Verificação de Locutores Independente de Texto: uma Análise de Robustez a Ruído

Hector N. B. Pinheiro, Tsang Ing Ren, George D. da C. Cavalcanti

Centro de Informática – Universidade Federal de Pernambuco (UFPE)
Recife – PE – Brazil

{hnbp,tir,gdcc}@cin.ufpe.br

Abstract. *This work focuses on the development of text-independent speaker verification systems. The main challenge in the development of these systems comes from the mismatches which may occur in the acquisition of the speech signals. The techniques proposed to reduce the mismatch effects are referred as compensation methods. They may operate in three domains: in the feature extraction process, in the estimation of the speaker models and in the computation of the decision score. This work presents a wide description of the main techniques used in the development of text-independent speaker verification systems and the main feature-, model- and score-based compensation methods. In the experiments, we present a comprehensive comparison between the conventional techniques and the alternatively compensation methods. Furthermore, two compensation methods are proposed: one operates in the model domain and the other in the score domain, which outperformed the main compensation techniques.*

Resumo. *Este trabalho foca no desenvolvimento de sistemas de verificação de locutores independente de texto, cujo principal desafio provém das chamadas incompatibilidades que podem ocorrer na aquisição dos sinais de voz. As técnicas propostas para suavizá-las são chamadas de técnicas de compensação e três são os domínios onde elas podem operar: no processo de extração de características do sinal, na construção dos modelos dos locutores e no cálculo do score final do sistema, utilizado na autenticação. Esse trabalho apresenta uma vasta revisão da literatura do desenvolvimento de sistemas de verificação independentes de texto, das técnicas de compensação de características, modelos e scores. Na fase de experimentação, uma análise comparativa das principais técnicas propostas na literatura é apresentada. Além disso, duas técnicas de compensação são propostas, uma do domínio de modelagem e outra do domínio dos scores, que por suas vezes apresentaram desempenhos superiores às principais técnicas da literatura.*

1. Introdução

O processo de identificação de um determinado indivíduo é realizado milhões de vezes, todos os dias, por organizações dos mais diversos setores. Perguntas como ”Quem é esse indivíduo?”ou ”É essa pessoa quem ela diz ser?”são realizadas frequentemente por organizações financeiras, sistemas de comércio eletrônico, sistemas de telecomunicações e por instituições governamentais. Para esses tipos de sistemas, a autenticação errônea de

um indivíduo pode trazer consequências desastrosas. Nesse contexto, o desenvolvimento de um sistema de identificação pessoal automático com uma alta acurácia tem se tornado cada vez mais crítica. Identificação biométrica é o processo de identificar um indivíduo a partir de características físicas ou comportamentais. Devido ao fato de muitas dessas características serem únicas, elas são capazes de proporcionar uma maior acurácia ao processo de identificação pessoal do que os métodos tradicionais (utilização de cartões, documentos ou senhas). Alguns exemplos dessas características são: face, impressão digital, íris, assinatura e voz. Reconhecimento de locutores é uma modalidade biométrica que propõe realizar o processo de identificação pessoal a partir das informações presentes unicamente na voz do indivíduo. Este trabalho foca no desenvolvimento de sistemas de verificação de locutores independente de texto. Nesse tipo de modalidade, o sistema deve executar uma autenticação (verificar a identidade alegada pelo usuário) independentemente do conteúdo fonético presente na locução. Sistemas de verificação de locutores

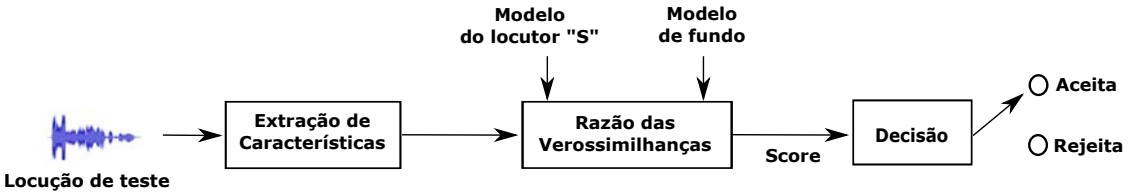

Figura 1. Arquitetura básica de um sistema de verificação de locutores independente de texto.

