteriori* para o arquivo de teste dado.

$$\hat{S} = \arg\max_{1 \leq k \leq S} p(\lambda_k \mid X_{TT}) = \arg\max_{1 \leq k \leq S} \frac{p(X_{TT} \mid \lambda_k)\, p(\lambda_k)}{p(X_{TT})} \tag{10}$$

sendo X_{TT} o conjunto de parâmetros calculados a partir do arquivo de teste, TT, e utilizando, na última passagem, a regra de Bayes.

Supondo que todos os locutores são igualmente prováveis, $p(\lambda_k) = cte., k = 1, ..., S$, e observando que $p(X_{TT})$ é constante para todos os locutores, conclui-se que a identificação do locutor pode ser realizada simplesmente calculando

$$\hat{S} = \arg\max_{1 \leq k \leq S} p(X_{TT} \mid \lambda_k) \tag{11}$$

Utilizando a independência entre os vetores de parâmetros formulada em (6) e utilizando o logaritmo, o cálculo de (11) é realizado por

$$\hat{S} = \arg\max_{1 \leq k \leq S} \sum_{t=1}^{T} \log p(\overline{x}_t \mid \lambda_k) \tag{12}$$

O uso do logaritmo é um artifício para evitar problemas numéricos, pois as probabilidades envolvidas em (11) são muito baixas.

Usualmente, como os arquivos de áudio contendo as vozes dos locutores não têm exatamente a mesma duração, a equação (12) é normalizada em relação ao tempo, passando à forma:

$$\hat{S} = \arg\max_{1 \leq k \leq S} \frac{\sum_{t=1}^{T} \log p(\overline{x}_t \mid \lambda_k)}{T} \tag{13}$$

C. Uso de GMMs para Modelagem de Voz

Devido às características particulares dos sinais de voz, as GMMs são uma excelente forma para sua modelagem [3].

O trato vocal humano é um sistema de forma variável; por meio de alterações nas dimensões de seus componentes, especialmente com o reposicionamento da língua, dos lábios e do palato mole, é possível produzir sons distintos. Desse modo são articulados os diferentes fonemas (classes de sons vocais), permitindo a composição de um grande número de palavras. Cada fonema, individualmente, apresenta certa variabilidade em sua realização, dependendo, dentre outras coisas, do contexto em que se insere (articulações anteriores e posteriores) e da entonação da frase.

Pode-se imaginar que, com uma escolha adequada de parâmetros da voz, as articulações de um mesmo fonema sejam representadas por vetores ($\vec{x}$) semelhantes entre si, que se agrupem no espaço dos parâmetros formando aglomerações (*clusters*) que podem ser bem modelados por uma distribuição Gaussiana. As articulações dos diferentes fonemas devem ser bem distintas entre si, de modo a formarem aglomerações disjuntas no espaço de parâmetros.

Assim, a utilização de GMMs em sistemas de ASR se justifica porque esse tipo de técnica é capaz de modelar os diferentes fonemas (idealmente uma componente da mistura para representar cada fonema) e também capaz de representar as variações possíveis na realização de um mesmo fonema (pela variabilidade permitida pela distribuição Gaussiana).

III. Protocolo de Avaliação de Desempenho

A avaliação do efeito da codificação no desempenho de sistemas ASR foi realizada por um conjunto simulações efetuadas no banco de dados de vozes descrito em [6], cujas características fundamentais são transcritas a seguir.

A. Descrição do Banco de Dados de Áudio

O banco de dados de vozes utilizado contém 30 locutores distintos, 15 do sexo masculino e 15 do sexo feminino ($S = 30$). Cada locutor foi gravado lendo um texto pré-definido; foi utilizado o mesmo texto para todos os locutores. Posteriormente, cada uma dessas gravações foi fracionada em 21 arquivos; os pontos de início e final dos cortes dos arquivos são os mesmo para todos os locutores. Assim, foram gerados 21 arquivos para cada locutor, que serão indicados por $A_{n,s}$, onde s indica o locutor e n o trecho do arquivo originalmente gravado.

