Uso dos parâmetros VVGP em sistemas de verificação de locutor baseados em i-vectors

Cristian Jesus Silva de Souza e Lee Luan Ling

***Resumo*— Neste trabalho, a fusão dos parâmetros MFCCs (*Mel-Frequency Cepstral Coefficients*) e VVGP (*Variable Variance Gaussian Parameter*) em sistemas de verificação de locutor independentes de texto é avaliada. Dois classificadores foram utilizados neste trabalho: o primeiro baseado em Mistura de Gaussianas; o segundo utilizando i-vectors. Os resultados indicam a validade do uso dos parâmetros VVGP em ambas metodologias.**

***Palavras-Chave*— i-vectors, MFCCs, verificação de locutor.**

***Abstract*— This paper reports on a text-independent speaker verification system that combines Mel-frequency cepstral coefficients (MFCCs) with the Variable Variance Gaussian Parameter (VVGP). We utilized classifiers based on Gaussian mixture models and i-vectors. Results show that the proposed framework with VVGP feature extraction can improve recognition accuracy.**

***Keywords*— i-vectors, MFCCs, speaker verification.**

I. Introdução

Um dos focos centrais dentro da área de reconhecimento de locutor tem sido direcionado na busca de técnicas para a extração de informação útil para caracterizar uma pessoa através de sua fala, e avanços significativos foram obtidos recentemente nesse campo, impulsionados pelo trabalho de Dehak [1].

Sistemas de reconhecimento de locutor visam identificar uma pessoa através da análise de uma amostra de sua fala. Dependendo do objetivo desejado, os sistemas de reconhecimento de locutor podem ser classificados em duas categorias: identificação de locutor e verificação de locutor. No processo de identificação, um locutor desconhecido é comparado com cada locutor cadastrado no sistema, e escolhe-se o mais semelhante. Nos sistemas de verificação de locutor, o locutor fornece sua identidade e o sistema decide aceitar ou recusar o usuário, dependendo da comparação com o seu padrão armazenado.

Além disso, os sistemas de reconhecimento de locutor, dependendo do conteúdo das locuções utilizadas, podem ser classificados em dependentes de texto ou independentes de texto. Em sistemas dependentes de texto, o usuário utiliza uma frase pré-definida (senha) para ser testada. Em sistemas independentes de texto, deseja-se reconhecer um usuário a partir de qualquer amostra de fala produzida por ele. O foco deste trabalho é o desenvolvimento de sistemas de verificação de locutor independentes de texto.

Cristian Jesus Silva de Souza e Lee Luan Ling. Departamento de Comunicações- DECOM, Faculdade de Engenharia Elétrica e Computação (FEEC), Universidade Estadual de Campinas- UNICAMP, Campinas, São Paulo, Brasil, E-mails: {cristian, lee}@decom.fee.unicamp.br. Este trabalho foi financiado pela Fundação de Amparo e Pesquisa do Estado do Amazonas (FAPEAM).

O desenvolvimento de um sistema de verificação de locutor independente de texto requer, necessariamente, a identificação de parâmetros que melhor caracterizem um sinal de fala. A tradicional metodologia baseada no uso da análise espectral de curto tempo para a obtenção dos vetores de características é bem sedimentada, no entanto outras perspectivas podem ser utilizadas como complemento às tradicionais.

Em sinais de fala, grande quantidade da informação se encontra concentrada nas partes não estacionárias do sinal, como é o caso das transições (de vogais a consoantes, de vogal a vogal, entre outras), o que torna os métodos baseados no uso da análise espectral de curto tempo pouco adequados para caracterizar estes comportamentos. A teoria multifractal é capaz de caracterizar estes tipos de mudanças rápidas, chamadas singularidades e modelar esse comportamento por meio de uma representação de multi-escalas. Em [2] o uso de um novo parâmetro multifractal como vetor de características do locutor, denominado VVGP foi proposto. Aqui, avaliamos sua complementariedade com parâmetros tradicionais, especificamente com os parâmetros MFCCs [3].