devem decidir se uma dada locução, X, foi gerada por um locutor específico, S, ou não. Dessa maneira, duas hipóteses são consideradas: H_0={X foi produzida por S} e H_1={X não foi produzida por S}. O sistema deve, portanto, ser capaz de modelar tais hipóteses, de modo que seja possível o cálculo da verossimilhança delas ocorrerem. A arquitetura básica de um sistema desse tipo é mostrada na Figura 1. Na fase de cadastramento ocorre a estimação dos modelos associados às hipóteses. O modelo associado à hipótese nula é referenciado como modelo do locutor e é estimado utilizando locuções de treino produzidas por S. Já o chamado modelo de fundo, associado à hipótese alternativa, é estimado utilizando locuções produzidas por diversos locutores, geralmente diferentes de S. Na fase de verificação (ou autenticação), as verossimilhanças da locução de teste com respeito a esses modelos são computadas, de modo que a decisão final seja realizada. O módulo de extração de características possui o objetivo de realizar uma descrição mais compacta do sinal de voz, através de vetores de características. Até então, há um domínio absoluto da utilização de extratores espectrais de tempo curto, em especial, aqueles que extraem os Coeficientes Mel-cepstrais (*Mel-frequency cepstral coefficients*, MFCCs) acompanhados de suas componentes dinâmicas (coeficientes delta de primeira e segunda ordem). A abordagem tradicional para a modelagem dos locutores consiste na técnica GMM-UBM [Reynolds et al. 2000]. Nela, ambas as hipóteses são modeladas através de modelos de misturas Gaussianas (*Gaussian Mixture Models*, GMMs). Primeiramente, um modelo de fundo universal (*Universal Background Model*, UBM) e independente de locutor é estimado através do algoritmo EM (emphExpectation-Maximization) utilizando locuções de diversos locutores. O modelo do locutor S é então produzido pela adaptação das médias das distribuições do UBM utilizando as locuções de S. Essa adaptação é realizada maximizando a probabilidade a posteriori das misturas do modelo (adaptação *Maximum A Poteriori*, MAP). Outra abordagem mais recente é chamada de GMM-SVM

[Campbell et al. 2006], onde cada locução é descrita por um ”supervetor GMM”, que descreve o GMM resultante da adaptação MAP do UBM utilizando a locução. Um SVM (*Support Vector Machine*) é então treinado para realizar a classificação das locuções, no espaço dos supervetores GMM.

2. Técnicas de compensação de ruído

O grande desafio no desenvolvimento de sistemas de reconhecimento de locutores consiste das chamadas incompatibilidades apresentadas entre os sinais de voz utilizados na estimação dos modelos e aqueles utilizados na autenticação. Muitas são as possíveis fontes de incompatibilidade, que vai desde o tipo de microfone utilizado até o ruído acústico presente no ambiente ou a qualidade do canal de comunicação por onde o áudio é transmitido. Tais fatores impactam na geração do sinal de voz, produzindo distorções que dificultam o reconhecimento. As técnicas desenvolvidas para suavizar tais distorções são chamadas de técnicas de compensação e elas operam sobre três domínios: na extração de características, na modelagem dos locutores e no cálculo dos *scores*. As técnicas de compensação de características se concentram na remoção dos efeitos do ruído sobre o processo de extração das características do sinal e podem operar em diferentes fases da extração. As principais técnicas são: Subtração de Média Cepstral; Normalização de Média Cepstral; Filtragem RASTA; Deformação de Características e Coeficientes de Amplitudes Descorrelacionadas. Além disso, elas podem ser combinadas entre si, se operarem sobre etapas distintas da extração de características. Já as técnicas de compensação de *scores* tentam normalizar os *scores* produzidos por um determinado locutor, de modo a facilitar a escolha de um limiar de classificação. As principais normalizações aplicadas são: Normalização Zero, Normalização de Teste e Normalização *Handset*. Por fim, as técnicas de compensação de modelo propõem estimar modelos de locutores que sejam capazes de lidar com possíveis distorções existentes nas locuções utilizadas para autenticação. As primeiras técnicas surgiram especificamente para tratar as incompatibilidades de canal, mais precisamente aquelas que surgem pela utilização de diferentes tipos de microfones. Porém, os modelos atuais se propõem a lidar com distorções de quaisquer tipos, de uma maneira geral. As principais técnicas de compensação de modelo são: Síntese do Modelo do Locutor e Mapeamento de Características, ambas aplicadas à modelagem GMM-UBM; Projeção dos Atributos Indesejáveis, aplicada à modelagem GMM-SVM e a técnica PUM-GMM, que combina os chamados modelos de união *a posteriori* (*Posterior Union Models* - PUMs) e teoria dos dados ausentes.