Como os pontos de corte dos arquivos foram os mesmos com relação ao conteúdo textual lido, tem-se que os arquivos $A_{k,s}$, k = cte. e $s = 1, ..., S$ contêm as vozes dos 30 locutores lendo um mesmo trecho do texto.

Essa forma de divisão dos arquivos foi escolhida por duas razões. Primeiramente, para que todos os trechos tivessem uma duração aproximadamente igual; a duração média aproximada dos trechos é de 30 segundos. Além disso, todas as simulações foram realizadas utilizando trechos de mesmo conteúdo textual; dessa forma, tentou-se minimizar a influência que diferentes conteúdos textuais dos arquivos poderiam ter na identificação do locutor.

Todas as gravações foram realizadas em ambientes acusticamente preparados, com microfones e placas de captura de áudio profissionais. Os arquivos foram adquiridos a uma taxa de amostragem de 22 kHz, quantização de 16 bits, em modo monaural.

B. Geração dos Bancos de Dados de Áudio Codificado

A partir desse banco de dados inicial, foram geradas versões dos arquivos de áudio codificadas no formato MP3 com diferentes taxas de codificação e diferentes freqüências de amostragem. A Tabela I, a seguir, lista as taxas de codificação e as respectivas freqüências de amostragem utilizadas nas simulações.

TABELA I
Taxas de codificações e freqüências de amostragem utilizadas

	Freqüência de Amostragem (kHz)		
	22	16	11
Taxa de Codificação (kbits/s)	PCM	PCM	PCM
	56	-	-
	40	-	-
	32	32	
	24	24	24
	-	20	20
	-	16	16
	-	-	8

As referências "PCM", na Tabela I e nas demais tabelas, indicam a realização de simulações com o áudio no formato *Pulse Code Modulation* linear, portanto, sem a codificação MP3. Para essas simulações, foi utilizado áudio com a freqüência de amostragem indicada e quantização de 16 bits.

Para a geração de arquivos com freqüências de amostragem (*Fa*) diferentes de 22 kHz, a mudança da *Fa* foi realizada sobre os arquivos originais (PCM) antes da codificação MP3. A mudança na *Fa* foi realizada com filtragem prévia do sinal de áudio, para eliminação de componentes acima do novo limite.

C. Pré-Processamento

Antes dos testes, todos os arquivos de áudio foram pré-processados da seguinte forma. Inicialmente, todos os arquivos foram normalizados de forma que sua amplitude de pico correspondesse a 100% do valor máximo de quantização.

Posteriormente, foram excluídos os trechos de silêncio dos arquivos. Essa eliminação foi realizada com o uso de um detector automático de silêncio baseado na medida da energia do sinal em janelas de 20 ms, com sobreposição de 15 ms (avanços de 5 ms) e limiar de silêncio definido manualmente

(um único valor para todos os arquivos) com base em testes práticos.

D. Extração de Parâmetros e Modelagem

Em todas as simulações, foram utilizados como parâmetros de modelagem os parâmetros mel-cepstrais, que reconhecidamente apresentam os melhores resultados em aplicações ASR [6]. Os parâmetros foram calculados a cada janela de 20 ms do áudio, sem sobreposição das janelas, através de bancos de filtros aplicados diretamente ao espectro de freqüências do sinal calculados nessa mesma janela. Foram extraídos, de cada janela, 12 parâmetros mel-cepstrais.

Aplicou-se ainda a normalização dos parâmetros detalhada em [4], tanto na fase de treinamento dos modelos quanto na de teste, como forma de aumentar o desempenho do sistema. Nessa normalização, são removidos os valores médios dos coeficientes mel-cepstrais.

Foram criados modelos GMM com diferentes números de componentes (1, 2, 4, 8, 16 e 32). Em todos os modelos, foi aplicada a restrição de diagonalidade à matriz de covariância, como forma de simplificar os cálculos, visto que essa restrição não ocasiona prejuízos significativos [5].

