Utilização de Técnicas de Aprendizagem de Máquina em Biometria de Voz

RESUMO EXPANDIDO - Disciplina de TCC290009

Henrique Hilleshein

Estudante do Curso de Engenharia de Telecomunicações

Mario Noronha Neto

Professor orientador

Semestre 2018-1

Resumo- *A necessidade do aumento de segurança nos sistemas atuais é algo crescente, uma forma de conseguir uma maior segurança é utilizando a biometria de voz. Este trabalho consta um introdutório de reconhecimento de falante para verificação. São passados os conceitos básicos de biometria de voz, reconhecimento de falante, extração de características de um sinal de voz e aprendizado de máquina. O objetivo principal do trabalho é testar modelos de aprendizado de máquina para a aplicação de reconhecimento de falante.*

Palavras-chave: Reconhecimento de falante. Aprendizado de máquina. Verificação.

1 Introdução

A importância da segurança e da identificação de pessoas é algo crescente na sociedade atual, fazendo surgir vários métodos para que se consiga suprir essa necessidade. Uma forma de atender essa necessidade é utilizando a biometria para identificar uma pessoa pelas suas características únicas. A biometria faz medidas das características fisiológicas ou comportamentais de um indivíduo para identificá-lo, podendo ser como por exemplo as características da face de uma pessoa(característica fisiológica) ou a maneira que uma pessoa anda(comportamental).

Existem tipos de biometria que são frequentemente usadas para segurança como a digital, íris e voz. Digital e íris se enquadram como uma biometria de característica fisiológica, onde se trata de algo imutável. No caso da voz pode-se fazer biometria das características fisiológicas da voz e também do comportamento, possibilitando métodos alternativos, pois a fala oferece muito mais flexibilidade e diferentes níveis para realizar o reconhecimento, o sistema pode forçar o usuário para que fale de

uma maneira particular e diferente para cada tentativa(FAUNDEZ-ZANUY; MONTE-MORENO, 2005). Uma outra vantagem da biometria de voz é que ela pode ser usada à distância, diferente das outras biometrias, um exemplo é quando se deseja autenticar ou identificar alguém via telefone.

A voz de cada pessoa possui características únicas igual a impressão digital, devido a formação orgânica da laringe, cordas vocais e todo o sistema responsável pela voz. Além das diferenças fisiológicas, também existem diferenças na maneira de falar de cada pessoa como sotaque, ritmo, entonação, padrão de pronúncia, escolha de vocabulário e assim adiante (KINNUNEN; HAIZHOU, 2010). Baseando-se nessas características é possível reconhecer um indivíduo, ou seja, reconhecer o falante.

O reconhecimento do falante comumente é utilizado no ramo da verificação ou identificação, mas também pode ser utilizado em outros ramos, como classificação, detecção, segmentação e etc. O objetivo da verificação é descobrir se a pessoa que está falando é realmente quem ela alega ser, nesse caso o sistema deve estar preparado para possíveis tentativas de enganação. No caso da identificação o objetivo é distinguir uma pessoa entre um grupo de pessoas previamente registradas (BOLES; RAD, 2017). Atualmente o ramo da verificação é o mais popular para reconhecimento de falante devido sua importância em segurança e controle de acesso(BEIGI, 2011), por esse motivo que será o ramo utilizado nesse trabalho.

O sistema de reconhecimento possui as modalidades dependente de texto, independente de texto ou uma variação entre elas. No caso da modalidade dependente de texto, a pessoa deve falar uma palavra ou um conjunto de palavras que são pré-estabelecidas pelo sistema e que já foram cadastradas previamente pelo falante para que seja feito o reconhecimento. Por outro lado, a modalidade independente de texto, o sistema deve identificar a pessoa que está falando independentemente do que está sendo falado, ou seja, o sistema aprende exclusivamente a voz da pessoa e não a combinação da voz com a texto(FAUNDEZ-ZANUY; MONTE-MORENO, 2005). Será utilizado o modalidade independente de texto neste trabalho, pois é o mais versátil e pode ser utilizado em todos os ramos de reconhecimento de falante(BEIGI, 2011).

Com a recente popularidade de aprendizagem de máquina e *deep learning*(aprendizagem profunda), aplicações de reconhecimento de falante estão ficando mais poderosas. O treinamento da máquina tem se tornado mais rápido e o número de falantes que podem ser qualificados se tornou maior (BOLES; RAD, 2017). Isso faz com que a aprendizagem de máquina se torne uma ferramenta de predição muito atraente para reconhecimento de voz.

A aprendizagem de máquina é um sistema de computador que automaticamente melhora com experiência e possui um processo de aprendizado, mas para isso é necessário que se tenha dados. O sistema utiliza os dados para encontrar padrões, conseguindo fazer um aprendizado à partir disso e fazer predições.(AYODELE, 2010).

O Banco de dados utilizado neste trabalho possui a fala de 20 pessoas no idioma inglês, chamado ELSDSR(*English Language Speech Database for Speaker Recognition*)(FENG; HANSEN, 2005).

