v. 4, n. 2, p. 33-37, 2015
ISSN 2237-9223

REVISTA BRASILEIRA DE CRIMINALÍSTICA

DOI: http://dx.doi.org/10.15260/rbc.v4i2.92

Fonética Forense: o uso da fusão de escores para verificação de locutor independente de texto

M.F. Silva [a*], D. Fernandes [a], M.C.F. de Castro [a]

[a] *Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS), Faculdade de Engenharia, Porto Alegre (RS), Brasil*

**Endereço de e-mail para correspondência: may.ferreira@gmail.com. Tel.: +55-51-92888079.*

Recebido em 19/06/2015; Revisado em 31/08/2015; Aceito em 31/08/2015

Resumo

Este artigo apresenta uma visão geral acerca de Verificação de Locutor Independente de Texto, demonstrando o funcionamento básico de um sistema baseado na aplicação do método da fusão de escores. Detectado um ponto a ser trabalhado dentro da etapa de extração de características, objetiva-se determinar coeficientes ou um conjunto destes, relevantes para discriminação do locutor, com o intuito de minimizar a EER (*Equal Error Rate*). Primeiramente, é redigida uma breve introdução ao projeto, contextualizando a proposta, e uma revisão sucinta do estado-da-arte. Logo, apresenta-se a metodologia aplicada e os resultados obtidos. Por fim, são feitas as considerações finais a respeito do trabalho e elencadas as perspectivas futuras em torno das pesquisas de Verificação de Locutor Independente de Texto. Com este trabalho atingiu-se uma redução de 4% na EER em comparação ao sistema de referência. Os resultados mostraram que a fusão de escores conduziu a resultados superiores aqueles obtidos com o procedimento usualmente adotado.

Palavras-Chave: Fonética Forense; Reconhecimento de Voz; Verificação de Locutor; Fusão de Escores; Modelo de Mistura Gaussiana.

Abstract

This article provides an overview about Text Independent Speaker Verification, showing the basic operation of a system based on the method of scores fusion. Detected a point to be worked within the feature extraction stage, the objective is to determine coefficients or a set of these, relevant to speaker discrimination, in order to minimize the EER (Equal Error Rate). First, a brief introduction to the project and a review of the state of the art are presented, contextualizing the proposal. Then, the methodology and the results obtained are presented and discussed. Finally, concluding remarks are made about the work and the future prospects around research in Text Independent Speaker Verification are listed. This work achieved a 4% reduction in EER compared to the reference system. The results showed that the scores fusion led to better performance than those obtained with the commonly adopted procedure.

Keywords: Forense Phonetic; Speech Recognition; Speaker Verification; Scores Fusion; Gaussian Mixture Models.

1. INTRODUÇÃO

Uma das principais áreas de pesquisa na atualidade é o reconhecimento biométrico. As características biométricas mais utilizadas são obtidas de faces, impressões digitais ou de voz. Este artigo se concentra na biometria através do reconhecimento de voz. Dentro da área de reconhecimento de voz, é possível destacar o reconhecimento do locutor (quem está falando), o qual apresenta aplicações relevantes na área forense.

A área de reconhecimento de locutor pode se ramificar em identificação e verificação, que muitas vezes se confundem. A identificação consiste em identificar um locutor dentro de um grupo, determinando quem é ele e se ele pertence ao grupo ou não. Já na verificação, deseja-se confirmar a identidade do locutor desconhecido, confrontando sua amostra de fala com a do locutor alvo (locutor que acredita-se ter dado origem ao áudio do locutor desconhecido) [1].

Uma das principais aplicações de verificação de locutor é a prática forense. Nas aplicações forenses, o uso da verificação de locutor auxilia os investigadores a direcionarem sua investigação, servindo como um recurso auxiliar, não podendo ser utilizado como prova final de culpa ou inocência.

Focando as aplicações forenses, o trabalho desenvolve o uso de verificação de locutor independente de texto, visto que o conteúdo das gravações (o que foi dito) não será de

conhecimento prévio, não usará uma senha ou frase determinada, e será composto de áudios de fala espontânea.