Para o treinamento dos modelos λ_s foram utilizados os arquivos $A_{21,s}$, com aproximadamente 30 segundos cada. Para os testes, foram utilizados os arquivos $A_{n,s}$, $n = 1, ..., 20$ de modo que não foram utilizados, nos testes, arquivos de áudio utilizados para o treinamento dos modelos.

Os modelos λ_s foram inicializados com valores aleatórios de parâmetros. Numa primeira fase, o treinamento desses modelos foi realizado com base no método do vizinho mais próximo (*nearest neighbor)*, a fim de obter uma primeira aproximação dos modelos ótimos. Essa primeira fase de treinamento é utilizada pois envolve um custo computacional muito inferior ao do treinamento GMM e promove uma boa aproximação do modelo final. Dessa maneira, é possível diminuir o número de ciclos de treinamento necessários na próxima etapa do treino e, com isso, diminuir o custo computacional total do processo. Essa fase foi limitada a 50 ciclos..

Concluída a primeira fase do treinamento, realizou-se uma etapa complementar em que os modelos são atualizados de acordo com o descrito na seção II.A. Essa nova fase proporciona o ajuste fino dos modelos. Foi utilizado o limite de treinamento de 25 ciclos (caso não ocorra a estabilização do modelo anteriormente). Experimentalmente, para os dados utilizados nesse trabalho, verificou-se que esse valor é suficiente para obter uma modelagem adequada se os modelos estiverem pré-treinados pelo *nearest neighbor.* Na realidade, em geral, os modelos não apresentavam alterações significativas após o décimo ciclo de treinamento; mesmo assim, foi definido um limite muito superior para garantir a boa modelagem.

E. Figuras de Mérito

Os testes de reconhecimento foram realizados utilizando modelos treinados com áudio codificado para cada uma das opções elencadas na Tabela I. Para cada um desses treinamentos, foram realizados testes com áudio codificado para todas as opções listadas na Tabela I. A realização das simulações cruzando as possíveis codificações tanto no treinamento quanto no teste objetiva verificar a ocorrência dos picos de desempenho análogos aos observados no caso de variações na relação sinal-ruído (SNR) [8].

Em princípio, o desempenho do sistema será máximo quando o treinamento dos modelos for realizado com áudio codificado de forma semelhante àqueles usados nos testes.. Entretanto, para o caso de adição de ruído, foi constatado experimentalmente que a utilização, no treinamento, de áudio levemente menos ruidoso que o utilizado nos testes leva à maximização do índice de acerto [8].

Em cada conjunto de avaliação, foram realizados 20 testes de reconhecimento (relativos aos 20 trechos distintos de áudio) com os 30 locutores do banco de dados, resultando num total de 600 análises de reconhecimento em cada conjunto de teste.

Foram ainda realizados testes com os modelos de diferentes números de componentes, a fim de observar a degradação da performance em função da simplificação dos modelos empregados.

O critério de sucesso do reconhecimento é o definido na equação (13). Entretanto, especialmente para os modelos de baixo número de componentes, foram realizadas análises relaxando um pouco essa condição e admitindo como positivas as situações em que o locutor correto fique classificado entre os α primeiros. Essas análises serviram para fundamentar a proposta de aceleração da identificação com uso de modelos GMM multi-resolução descritos em [9].

IV. RESULTADOS DE SIMULAÇÕES COMPUTACIONAIS

Os resultados obtidos foram dispostos nas Tabelas II a IV, a seguir. Essas tabelas contêm os percentuais de reconhecimentos corretos para as diferentes codificações de áudios utilizados nos treinamentos dos modelos e nos testes. Cada tabela refere-se a uma freqüência de amostragem do áudio.