A voz não é colocada diretamente como entrada da máquina que irá aprender, é ne-

cessário que a informação da voz passe por um pré-processamento e para isso é utilizada uma técnica para se obter características do sinal de voz. Existem diversas técnicas que podem ser usadas para extrair essas características, como LPC(*Linear predictive coding*), MFCC(Mel frequency cepstral coefficients) e PLP(Perceptual Linear Coding)(FAUNDEZ-ZANUY; MONTE-MORENO, 2005). Colocando uma informação pré-processada como entrada de uma máquina que já terminou seu processo de aprendizado, se obtém como saída a predição que o falante é quem alega ser ou se obtém como saída quem é o falante, dependendo se o interesse da aplicação é verificação ou identificação. MFCC é a técnica mais popular e normalmente usada na maioria das aplicações de extração de características de sinal de voz (RAMGIRE; JAGDALE, 2016), por esse motivo será a técnica utilizada neste trabalho.
Existem ferramentas disponíveis para o desenvolvimento de sistemas de reconhecimento de falante, como Spear(KHOURY; SHAFEY; MARCEL, 2014) e Alize(BONASTRE; WILS; MEIGNIER, 2005), mas também há aplicações prontas como *Speaker Recognition API* da Microsoft Azure e o *Ok Google* do sistema operacional *Android*. Essas ferramentas foram testadas e são comparadas com modelos desenvolvidos nesse trabalho na linguagem de programação Python. O maior foco deste trabalho é a aplicação de técnicas de aprendizagem de máquina para verificação de falante com modalidade independente de texto.

2 Metodologia

A metodologia consiste em 5 etapas que permitem montar um sistema de reconhecimento de falante. As etapas são estudar as características de um sinal de voz, localizar ferramentas que auxiliam na geração de modelos, definir bancos de dados a serem utilizados, realizar o pré-processamento do sinal de voz e implementar técnicas de aprendizagem de máquina para a biometria de voz

2.1 Estudar as características de um sinal de voz

São estudados primeiramente as características do sinal de voz que o sistema possuirá e como extrai-las para que se consiga utilizar técnicas de aprendizagem de máquina com elas. É necessário encontrar o que é importante em um sinal de áudio e descartar o que é desnecessário.

2.2 Localizar ferramentas que auxiliam na geração de modelos

Neste trabalho os algoritmos de aprendizado de máquina não são implementados, são gerados modelos utilizando esses algorítimos. Então é necessário estudar ferramentas que utilizam algorítimos aprendizado de máquina para gerar os modelos matemático/estatísticos que fazem a representação do falante.

2.2.1 Definir bancos de dados a serem utilizados

É necessário definir o banco de dados que será utilizado e quais características ele irá possuir, como a qualidade do áudio, ambiente onde foi gravado o áudio e qual o conteúdo do áudio.

2.3 Realizar o pré-processamento do sinal de voz

É necessário fazer um pré-processamento do sinal de voz, são extraídas as informações do sinal de áudio utilizando uma técnica de extração de característica de sinal de voz com o MFCC e escolhidos quais características que serão usadas.

2.4 Implementar técnicas de aprendizagem de máquina para a biometria de voz

Com todo o pré-processamento feito, são implementados os modelos de aprendizagem de máquinas. Para cada algorítimo de aprendizado de máquina são feitos diversos modelos utilizando técnicas de máxima probabilidade para localizar o melhor modelo do respectivo algoritmo que está sendo utilizado. Para localizar o melhor modelo normalmente não há um passo a passo bem definido, mas normalmente são utilizadas algumas técnicas para localizar uma convergência de modelo que traga uma alta probabilidade de acerto.

Referências

AYODELE, T. O. Machine learning overview. In: *New Advances in Machine Learning*. [S.l.]: InTech, 2010.

BEIGI, H. *Fundamentals of Speaker Recognition*. New York, NY, USA: Springer, 2011.

BOLES, A.; RAD, P. Voice biometrics: Deep learning-based voiceprint authentication system. In: *2017 12th System of Systems Engineering Conference (SoSE)*. [S.l.: s.n.], 2017. p. 1–6.

BONASTRE, J. F.; WILS, F.; MEIGNIER, S. Alize, a free toolkit for speaker recognition. In: *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.* [S.l.: s.n.], 2005. v. 1, p. 737–740. ISSN 1520-6149.

FAUNDEZ-ZANUY, M.; MONTE-MORENO, E. State-of-the-art in speaker recognition. *IEEE Aerospace and Electronic Systems Magazine*, v. 20, n. 5, p. 7–12, 2005.

FENG, L.; HANSEN, L. K. A new database for speaker recognition. 01 2005.

KHOURY, E.; SHAFEY, L. E.; MARCEL, S. Spear: An open source toolbox for speaker recognition based on Bob. In: *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. [s.n.], 2014. Disponível em: <http://publications.idiap.ch/downloads/papers/2014/Khoury_ICASSP_2014.pdf>.

KINNUNEN, T.; HAIZHOU, L. An overview of text-independent speaker recognition: From features to supervectors. *Speech communication*, v. 52, n. 1, p. 12–40, 2010.

RAMGIRE, J. B.; JAGDALE, S. M. A survey on speaker recognition with various feature extraction and classification techniques. *International Research Journal of Engineering and Technology*, v. 3, n. 4, p. 709–712, 2016.