Basicamente, existem duas fases: a fase de treino e a fase de teste. A fase de treino consiste no modelamento dos locutores e do *background* (modelo gerado com base em um grande número de locutores) e a fase de teste, em modelar a declaração desconhecida e compará-la com a do locutor alvo, decidindo se a fala desconhecida pertence a tal locutor (aceita) ou não pertence (rejeita). Juntas, as fases de treino e teste constituem um processo de classificação. Na fase de treino, o locutor disponibiliza amostras de sua fala que serão usadas para treinar o modelo deste locutor.

Para o modelamento do locutor é necessário extrair suas características. As características mais utilizadas atualmente são os MFCC (*Mel Frequency Cepstral Coeficients*) [2].

Ao nível do modelamento estatístico também podem ser encontrados diversos estudos, destacando-se os *Gaussian Mixture Models* (GMM) [3], e as *Suport Vector Machines* (SVM) [4].

Dentro deste cenário, encontra-se a extração de características como um ponto a ser melhor desenvolvido, já que a utilização de apenas um tipo de característica pode não ser capaz de representar o locutor com o melhor desempenho.

A proposta deste trabalho vem a ser a de selecionar e combinar características, através dos coeficientes do vetor característico, analisando a taxa de erros resultante. A fim de buscar um melhor desempenho do sistema, comparou-se a performance com a fusão a nível escores (devido ao uso de características distintas) com a concatenação de características em um único vetor.

2. VERIFICAÇÃO DE LOCUTOR INDEPENDENTE DE TEXTO

Para a tarefa de verificação de locutor encontramos duas fases distintas: a fase de treino (Fig. 1) e a fase de teste (Fig. 2).

Durante a fase de treino, é realizado o modelamento dos locutores, a fim de identificar os parâmetros específicos de cada locutor e armazená-los no banco de dados do sistema. Conforme a Fig. 1, a partir dos dados de fala de cada locutor é feita a extração de características, ou seja, a busca dos coeficientes que melhor distinguem o indivíduo, gerando assim um vetor característico. Esses coeficientes, por sua vez, servirão para o modelamento estatístico do locutor. O modelo do *background* também é determinado através do mesmo método. O *background* vem a ser a junção de falas de vários locutores, através de amostras de cada locutor, criando um modelo único e universal (*Universal Background Model* – UBM) [3].

Figura 1. Fase de Treino do Sistema de Verificação de Locutor. Fonte: Adaptado de [5].

Já na fase de teste, também conhecida como fase de verificação, é realizada a extração de características da fala de um locutor desconhecido. Como mostra a Fig. 2, os coeficientes gerados para este indivíduo serão comparados com o modelo do locutor alvo (locutor a ser verificado) e com o *background* (modelo universal), indicando se está mais próximo de serem do locutor (aceita) ou do *background* (rejeita), realizando um casamento de padrões.

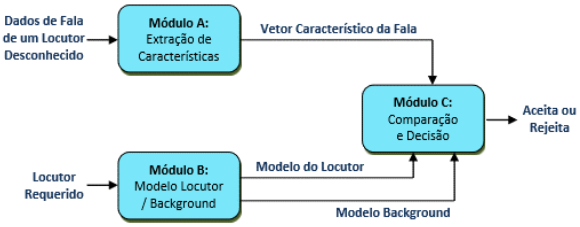

Figura 2. Fase de Teste do Sistema de Verificação de Locutor. Fonte: Adaptado de [5].

Para aplicações na área forense, os sistemas de verificação de locutor independentes de texto são os mais utilizados[12].

2.1. Extração de Características

O módulo de extração de características é responsável por transformar o sinal de fala em um vetor característico. É comumente usado o termo *features* para definir características. O sinal de fala pode ser representado por uma sequência de vetores característicos. As *features* são utilizadas para gerar os modelos de locutores.