Uma descrição detalhada de cada uma dessas técnicas é apresentada em [Pinheiro 2015]. Nesse trabalho, tais técnicas são avalidas e comparadas entre si, apresentando assim um panorama da eficácia que as compensações proporcionam em conjunto ás modelagens GMM-UBM e GMM-SVM. Além disso, duas técnicas de compensação são propostas nesse trabalho. A primeira delas é uma técnicas de compensação de *scores* que realiza a normalização dos *scores* através da distribuição normal acumulada dos *scores* produzidos pelo locutor correspondente à alegação. A segunda delas, do domínio dos modelos, propõe uma modelagem baseada no PUM-GMM [Ming et al. 2007]. A formulação apresentada pelos autores da técnica não é completamente adequada para verificação de locutores independente de texto e considera uma abordagem *a posteriori* de uma quantidade comumente elevada de modelos de locutores (diferentes do locutor associada à alegação) para a formação do modelo de união necessário para a escolha do conjunto de

características adequado para a autenticação (seguindo a teoria dos dados ausentes). A técnica proposta apresenta uma formulação mais apropriada para a tarefa, combinando os dois conceitos utilizados pelos autores com um tipo de modelagem utilizando modelos de fundo para a formação do modelo de união. As duas técnicas propostas foram também comparadas às técnicas de compensação presentes na literatura e se mostraram bastante eficazes para seus propósitos. Detalhes desses resultados são mostrados mais adiante. Por conta de limitações ao escopo do trabalho [Pinheiro 2015], outra técnica de compensação de modelo proposta pelos autores não foi apresentada. Tal técnica é referenciada como Type-2 Fuzzy GMM-UBM, que combina os chamados GMMs difusos de tipo 2 [Tsang et al. 2012, Pinheiro et al. 2013] a métodos de treinamento multicondicional [Pinheiro et al. 2014] para estimar modelos difusos robustos a incompatibilidades. O modelo proposto foi avaliado utilizando o mesmo protocolo apresentado [Pinheiro 2015] e demonstrou desempenhos superiores à modelagem tradicional GMM-UBM [Pinheiro et al. 2016], apresentando um ganho de desempenho de $31,5\%$ em ambientes com alto grau de ruído.

3. Experimentos

Em [Pinheiro 2015], experimentos foram realizados para analisar a robustez dos sistemas de verificação de locutores na presença de incompatibilidades provenientes de ambientes ruidosos. Para isso a base de dados pública *MIT Mobile Device Speaker Verification Corpus* foi utilizada. Ela apresenta locuções provenientes de 48 locutores gravadas em três sessões distintas (uma utilizada para treinamento e outras duas para teste). Um total de 54 locuções (para cada locutor em cada sessão) foram geradas em três ambientes distintos: um escritório silencioso, o *hall* de entrada de um prédio, com nível mediano de ruído, e no cruzamento de ruas movimentadas, com alto nível de ruído de ambiente. Os sistemas foram então treinados utilizando as locuções provenientes do local silencioso (sem ruído) e testado em todos os ambientes. Dessa maneira foi possível a análise da robustez dos sistemas quando há alta incompatibilidade de ruído entre as locuções. Os experimentos foram divididos em quatro etapas, com objetivos específicos. Em cada uma delas, as melhores configurações de sistemas foram identificadas. Essa identificação foi realizada utilizando um teste estatístico pareado, o *Wilcoxon Signed-Rank*, através dos desempenhos encontrados nos pontos de operação que apresentam taxas de falsa aceitação e rejeição iguais, chamado de *Equal Error Rate* (EER).