De modo geral, pode-se verificar que os desempenhos máximos foram obtidos quando o modelo foi treinado com áudio de taxa de codificação igual à dos arquivos de teste. Esse fato está destacado pelos itens em negrito nas tabelas, que, na maioria dos casos, estão na diagonal.

TABELA II
Percentual de reconhecimentos corretos, 22 kHz, 32 componentes.

Taxa de Codificação (kbits/s)		TESTE				
		PCM	56	40	32	24
TREINO	PCM	**100,0**	100,0	99,0	88,7	70,2
	56	100,0	**100,0**	98,0	90,0	68,7
	40	99,5	99,8	**100,0**	99,3	71,3
	32	98,3	99,8	98,7	**100,0**	94,5
	24	84,7	89,0	85,0	97,2	**98,5**

TABELA III
Percentual de reconhecimentos corretos, 16 kHz, 32 componentes.

Taxa de Codificação (kbits/s)		TESTE				
		PCM	32	24	20	16
TREINO	PCM	**100,0**	99,7	86,8	83,0	66,0
	32	100,0	**99,8**	86,2	84,2	64,0
	24	93,0	94,0	**99,8**	92,2	70,7
	20	90,5	95,5	95,2	**99,5**	94,3
	16	77,2	73,2	82,5	93,3	**99,8**

TABELA IV
Percentual de reconhecimentos corretos, 11 kHz, 32 componentes.

Taxa de Codificação (kbits/s)		TESTE				
		PCM	20	20	16	8
TREINO	PCM	**100,0**	**100,0**	**100,0**	98,5	6,7
	24	99,7	99,7	99,7	96,5	8,3
	20	99,7	99,8	99,8	98,5	6,7
	16	98,7	98,5	97,8	**99,8**	8,0
	8	9,3	9,7	10,2	11,2	**92,7**

Esse comportamento é mais bem visualizado no gráfico da Figura 1. Nesse gráfico, cada traço corresponde ao desempenho do modelo treinado com áudio de taxa de codificação específica (em kb/s), conforme indicado na legenda; todos com freqüência de amostragem de 16 kHz. Todos os traços dessa figura referem-se a modelos de 32 componentes. Para inclusão dos dados referentes às simulações com áudio de teste no formato PCM, foi utilizado, arbitrariamente, apenas para a ilustração no gráfico, o valor 35 kb/s.

No gráfico da Figura 1, são visíveis os picos, para cada traço, nos pontos em que a codificação do áudio de treinamento se iguala à codificação do áudio de teste.

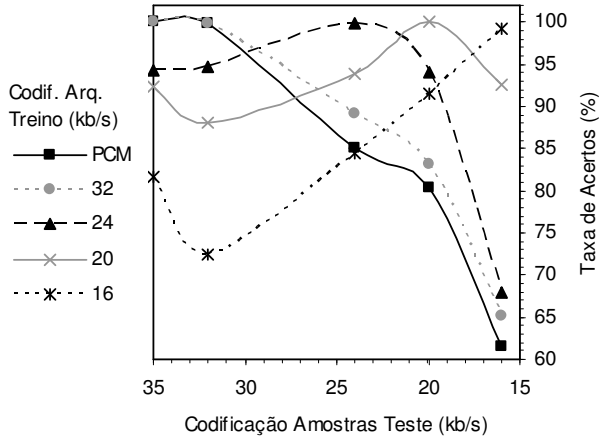

Figura 1: Taxas de acertos para modelo de 32 componentes, *Fa*=16 kHz, para arquivos de treinamento com diversas codificações.

Considerando os melhores modelos para cada caso de codificação, observa-se que as taxas de identificação corretas ficam sempre acima de 90%. O gráfico da Figura 2, a seguir, exibe as melhores taxas de identificação para uma determinada codificação de arquivos de teste. Nesse gráfico, cada traço corresponde a uma freqüência de amostragem determinada (11, 16 ou 22 kHz), conforme indicado na legenda.