II. Sistemas de Verificação de Locutor

Em verificação de locutor, o sistema deve comparar a amostra adquirida com uma referência e determinar se pertence à mesma pessoa ou não. Matematicamente, dado um modelo R e um conjunto de vetores de características $X = \{x_1, ..., x_T\}$, um grau de similaridade $s(X, R) \in \Re$ é obtido.

A. Abordagem clássica baseada em Mistura de Gaussianas

Entre as metodologias usadas, o modelo estatístico Mistura de Gaussianas GMM (*Gaussian Mixture Model*) destacou-se na literatura de reconhecimento de locutor [4]. GMM assume que os vetores de características de um locutor seguem uma densidade de probabilidade que é uma combinação de funções gaussianas multidimensionais. Nessa abordagem, as gaussianas representam as classes fonéticas que compõem o som produzido.

Em verificação de locutor o objetivo é determinar se um conjunto de vetores de características $X = \{x_1, ..., x_T\}$ pertence ou não ao possível locutor s. Assim duas hipóteses são levantadas:

- H_0: X pertence ao locutor s
- H_1: X não pertence ao locutor s

As hipóteses H_0 e H_1 precisam ser estimadas. A hipótese H_0 pode ser representada por um modelo GMM λ_s obtido da locução de cadastro do locutor s. Contudo, na prática não possuímos material suficiente de cada locutor. Assim, classes

fonéticas não encontradas no conjunto de treinamento não serão representadas.

Um solução proposta para esse problema é o uso de técnicas de adaptação. A ideia consiste em implementar um método que gere um modelo específico para um determinado locutor a partir de um modelo independente de locutor, denominado UBM (*Universal Background Model*).

O modelo UBM é um GMM treinado a partir de horas de fala de diferentes locutores (locuções de impostores) e idealmente representa toda a variabilidade fonética existente. Geralmente uma base de locutores é utilizada exclusivamente para compor o UBM, pois o modelo UBM representa a classe de impostores, ou seja, todo o espaço de vetores de características que não pertencem aos locutores cadastrados. Assim, os modelos de cada locutor são obtidos a partir do UBM em um processo de adaptação denominado estimação de maximum a posteriori (MAP) [4]. A técnica MAP é baseada na estimação de máxima verossimilhança. De posse dos dados de adaptação, a técnica MAP estima os parâmetros do novo modelo tais que a verossimilhança desse modelo, dados os dados de adaptação, seja máxima.

Além disso, o UBM λ_Ω tem a função de representar a hipótese H_1. A decisão é baseada na razão de verossimilhanças dada por

$$S(X, \lambda_s) = log \quad \frac{p(X \mid \lambda_s)}{p(X \mid \lambda_\Omega)} \begin{cases} \geq \theta & \text{aceita} \\ < \theta & \text{rejeita} \end{cases} \tag{1}$$

B. Joint Factor Analysis

A variabilidade observada nas características extraídas de um mesmo locutor (variância intra-classe) é considerado o principal problema em verificação de locutor. Nesse sentido, diversos trabalhos foram direcionados com o objetivo de compensar esse efeito. Entre eles, destacou-se a técnica denominada JFA (*Joint Factor Analysis*) [5] [6].

JFA representa amostras de um sinal de fala por um único vetor, denominado supervetor. Supervetores, como o nome sugere, são vetores simples de alta dimensão D obtidos a partir de um conjunto de vetores de características $X = \{x_1, ..., x_T\}$ de dimensão d, onde $D >> d$. Assim, vários supervetores podem ser obtidos a partir de diferentes locuções pronunciadas por um locutor. JFA considera a variabilidade entre os supervetores obtidos e modela explicitamente essa variação.

Inicialmente, assume-se que um supervetor M pertencente a um locutor pode ser decomposto na soma de dois supervetores, s (*speaker supervector*) e c (*channel supervector*) como mostrado em

$$M = s + c \tag{2}$$

sendo s e c estatisticamente independentes, ambos seguindo uma distribuição gaussiana. O supervetor s carrega informação somente da identidade do locutor, enquanto c possui informação a respeito da especifica amostra de fala da qual o supervetor foi obtido.