A primeira etapa consistiu na comparação dos métodos convencionais, sem a presença de nenhum tipo de compensação. Basicamente, foram utilizadas diferentes configurações de coeficientes MFCC como possibilidades de características, e as duas modelagens: GMM-UBM e GMM-SVM, variando o número de misturas dos GMMs. Um total de 21 sistemas foram testados para ambas as modelagens ao variar a utilização dos coeficientes dinâmicos e ao variar a quantidade de distribuições dos modelos (potências de 2, de 16 a 1024). Nesse experimento foi constatada a superioridade estatística da modelagem GMM-SVM em relação ao GMM-UBM, quando as melhores configurações encontradas para as modelagens foram comparadas.

Já na segunda etapa dos experimentos foram avaliadas as técnicas de compensação de características propostas na literatura e as possíveis combinações entre elas. Ao variar mais uma vez a quantidade de misturas dos modelos e a utilização dos coeficientes dinâmicos dos MFCCs, um total de 294 configurações foram testadas para ambas

as modelagens. Para a modelagem GMM-UBM, um total de 12 configurações foram identificadas como as melhores e cada uma das técnicas da literatura aparecem em alguma delas. Além disso, constatou-se que, para essa modelagem, combinar técnicas de compensação de características aumenta significativamente a robustez dos sistemas (todas as 12 configurações apresentam combinações de técnicas). Por outro lado, na modelagem GMM-SVM, apenas 4 configurações foram consideradas as melhores e absolutamente todas elas apresentam apenas técnicas de compesação baseadas na normalização das características (subtração e normalização de média cepstral). Tais técnicas de normalização aparentam ser significativamente mais adequadas à abordagem GMM-SVM do que as demais. Por fim, nessa etapa foi constatada uma equivalência estatística entre os desempenhos encontrados pelas melhores configurações das modelagens GMM-UBM e GMM-SVM, o que mostra que, ao utilizar técnicas de compesação de características, a modelagem GMM-UBM é capaz de equiparar-se em desempenho ao GMM-SVM.

Na terceira etapa dos experimentos foram avaliadas as técnicas de compensação de *scores* presentes na literatura e a técnica proposta em [Pinheiro 2015]. Para cada tipo de modelagem, as configurações utilizadas (de características e de modelagem) foram fixadas para aquelas que apresentaram as menores taxas de EER nas etapas anteriores dos experimentos. Em ambas as modelagens, foram analisados os casos em que há ou não a compensação de características. Para a modelagem GMM-UBM, foi constatada uma superioridade da técnia Normalização Zero, para ambos os casos em que compensação de características ocorre ou não. Nesse caso, a aplicação da técnica proposta diminui a taxa de EER, mas não apresentou ganho estatisticamente significativo. Já na modelagem GMM-SVM, a técnicas proposta foi aquela que apresentou o maior ganho de desempenho e se mostrou estatisticamente superior às demais. Como isso ocorreu para ambos os casos em que a compensação de características ocorre ou não, a técnica proposta se mostrou uma excelente alternativa para ser utilizada em conjunto à modelagem GMM-SVM.