O objetivo da seleção de características é encontrar uma transformação para um espaço de características de dimensão relativamente baixa que preserve as informações pertinentes, permitindo comparações significativas por meio de medidas simples de similaridade [1].

2.1.1. MFCC

Os coeficientes MFCC (*Mel Frequency Cepstral Coeficients*) são os mais utilizados em Verificação de Locutor [7, 8]. Uma transformada de Fourier é aplicada a cada janela do sinal de voz, sendo ignorada a informação de fase e permanecendo somente o espectro de magnitude. É comum converter a amplitude do sinal sonoro para a escala logarítmica de decibéis, e converter os coeficientes espectrais do vetor característico para a escala MEL (similar à escala de frequência do ouvido humano) [9].

Normalmente, é feita a extração de apenas 12 coeficientes MFCC por janela [10] e os vetores característicos são gerados a cada 10 ms, porque se assume que neste período de tempo o sinal de fala é estacionário. Uma das características extraída é energia. A fim de evitar maus resultados devido ao uso desta informação, a componente energia é removida [11].

2.1.2. Coeficientes DELTA E DELTA/DELTA

A análise independente dos coeficientes cepstrais pode perder informações importantes, como a coarticulação.

A primeira derivada, ou delta, pode ser aproximada pela Eq. 1 [7], onde tipicamente P=2 e $\vec{x}_t$ é o vetor em cada frame t:

$$\overrightarrow{d_t} = \frac{\sum_{p=1}^{P} p(\vec{x}_{t+p} - \vec{x}_{t-p})}{2\sum_{p=1}^{P} p^2} \quad (1)$$

Substituindo-se o valor de $\vec{x}_t$ por $\overrightarrow{d_t}$, obtém-se a segunda derivada (delta-delta) ou parâmetros de aceleração. Esses resultados são concatenados com o vetor original MFCC.

2.2. Modelo de Mistura Gaussiana (GMM)

O GMM (*Gaussian Mixture Models*), representado por λ, nada mais é que o uso de uma mistura finita de distribuições gaussianas para aproximação (modelamento) da função densidade de probabilidade de interesse. O objetivo é modelar o locutor por meio de um modelo de distribuição estatística das *features* do locutor, através de uma mistura de gaussianas. Conforme [5], para um vetor de características D-dimensional, a densidade da mistura para posterior obtenção de uma função de verossimilhança é definida como segue:

$$p(\vec{x}|\lambda) = \sum_{i=1}^{M} w_i \cdot p_i(\vec{x}) \quad (2)$$

A densidade é, portanto, uma combinação linear ponderada de M densidades gaussianas unimodais $p_i(\vec{x})$, cada uma parametrizada por um vetor média $\vec{\mu}_i$ (Dx1) e uma matriz covariância Σ_i (DxD):

$$p_i(\vec{x}) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} e^{-(1/2)(\vec{x}-\vec{\mu}_i)'(\Sigma_i)^{-1}(\vec{x}-\vec{\mu}_i)} \quad (3)$$

A soma dos pesos da mistura w_i deve satisfazer $\sum_{i=1}^{M} w_i = 1$. Coletivamente, os parâmetros do modelo de densidade são simbolizados por $\lambda = (w_i, \vec{\mu}_i, \Sigma_i)$, i=(1... M). Normalmente, utilizam-se apenas matrizes de covariância diagonais, principalmente por serem mais eficientes computacionalmente.

A *score function* do sistema determina o valor a ser comparado com o limiar de decisão (θ) para determinar se a declaração pertence ao locutor alvo. A equação (4) apresenta a forma de determinação da *score function* (Λ). Sendo $\Lambda > \theta$, o sistema determina que a fala seja do locutor alvo (aceita), e, sendo $\Lambda < \theta$, o sistema determina que não seja do locutor alvo (rejeita).

$$\Lambda = \log p(X|\lambda_{alvo}) - \log p(X|\lambda_{ubm}) \quad (4)$$

Uma técnica comumente utilizada é gerar o modelo do locutor a partir do modelo universal, adaptando os parâmetros do UBM através de adaptação Bayesiana. Esta técnica é chamada de MAP (*Maximum a Posteriori*) e normalmente são adaptadas apenas as médias, permanecendo iguais os outros parâmetros [3].