Assumimos que a distribuição de s é função de variáveis latentes definida por

$$s = m + Vy + Dz \tag{3}$$

O supervetor s é obtido pela combinação de duas técnicas: adaptação MAP, representada pela matriz D, e *eigenvoice*, representada pela matriz V. A variável m é um supervetor obtido do UBM.

A variabilidade de canal/sessão é representada pelo supervetor c definido por

$$c = Ux \tag{4}$$

onde x (*channel factors*) é uma variável latente, de distribuição normal e U (*eigenchannel matrix*) é uma matriz retangular de *low rank*.

Combinando as equações 3 e 4 na equação 2 obtemos a equação

$$M = m + Vy + Dz + Ux \tag{5}$$

Intuitivamente, as componentes c e s são representadas por um conjunto de fatores de baixa dimensão que operam ao longo de componentes principais. Então, dado o material de treinamento de um locutor e as matrizes U, V e D, as variáveis x, y e z são estimadas para determinada amostra de fala. Em seguida, o supervetor c é subtraído, e somente o supervetor s é usado como modelo do locutor. Os scores são obtidos usando as matrizes e as variáveis latentes.

Geralmente, as matrizes U, V e D são obtidas a partir de uma base de dados com uma grande quantidade de locutores, onde cada um tenha pronunciado diferentes locuções.

C. I-vectors

Em [1] Dehak comprovou que a matriz U também contém informação útil na discriminação de locutores. Em seu trabalho, apenas uma matriz, denominada *total variability* T é usada, onde ambas variabilidades inter e intra classe são representadas. Nessa nova abordagem, a equação 5 pode ser reescrita por

$$M = m + Tw \tag{6}$$

e pode ser interpretada como uma Análise de Componentes Principais (PCA) que projeta os vetores de características nesse novo espaço definido pela matriz T. Assim, dados o supervetor de um locutor e a matriz T, a variável w é estimada. Os componentes da variável w são denominados i-vectors.

Quando locuções de fala são representadas por i-vectors, a tarefa de um sistema de verificação de locutor é simplesmente determinar se dois i-vectors possuem a mesma informação ou não, em relação ao locutor. Um i-vector representa a amostra de um locutor cadastrado e o outro representa uma amostra de teste. Assim, se o sistema de verificação conclui que os dois i-vectors carregam a mesma informação, então ambos pertencem ao mesmo locutor. A dimensão dos i-vectors geralmente é reduzida após a aplicação da técnica LDA (*Linear Discriminant Analysis*) [7] [1] com o objetivo de maximizar a variabilidade inter-classe e minimizar a variabilidade intra-classe. O processo de decisão é baseado na metodologia denominada PLDA (*Probabilistic Linear Discriminant Analysis*) [8].

D. PLDA

A técnica PLDA originalmente foi proposta para reconhecimento de face [8] e posteriormente para verificação de locutor em [9]. Na abordagem PLDA, dado um i-vector w, duas componentes são obtidas: uma componente I que dependa somente da identidade da pessoa e não da amostra de voz em si; e em outra componente L, que difere para cada amostra de fala de um locutor, ou seja, representa a intra-variabilidade.

Assim, um i-vector $w_{s,r}$ pode ser representado por

$$w_{s,r} = m + Sx_s + Cy_{s,r} + e_{s,r} \tag{7}$$

onde r representa uma dada locução do locutor s, da qual o i-vector foi obtido.

A componente I é representada por $m + Sx_s$, enquanto a componente L é representada por $Cy_{s,r} + e_{s,r}$. A variável m representa a média de todos os i-vectors de treinamento.

Dados dois i-vectors, w_1 (de cadastro) e w_t (de teste), o $score$ é obtido pela razão de verossimilhança dada por [8]

$$score(w_1, w_t) = \frac{p(w_1, w_t|H_1)}{p(w_1|H_0)p(w_t|H_0)} \tag{8}$$

onde a hipótese H_1 indica que ambos i-vectors pertencem ao mesmo locutor, enquanto que a hipótese H_0 indica que os i-vectors pertencem a locutores diferentes.

III. Cascatas multiplicativas

A obtenção dos parâmetros VVGP baseia-se no modelo multiplicador Gaussiano de variância variável (VVGM). Nessa seção, inicialmente apresentamos conceitos relacionados às cascatas multiplicativas. Em seguida, é feita a descrição do modelo VVGM usado para a obtenção de parâmetros VVGP.