Na quarta e última etapa dos experimentos, foram avaliadas as técnicas de compensação de modelos propostas na literatura e a técnica proposta em [Pinheiro 2015]. Primeiramente, a técnica Projeção de Atributos Indesejáveis foi comparada à modelagem convencional GMM-SVM, uma vez que ela é aplicável somente a essa abordagem. Nesse caso as melhores configurações da modelagem GMM-SVM observadas na segunda etapa dos experimentos foram utilizadas. Porém, nenhum ganho estatisticamente significativo foi observado em nenhuma delas. Em seguida, a técnica de compensação PUM-GMM e a técnica proposta foram comparadas com a técnica GMM-UBM. Mais uma vez, as melhores configurações encontradas para essa modelagem na segunda etapa dos experimentos foram utilizadas. Nesse caso, o desempenho encontrado pela técnica PUM-GMM foi significativamente superior ao da abordagem GMM-UBM, com ganho de desempenho de $15,5\%$. O mesmo foi observado para a técnica proposta, que provocou um ganho de desempenho de $11,1\%$. Porém, como observado anteriormente, a técnica PUM-GMM não é completamente adequada para verificação de locutores e por essa razão requer uma quantidade elevada de modelos de outros locutores, o que aumenta o custo computacional na fase de autenticação. Nesse contexto, a formulação apresentada em [Pinheiro 2015] utiliza o modelo de fundo para esse propósito, adequando os conceitos do PUM-GMM para o contexto de verificação. Tais resultados mostram, portanto, que a formulação proposta é tão eficaz quanto a da técnica da literatura, requerendo, porém, um custo computacional bem inferior.

4. Contribuições

Em [Pinheiro 2015] podemos destacar as seguintes contribuições: (i) descrição dos principais métodos utilizados no desenvolvimento de sistemas de verificação de locutores independente de texto; (ii) descrição dos principais métodos de compensação de ruído presentes na literatura; (iii) avaliação e comparação dos principais métodos utilizados no desenvolvimento dos sistemas; (iv) avaliação e comparação dos principais métodos de compensação presentes na literatura; (v) proposta de um método de compensação de *scores* baseado na Distribuição Normal Acumulada; (vi) proposta de um método de compensação de modelos baseado no treinamento multicondicional, na Teoria dos Dados Ausentes e na modelagem GMM-UBM; (vii) desenvolvimento de um algoritmo quadrático que realiza o cálculo da probabilidade de um Modelo de União a Posteriori; (viii) avaliação e comparação entre os métodos propostos e os métodos presentes na literatura.

Referências

Campbell, W. M., Sturim, D. E., and Reynolds, D. A. (2006). Support vector machines using GMM supervectors for speaker verification. *IEEE Signal Processing Letters*, 13(5):308–311.

Ming, J., Hazen, T. J., Glass, J. R., and Reynolds, D. A. (2007). Robust speaker recognition in noisy conditions. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(5):1711–1723.

Pinheiro, H. N., Ren, T. I., Cavalcanti, G. D., Jyh, T. I., and Sijbers, J. (2013). Type-2 fuzzy GMM-UBM for text-independent speaker verification. In *International Conference on Systems, Man, and Cybernetics (SMC)*, pages 4328–4331. IEEE.

Pinheiro, H. N., Ren, T. I., Cavalcanti, G. D., Jyh, T. I., and Sijbers, J. (2014). Type-2 fuzzy GMMs for robust text-independent speaker verification in noisy environments. In *International Conference on Pattern Recognition (ICPR)*, pages 4531–4536. IEEE.

Pinheiro, H. N. B. (2015). *Verificação de Locutores Independente de Texto: uma Análise de Robustez a Ruído*. Dissertação (Mestrado em Ciência da Computação) - Centro de Informática, Universidade Federal de Pernambuco, Recife, Brazil.

Pinheiro, H. N. B., Vieira, S. R. F., Ren, T. I., Cavalcanti, G. D. C., and Mattos Neto, P. S. G. (2016). Type-2 fuzzy GMM for text-independent speaker verification under unseen noise conditions. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Reynolds, D. A., Quatieri, T. F., and Dunn, R. B. (2000). Speaker verification using adapted gaussian mixture models. *Digital signal processing*, 10(1):19–41.

Tsang, R., Gabriel, D., Pinheiro, H. N., and Cavalcanti, G. D. (2012). Speaker verification using type-2 fuzzy gaussian mixture models. In *International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2336–2340. IEEE.