3. SISTEMA PROPOSTO

Foi implementado um sistema de Verificação de Locutor Independente de Texto. Num primeiro momento, foram utilizados os coeficientes MFCC para representação das características de cada locutor e GMM como modelo estatístico.

Após, foi desenvolvido um método para gerar coeficientes através de um filtro auto regressivo, que utiliza o método da covariância modificada, gerando coeficientes que modelam a variação dos coeficientes MFCC através do tempo. O método da covariância modificada estima os parâmetros minimizando o erro preditivo, por via da minimização dos erros preditivos posterior e anterior [12].

Com a obtenção destes coeficientes, foi possível implementar o sistema de outras duas diferentes formas, apenas com os novos coeficientes sendo utilizados para gerar o modelo do locutor, e também concatenando os coeficientes MFCC com os novos coeficientes e através deste único vetor gerando o modelo do locutor.

Além disso, foi realizada uma fusão, a nível de escore, do sistema utilizando apenas coeficientes MFCC e do sistema utilizando os novos coeficientes. Os coeficientes delta também foram testados para comparação, inclusive através da fusão.

Outra análise do trabalho foi com relação a variações de relação sinal-ruído, verificando o comportamento dos sistemas em diferentes situações.

3.1. Banco de Dados

O banco de dados elaborado incluiu áudios de falas em português brasileiro, de indivíduos do gênero masculino,

extraídos de forma espontânea de um programa de rádio sobre temas diversos.

Para composição do *background* foram usados 120 locutores, com 30 s de fala cada um, totalizando 1 hora de falas.

3.2. Resultados

Os testes foram feitos com 5 diferentes sistemas (MFCC, MFCC-LPC, MFCC+LPC, MFCC-Δ-ΔΔ e MFCC+Δ+ΔΔ), que serão detalhados a seguir, e para cada sistema foram realizados 16 diferentes testes (com situações diferentes em relação à presença de ruído nos áudios do *background* e nos áudios de teste).

O sistema de referência construído com base no estado-da-arte, o sistema MFCC, apresentou os resultados da terceira coluna da Tab. 1. Observando a Tab. 1, os valores estão de acordo com a literatura, onde a taxa de erro é baixa para áudios de ótima qualidade e aumenta conforme a relação sinal ruído presentes nos áudios.

As pesquisas realizadas [5, 10, 13] também afirmavam que quanto mais próximos fossem os áudios do *background* dos áudios de teste, melhores seriam os resultados do sistema. Isso se comprova na Tab.1 ao avaliarmos cada grupo. Focando o grupo 4, pode-se perceber uma enorme diferença, já que, com 20dB de SNR (áudios de baixíssima qualidade) a taxa de erro diminui de 21,94% para 7,45%, simplesmente porque os áudios do *background* têm a mesma relação sinal-ruído dos áudios de teste.

Implementou-se o sistema MFCC-Δ-ΔΔ concatenando-se 12 coeficientes delta (Δ) e 12 coeficientes delta-delta (ΔΔ) ao vetor característico original com 12 coeficientes MFCC. Os resultados encontram-se na quarta coluna da Tab. 1.

Tabela 1. EER (%) dos sistemas implementados.