A. Cascasta multiplicativa binomial

Uma cascata é um processo em que um dado conjunto é dividido em porções sucessivamente menores obedecendo a uma regra geométrica, e ao mesmo tempo, a medida associada a este mesmo conjunto é dividida de acordo com uma outra regra.

A cascata multiplicativa binomial é o método mais simples de se obter um processo multifractal, consistindo de um procedimento iterativo no intervalo compacto $[0:1]$.

Sejam m_0 e m_1 (multiplicadores da cascata), dois números positivos cuja soma é 1. Como mostrado na Fig. 1, no estágio $n = 0$ da cascata, a medida inicial μ_0 do processo com valor aleatório entre $I_0 = [0, 1]$ é calculada. No estágio $n = 1$, a medida μ_0 distribui massa, sendo, m_0 no subintervalo $I_{00} = [0, 1/2]$ e massa igual a m_1 em $I_{01} = [1/2, 1]$. Em $n = 2$, esse processo é repetido em ambos intervalos I_{00} e I_{01}, obtendo quatro sub-intervalos I_{000}, I_{010}, I_{010} e I_{011}, gerando assim:

$$\begin{aligned} \mu[0, 1/4] &= m_0 m_0 \\ \mu[1/4, 1/2] &= m_0 m_1 \\ \mu[1/2, 3/4] &= m_1 m_0 \\ \mu[3/4, 1] &= m_1 m_1 \end{aligned} \tag{9}$$

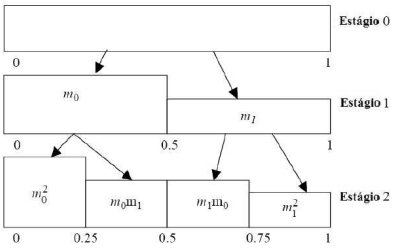

Fig. 1. Processo de construção da cascata binomial.

Este processo de divisão preserva em cada estágio a massa dos intervalos diádicos, por isso é chamado de cascata conservativa.

B. Modelo Multifractal VVGM

Na seção anterior, consideramos a cascata multiplicativa binomial com multiplicadores fixos m_0 e m_1. Ao considerar os multiplicadores da cascata como variáveis aleatórias independentes em [0;1], com densidade de probabilidade $f_R(x)$, obtém-se uma estrutura mais geral do que a determinística, em que os multiplicadores são valores fixos.

O modelo VVGM, originalmente utilizado na modelagem de intervalos de tempo de chegada de tráfego LAN em banda larga, assume que os multiplicadores possuem uma distribuição de probabilidade Gaussiana com média r=0,5 e variância variável para cada nível da cascata [10].

O algoritmo para estimação dos multiplicadores pode ser encontrado em [2] [10].

C. Extração dos Parâmetros VVGP

Os parâmetros VVGP são obtidos como descrito abaixo:

- Divisão do sinal de fala em quadros, utilizando uma janela retangular de 30ms, com deslocamentos de 15 ms.
- Uso do módulo dos valores do sinal em cada quadro, pois para obter os parâmetros VVGP o sinal a ser processado deve ser positivo.
- Para cada quadro, a média correspondente é obtida. Em seguida, esse valor é somado ao sinal para evitar sinais resultantes próximos a zero.
- Dependendo do comprimento da janela e da frequência de amostragem utilizados, quadros com diferentes quantidades de amostras são obtidos. Na obtenção dos parâmetros VVGP, é preciso limitar cada quadro ao número de amostras máximo que seja potência de 2. Por exemplo, dado um sinal de entrada com uma taxa de amostragem de 22,05 kHz, uma janela de 30 ms contém 662 amostras, porém somente serão usadas 2^N amostras, nesse caso, $2^9 = 512$ amostras.
- O processo de estimação dos multiplicadores é realizado. Em seguida, para cada estágio da cascata são obtidos seus respectivos histogramas. Dado um histograma, é calculada a variância da distribuição. Assim, para um quadro, N-2 coeficientes VVGP são determinados, dados pelas variâncias dos multiplicadores estimados.