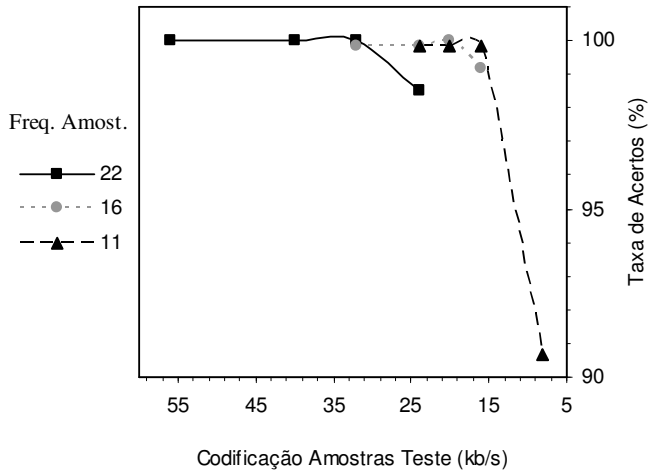

Figura 2: Melhores taxas de acertos para modelos de 32 componentes.

Para elaborar o gráfico da Figura 2, foram considerados os valores máximos das colunas das Tabelas II a IV (elementos em negrito). Dessa forma, obteve-se o valor máximo de desempenho do sistema para uma determinada condição de arquivos de teste.

A análise do gráfico da Figura 2 permite concluir que, mesmo para áudio codificado no padrão MP3 a baixas taxas de bits por segundo (8kb/p), é possível construir um sistema de reconhecimento automático de locutor com índice de acerto superior a 92%. Mais ainda, para codificações de 16 kp/s ou superiores, o índice de acerto é superior a 98%.

Com relação à variação do número de componentes dos modelos, observou-se que, entre 32 e 8 componentes, o desempenho dos sistemas apresentou poucas variações. Para números de componente ainda menores, a queda de desempenho foi mais acentuada.

O gráfico da Figura 3, a seguir, permite verificar o desempenho de modelos com diversos números de componentes para diversas taxas de codificação do áudio de teste. Nesse gráfico, todos os dados são relativos a simulações com modelos treinados com áudio PCM. Para inclusão dos dados referentes às simulações com áudio de teste no formato PCM, foi utilizado, arbitrariamente, apenas para a ilustração no gráfico, o valor de 64 kb/s.

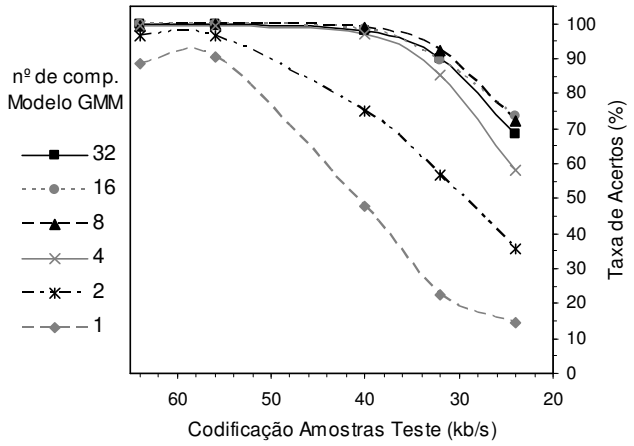

Figura 3: Taxas de acertos para modelos com diversos número de componentes (1, 2, 4, 8, 16, 32), *Fa*=22 kHz , áudio de treinamento no formato PCM.

A análise do gráfico da Figura 3 permite concluir que, para modelos com mais de 8 componentes, as taxas de identificação não apresentam ganhos significativos. Mesmo os modelos com apenas 4 componentes apresentaram desempenho muito bom, com taxas de identificação corretas próximas às dos modelos mais complexos. Apenas quando a codificação dos arquivos de teste afasta-se muito da utilizada no treinamento, o desempenho dos modelos com 4 componentes apresenta queda significativa em relação aos modelos de 8, 16 e 32 componentes.