Situação		EER (%)				
Background	**Teste**	**MFCC**	**MFCC-Δ-ΔΔ**	**MFCC-LPC**	**MFCC+LPC**	**MFCC+Δ+ΔΔ**
Sem Ruído	Sem Ruído	1,68	1,45	6,82	1,45	1,29
Sem Ruído	60dB	2,10	1,78	7,45	1,94	1,45
Sem Ruído	40dB	6,32	5,60	15,55	5,38	3,24
Sem Ruído	20dB	50,25	48,94	44,39	44,57	41,97
60dB	Sem Ruído	1,83	1,45	7,57	1,61	1,45
60dB	60dB	1,54	1,45	6,57	1,62	1,43
60dB	40dB	3,80	3,52	14,29	3,40	2,10
60dB	20dB	48,13	45,76	43,11	43,59	37,93
40dB	Sem Ruído	4,88	4,09	11,50	4,05	4,05
40dB	60dB	4,53	3,88	11,18	3,72	3,52
40dB	40dB	2,91	2,26	8,75	2,10	1,62
40dB	20dB	39,38	41,09	39,40	39,38	27,06
20dB	Sem Ruído	28,36	26,09	43,76	27,87	20,74
20dB	60dB	28,03	41,16	25,28	26,25	20,68
20dB	40dB	21,94	38,18	23,02	21,39	16,27
20dB	20dB	7,45	20,29	10,21	8,87	4,34

O método de extração de características desenvolvido foi o Sistema MFCC-LPC. Ele usa um preditor linear de 2ª ordem, que prediz o comportamento temporal dos coeficientes MFCC. Sendo 12 os coeficientes MFCC, o LPC gera 24 coeficientes que são concatenados ao vetor característicos com os coeficientes MFCC. O novo vetor característico tem um total de 36 coeficientes. Na quinta coluna da Tab. 1 encontram-se os resultados deste sistema. A janela de predição foi feita a cada 12 coeficientes MFCC, visto que o aumento da janela não proporcionou nenhum ganho ao sistema.

Analisando a Tab. 1, verifica-se que o sistema MFCC-LPC apresenta piores resultados, muito acima do sistema de referência, o sistema MFCC. Isto se justifica pelo fato de os coeficientes MFCC e os coeficientes LPC possuírem média e desvio padrão diferentes, ou seja, seus valores não estão normalizados.

A fusão é dada através da seguinte equação:

$$Score_{fusão} = w \cdot Score_{MFCC} + (1 - w) \cdot Score_{LPC} \quad (5)$$

O sistema que realiza a fusão é o sistema MFCC+LPC, cujo melhor fator de ponderação w escolhido foi de w=0,6. Os resultados são apresentados na sexta coluna da Tab. 1. A partir desses dados confirma-se que o sistema desenvolvido, através do método da fusão, traz melhor performance ao sistema, diminuindo a taxa de erro numa média de 0,81%. A performance do sistema MFCC+LPC se iguala à do sistema MFCC-Δ-ΔΔ.

Por fim, foi realizado um último teste, no qual aplicou-se a fusão com os coeficientes delta e delta/delta. O valor

de w=0,35 gerou os melhores resultados.

A partir dos dados da sétima coluna da Tab. 1 conclui-se que os coeficientes delta/delta-delta são muito mais eficientes em conjunto com os coeficientes MFCC se forem combinados a nível de *score*, ao invés de serem concatenados no vetor característico, como normalmente é feito. A taxa de erro diminui numa média de 4%. A performance do sistema MFCC+Δ+ΔΔ se mostra a melhor dentre os sistemas discutidos até aqui.

A pior performance foi a do sistema MFCC-LPC, que resultou em um aumento da taxa de erro de 6,6% em média com relação ao sistema de referência (sistema MFCC). Os sistemas MFCC-Δ-ΔΔ e MFCC+LPC apresentaram resultados bem próximos, o primeiro com uma redução média de 0,65% na taxa de erro, e o segundo, com uma redução média de 0,81% na taxa de erro. Ambos apresentam bons resultados. Finalmente, o sistema MFCC+Δ+ΔΔ superou os outros sistemas, com a menor taxa de erro.

4. CONCLUSÕES

O objetivo principal do trabalho foi alcançado através da seleção dos coeficientes LPC, proporcionando características do sinal de voz, da variação temporal dos coeficientes MFCC, e o conjunto desses coeficientes, a partir do método da fusão, possibilitou obter melhores resultados que o estado-da-arte em Verificação de Locutor Independente de Texto.

Criou-se um banco de falas em português que foi de extrema importância para a realização do trabalho.