IV. MATERIAIS E MÉTODOS

A. Base de dados

A base Ynoguti [11] é composta por sinais de fala de 71 locutores (50 homens e 21 mulheres), digitalizadas a 22,05 kHz e com 16 bits/amostra. Utilizamos a base Ynoguti para o cadastro dos locutores e para a realização dos testes.

Cada modelo dos 71 locutores de Ynoguti foi obtido a partir de 20 locuções de treinamento, totalizando 70 s de duração na média. O sistema foi testado usando 10 locuções de cada locutor, cada uma com duração entre 3 e 4 s. Cada locução de teste foi comparada com cada locutor cadastrado no sistema. Assim foram realizados 50410 experimentos, dos quais 710 eram locutores verdadeiros.

A base de dados Spoltech [12] é composta por 480 locutores, contudo neste trabalho selecionamos apenas 280. Os sinais de fala foram gravados a uma taxa de amostragem de 44,1 kHz, 16 bits/amostra, utilizando um microfone conectado a uma placa de som em um computador. Utilizamos a base Spoltech exclusivamente para o treinamento do UBM. Os arquivos da base Spoltech foram convertidos para a mesma taxa de amostragem da base Ynoguti.

B. Extração de parâmetros

Para a obtenção dos parâmetros MFCCs inicialmente aplica-se um filtro de pré-ênfase no sinal de fala, divide-se o sinal em quadros de 30 ms, com deslocamento de 15 ms. Cada quadro obtido é multiplicado por uma janela de Hamming, com o objetivo de minimizar as transições abruptas nos extremos do sinal segmentado. A análise de Fourier de curto tempo é aplicada ao sinal janelado por meio da transformada rápida de Fourier FFT (*Fast Fourier Transform*).

Em seguida, os valores dos módulos ao quadrado da FFT são agrupados em bandas críticas e ponderados por uma função triangular. A dimensionalidade da informação extraída do sinal de fala é reduzida pela aplicação do banco de filtros na escala mel. Em seguida, é calculado o logaritmo do módulo ao quadrado do espectro resultante.

E finalmente é calculada a DCT (*Discrete Cossine Transform*). Ressalta-se que neste trabalho somente os primeiros 12 coeficientes da DCT são usados.

Os parâmetros VVGP foram obtidos conforme descrito na subseção III-C. Em todos experimentos, a ordem dos vetores de características do parâmetro VVGP foi 7.

Após obter o vetor de parâmetros MFCCs e VVGP, os quadros que representam silêncio são descartados, utilizando um detector de atividade de voz (VAD) [13].

C. Experimentos

Inicialmente um sistema UBM-GMM foi implementado. O UBM foi treinado a partir de sinais de fala retirados da base Spoltech. Utilizamos 140 locutores masculinos e 140 locutores femininos, onde cada locutor possui em média um minuto de material de treinamento. Dois UBMs dependentes de gênero foram treinados separadamente utilizando 512 gaussianas. Em seguida os dois modelos UBMs foram combinados e um UBM final com 1024 gaussianas foi obtido.

Nos sistemas baseados em i-vectors, a matriz T foi obtida utilizando 300 fatores. Em seguida, para cada locução de treinamento, um i-vector projetado a partir da matriz T foi obtido. Assim, 20 i-vectors foram obtidos para cada um dos 71 locutores. Para compor o i-vector final, simplesmente calculamos a média dos i-vectors disponíveis. Em seguida, para cada locução de teste, um i-vector também foi obtido. Assim, para cada locutor, obtemos 10 i-vectors. Aplicamos a técnica LDA, reduzindo a dimensionalidade dos i-vectors de 300 para 70.

Cada i-vector de teste foi comparado com cada i-vector de cadastro. O processo de decisão é baseado no valor de *score* definido pela equação 8, utilizando a técnica PLDA.

A variabilidade intra-classe geralmente é modelada por uma base de dados independente, contudo, neste trabalho devido à disponibilidade de dados de treinamento, os modelos de intra-variabilidade (matrizes T e LDA) foram obtidos pelo próprio material de cadastro dos locutores.