Os modelos com 1 ou 2 componentes apresentaram desempenho significativamente inferior ao dos demais, mesmo quando os arquivos de teste estavam codificados no mesma forma dos de treino. Essa diferença acentua-se rapidamente à medida que a codificação dos arquivos de teste utilizados afasta-se da utilizada no treinamento.

Apesar de o gráfico representar uma situação particular, esse mesmo comportamento foi observado em todas as demais simulações realizadas.

V. CONCLUSÃO

Neste trabalho, foi avaliado o efeito da redução da taxa de codificação (bits/s) de áudio no formato MP3 no desempenho dos sistemas de ASR utilizando GMMs. Foram feitas diversas simulações, variando a taxa de amostragem do sinal e buscando identificar métodos capazes de minimizar a degradação da performance do reconhecimento causada pela codificação. Verificou-se que as técnicas de reconhecimento automático de locutor por modelos de mistura de gaussianas podem ser utilizadas com sucesso mesmo nos casos de arquivos codificados no formato MP3.

Os resultados das simulações evidenciaram que, para sinais de áudio no formato MP3 codificados com taxas iguais ou superiores a 16 kb/s, é possível obter percentuais de acerto na identificação do locutor semelhantes aos obtidos com áudio PCM: taxas de acerto superiores a 98% para grupos de 30 locutores.

Para áudio no formato MP3 codificado com taxas de 8 kb/s há significativa degradação no desempenho, embora ainda sejam obtidos acertos em mais de 90% dos casos.

Além dos testes com áudio codificado a diferentes taxas, também foram realizados testes com modelos de diferentes números de componentes gaussianas. Verificou-se que a utilização de modelos com mais de 8 componentes não levou a ganhos significativos no desempenho. Mesmo modelos com apenas 4 componentes apresentam percentuais de acerto no reconhecimento muito próximos aos dos modelos mais complexos, desde que os modelos sejam treinados com áudio com codificação semelhante à dos arquivos de teste.

Em qualquer situação, para a maximização do desempenho do reconhecimento, é necessário treinar os modelos com áudio codificado com a mesma taxa de kb/s dos arquivos utilizados para teste.

REFERÊNCIAS

[1] J.E. Luck, *Automatic speaker verification using cepstral measurement*, Journal Acoustic Society of America, Vol. 46, pp. 1026-1032, 1969.

[2] A. Dempster, N. Laird e D. Rubin, *Maximum likelihood from inclomplete data via the EM algorithm*, Journal Royal Statistical Society, Vol. 39, pp. 1-38, 1977.

[3] D.A.Reynolds, *A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification*, Ph. D. Thesis, Georgia Inst. of Tech, 1992.

[4] D.A. Reynolds e R.C. Rose, *Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models*, IEEE Trans. Speech and Audio Proc., Vol. 3, no. 1, pp 72-83, 1995.

[5] D.A. Reynolds, T.F. Quatieri e R.D.Dunn, *Speaker Verification Using Adapted Gaussian Mixture Models*, Digital Signal Processing, Vol. 10, nos 1-3, pp. 19-41, 2000.

[6] F.Q. D'Almeida e F.A.O Nascimento, *Comparação de Desempenho de Parâmetros da Fala em Sistemas de Reconhecimento Automático de Locutor*, Congresso Brasileiro de Automática – CBA, 2006.

[7] S. Sigurdsson, K.B. Petersen e T. Lehn-Schiøler, *Mel Frequency Cepstral Coefficients: An Evaluation of Robustness of MP3 Encoded Music*, 7th International Conference on Music Information Retrieval, 2006.

[8] F.Q. D'Almeida e F.A.O Nascimento, *Efeitos do Ruído em Sistemas de Reconhecimento Automático de Locutor via GMM*, não publicado.

[9] F.Q. D'Almeida e F.A.O Nascimento, *Reconhecimento Automático de Locutor Utilizando Modelos GMM Multi-Resolução*, não publicado.