Como resultado direto do trabalho, identificou-se nos testes realizados, assim como foi comentado em publicações [5, 10, 13], que o ideal é que o *background* escolhido fosse o mais próximo possível do áudio desconhecido. No trabalho os testes foram feitos com diferenças na relação sinal-ruído presente nos áudios e constatou-se através das simulações que a melhor situação é quando temos os áudios no *background* com a mesma relação sinal-ruído do áudio testado, comprovando-se o que foi sugerido nas publicações [5, 10, 13].

Durante o processo de simulação dos sistemas observou-se que o método inicialmente proposto, o sistema MFCC-LPC, não forneceu os resultados esperados. Contudo, ao implementa-lo através do método da fusão, como feito no sistema MFCC+LPC, o sistema se equivaleu aos métodos mais avançados já desenvolvidos, como o caso do sistema MFCC-D-DD.

Enfim, como um ganho adicional, aplicou-se a fusão a um método já consagrado na literatura, o delta-delta, realizando a fusão do sistema MFCC com o delta. Esta diversificação do método permitiu implementar um sistema com uma performance muito melhor, reduzindo consideravelmente a taxa de erro e aumentando a confiabilidade do sistema.

Considerando-se os resultados obtidos no presente trabalho, propõe-se como trabalhos futuros a implementação de um sistema de detecção de níveis de sinal ruído no sinal gravado para que seja utilizado o mesmo nível na composição do *background*.

Além disso, a inserção de informações linguísticas da fala (vocabulário pessoal, pronúncia, uso de fonemas e siglas) junto aos métodos de modelamento dos locutores já existentes seria de grande relevância, pois características específicas do locutor encontram-se presentes ali. Essas características também poderiam ser conjugadas através do método da fusão apresentado neste trabalho.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] J.P.C. JR. Speaker Recognition: a tutorial. *Proceedings of the IEEE* **85(9)**, 1437-1462, 1997.

[2] M.A. Hossan; S. Memon; M.A. Gregory. A novel approach for MFCC feature extraction. *4th International Conference on Signal Processing and Communication Systems (ICSPCS)*, 2010.

[3] D.A. Reynolds; T.F. Quatieri; R.B. Dunn. Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*. **10**, 19-41, 2000.

[4] W.M. Campbell, D.E. Sturim, D.A. Reynolds. Support vector machines using GMM supervectors for speaker verification. *Signal Proces. Letters* **13(5)**, 308-311, 2006.

[5] F. Bimbot, J.F. Bonastre, C. Fredouille, G. Gravier, I.M. Chagnolleau, S. Meignier, T. Merlin, J.O. Garcia, D.P. Delacrétaz, D.A. Reynolds. A tutorial on Text-Independent Speaker Verification. *Eurasip Journal on Applied Signal Processing*. **4**, 430-451, 2004.

[6] J. P. Campbell, W. Shen, W.M. Campbell, R. Schwartz, J.-F. Bonastre, D. Matrouf. Forensic speaker recognition. *Signal Processing Magazine, IEEE* **26(2)**, 95-103, 2009.

[7] S.Z. Li; A.K. Jain. Encyclopedia of Biometrics. Springer, 2009.

[8] U. Bhattacharjee; K. Sarmah. A Multilingual Speech Database for Speaker Recognition. *IEEE International Conference on Signal Processing, Computing and Control (ISPCC)*, 2012.

[9] W.M. Hartmann. Signals, sound and sensation. *Springer Science & Business Media*, 1997.

[10] R. Togneri; D. Pullella. An overview of Speaker Identification: Accuracy and Robustness Issues. *IEEE Circuits and Systems Magazine*, **11(2)**, 23-61, 2011.

[11] K. Modi; L. Saul. Text Independent Speaker Verification System, 2006.

[12] A.C. Gonçalves. Processamento Digital de Sinais. Estimação Paramétrica. *Universidade Federal do Paraná*, 2007.

[13] D.A. Reynolds. Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication* **17**, 91-108, 1995.