D. Fusão

Nesse trabalho dois métodos de fusão foram utilizados:

1) Fusão 1: o método consiste em utilizar os parâmetros MFCC e VVGP individualmente. Em seguida, um classificador gera dois sub-scores independentes s_1 e s_2. O score final s é obtido a partir da combinação dos sub-scores utilizando a equação

$$s = w_1 s_1 + w_2 s_2 \tag{10}$$

onde w_1 e w_2 correspondem aos pesos dados aos parâmetros MFCCs e VVGP, respectivamente. Nos nossos experimentos, verificamos que o melhor resultado foi obtido ao utilizar valores de $w_1 = 0,7$ e $w_2 = 0,3$.

2) Fusão 2: o método consiste em combinar os parâmetros VVGP e MFCC em um único vetor de características. Assim, ambos parâmetros foram extraídos através dos mesmos quadros de fala.

E. Medidas de Desempenho

Para avaliar o desempenho dos sistemas, consideramos o valores de EER (*equal error rate*). EER é o ponto de operação onde a taxa de falsa aceitação (FA) é igual a taxa de falsa rejeição (FR). Quanto menor for esse valor melhor é o desempenho do sistema. Além do valor de EER, a função DCF (*Detection Cost Function*) também foi usada, definida por

$$DCF(\theta) = C_{FR} P_{ver} FR(\theta) + C_{FA} P_{imp} FA(\theta) \tag{11}$$

Os valores de C_{FR} e C_{FA} são os valores de custos associados aos erros de falsa rejeição e falsa aceitação respectivamente. Os valores de P_{ver} e P_{imp} correspondem à probabilidade de ocorrerem testes de aceitação e de rejeição e θ é o valor do limiar. Em nossas experimentos utilizamos os valores de $C_{FR} = 10$, $C_{FA} = 1$, $P_{ver} = 0,01$ e $P_{imp} = 1 - P_{ver} = 0,99$. Assim a função DCF representada na equação 11 pode ser reescrita por

$$DCF(\theta) = 0,1 FRR(\theta) + 0,99 FAR(\theta) \tag{12}$$

Além disso, apresentamos o desempenho dos sistemas operando em diferentes valores de limiares, por meio da curva DET (*Detection Error Tradeoff*). Neste trabalho, utilizamos as rotinas disponíveis em MSR Identity Toolbox [14].

V. Resultados e Discussão

Nesta seção o desempenho de sistemas de verificação de locutor baseados na fusão de parâmetros MFCCs e VVGP são apresentados.

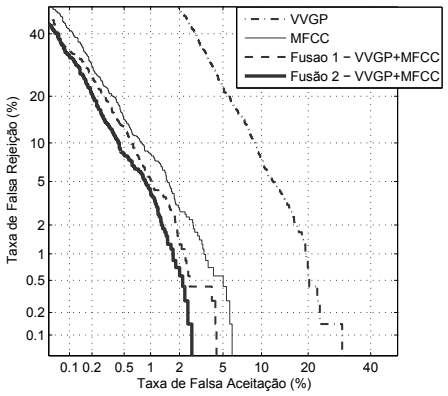

Fig. 2. Curvas DETs em sistemas de verificação de locutor UBM-GMM.

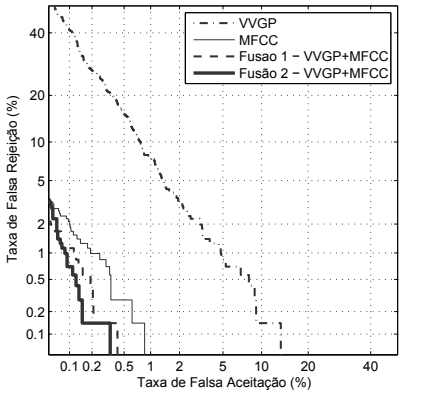

Fig. 3. Curvas DETs em sistemas de verificação de locutor baseados em i-vectors.

TABELA I
VALORES DE EER E MINDCF PARA DIFERENTES CONFIGURAÇÕES.

Método	EER (%)	MinDCF
VVGP (UBM-GMM)	9,29	6,69
MFCC (UBM-GMM)	2,39	1,72
Fusão 1- VVGP+MFCC (UBM-GMM)	1,86	1,46
Fusão 2- VVGP+MFCC (UBM-GMM)	1,51	1,28
VVGP (i-vector)	2,46	1,64
MFCC (i-vector)	0,34	0,26
Fusão 1- VVGP+MFCC (i-vector)	0,20	0,19
Fusão 2- VVGP+MFCC (i-vector)	0,14	0,16

Na Fig. 2, visualizamos o efeito nas curvas DET em um sistema UBM-GMM ao utilizar os parâmetros VVGP. Verificamos o ganho de desempenho, uma vez que os valores de EER e de MinDCF diminuíram.

Em seguida, notamos pela Fig. 3 que ao usar i-vectors, os desempenhos dos sistemas apresentaram valores de erros mínimos. Além disso, verificamos que em ambas metodologias, UBM-GMM e i-vector PLDA, o uso dos parâmetros VVGP contribui para o desempenho dos sistemas, como indicado na Tabela I.

VI. Conclusões

Neste trabalho, o objetivo principal foi avaliar o uso do parâmetro VVGP em sistemas de verificação de locutor. Assim, dois métodos de fusão foram testados; utilizando a metodologia clássica baseada em UBM-GMM, e em sistemas do atual estado da arte utilizando i-vectors. Verificamos que os dois métodos de fusão apresentaram bons desempenhos, indicando que os parâmetros MFCCs e VVGP são complementares entre si.

Contudo, outras técnicas de fusão podem ser utilizadas, além da necessidade de testar o parâmetro VVGP em situações adversas, como, por exemplo, em presença de ruído telefônico.

Referências

[1] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, e P. Ouellet, "Front-end factor analysis for speaker verification", *IEEE TASLP*, vol. 19, pp. 788-798, Maio 2011.

[2] L. L. Ling e D. C. Gonzalez, "Improving MFCC Based ASI System Performance Using Novel Multifractal Cascade Features", *In the Fifth International Conference on Intelligent Control and Information Processing (ICICIP2014)*, 2014, Dalian.

[3] S. Davis e P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", *IEEE Trans. Acoustics, Speech, and Signal Processing*, pp. 357-366, 1980.

[4] D. Reynolds, T. F. Quatieri e R. B. Dunn. "Speaker verification using adapted gaussian mixture models", *Digital Signal Processing*, vol. 10, pp. 19-41, 2000.

[5] P. Kenny. Joint factor analysis of speaker and session variability: theory and algorithms, Technical Report CRIM, 2006.

[6] P. Kenny, G. Bouliange, P. Ouellet e P. Dumouchel, Speaker and session variability in GMM-based speaker verification, *IEEE Trans. Audio, Speech and Language Processing 15*, 2007.

[7] K. Fukunaga. Introduction to Statistical Pattern Recognition. 2nd ed. New York: Academic Press, 1990, cap. 10.

[8] S. J. D. Prince e J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity", *in Proc. IEEE ICCV*, Rio de Janeiro, Brasil, Out 2007.

[9] P. Kenny. "Bayesian speaker verification with heavy tailed priors", *In: Proceedings of Speaker Odyssey,* 2010.

[10] M. Krishna, V. Grade e U. Dessay, Multifractal Based Network Traffic Modeling. *Bombay: Kluwer Academic Publishers*, 2003.

[11] C. Ynoguti e F. Violaro. "Reconhecimento de fala contínua usando Modelos ocultos de Markov". PhD tese, Universidade Estadual de Campinas, 1999.

[12] "Advancing human language technology in Brazil and the United states through collaborative research on Portuguese spoken language systems." Federal University of Rio Grande do Sul, University of Caxias do Sul, Colorado University, and Oregon Graduate Institute, 2001.

[13] T. Kinnunen e H. Li, "An overview of text-independent speaker recognition: from features to supervectors", *Speech Communication*, vol. 52, no. 1, pp. 12-40, Jan. 2010.

[14] S. Sadjadi, M. Slaney e L. Heck, "MSR Identity Toolbox v1. 0: A MATLAB Toolbox for Speaker Recognition Research", *Speech and Language Processing Technical Committee Newsletter*, 2013.