Universidade Federal
do Rio de Janeiro
Escola Politécnica

DIARIZAÇÃO E IDENTIFICAÇÃO DE LOCUTOR EM CONTEÚDO DE VÍDEO BASEADA EM ANÁLISE DE EXPRESSÃO FACIAL VIA APRENDIZADO DE MÁQUINA SUPERVISIONADO

Renan Fasolato Basilio

Projeto de Graduação apresentado ao Curso de Engenharia de Computação e Informação da Escola Politécnica, Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Engenheiro.

Orientador: Geraldo Zimbrão da Silva

Rio de Janeiro
Fevereiro de 2020

DIARIZAÇÃO E IDENTIFICAÇÃO DE LOCUTOR EM CONTEÚDO DE VÍDEO BASEADA EM ANÁLISE DE EXPRESSÃO FACIAL VIA APRENDIZADO DE MÁQUINA SUPERVISIONADO

Renan Fasolato Basilio

PROJETO DE GRADUAÇÃO SUBMETIDO AO CORPO DOCENTE DO CURSO DE ENGENHARIA DE COMPUTAÇÃO E INFORMAÇÃO DA ESCOLA POLITÉCNICA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE ENGENHEIRO DE COMPUTAÇÃO.

Examinado por:

_____
Prof. ,

_____
Prof. ,

_____
Prof. ,

RIO DE JANEIRO, RJ – BRASIL
FEVEREIRO DE 2020

Basilio, Renan Fasolato

Diarização e Identificação de Locutor em Conteúdo de Vídeo Baseada em Análise de Expressão Facial via Aprendizado de Máquina Supervisionado/Renan Fasolato Basilio. – Rio de Janeiro: UFRJ/ Escola Politécnica, 2020.

IV, 3 p. 29, 7cm.

Orientador: Geraldo Zimbrão da Silva

Projeto de Graduação – UFRJ/ Escola Politécnica/ Curso de Engenharia de Computação e Informação, 2020.

Referências Bibliográficas: p. 3 – 3.

1. Aprendizado Supervisionado. 2. Aprendizado de Máquina. 3. Diarização de Locutor. I. da Silva, Geraldo Zimbrão. II. Universidade Federal do Rio de Janeiro, Escola Politécnica, Curso de Engenharia de Computação e Informação. III. Título.

To-Do: Write Dedication

Capítulo 1

Introdução

1.1 Descrição do Problema

A diarização de locutor consiste no processo de identificar os diferentes locutores em um conteúdo multimídia, de forma a separa-los temporalmente, definindo quando quem falou, e produzindo um tipo de roteiro para o mesmo.

Tradicionalmente, tenta-se resolver esse problema analizando exclusivamente o áudio, através da extração de *features* na forma de vetores denominados *I-vectors*, e a clusterização destes vetores. Porém, trata-se de um problema difícil; o timbre, principal característica sonora responsável pela identificação do locutor pelo ser humano, é de caráter neurológico[1], produzido pela decomposição da onda sonora em seus harmônicos pelo trato auditivo. E, ainda, como propriedade intrínseca da etapa de clusterização, a grande maioria desses algoritmos depende do conhecimento prévio do número de locutores que participam do áudio.

Dadas essas limitações, temos que o desempenho dos algoritmos considerados estado da arte é insuficiente, com taxa de erro de diarização (daqui em diante chamada de DER, do inglês *Diarization Error Rate*) de cerca de 20%[2]. Portanto, a busca de outras técnicas capazes de prover um melhor desempenho nos leva a considerar também outros sinais do conteúdo multimídia, como o de vídeo dos locutores individuais que acompanha o texto.

1.2 Motivação

Em muitas situações, no processo de transcrição de áudio e vídeo, é interessante obter também a informação de quem está falando. Com essa informação seria possível roteirizar a mídia, viabilizando uma melhor formatação do texto transcrito. Além disso, essa informação adicional permite viabilizar novos critérios de busca sobre o texto transcrito, tornando possível filtrar os resultados por locutor.

1.3 Escopo

Referências Bibliográficas

[1] A. J. Oxenham, "Pitch Perception," *Journal of Neuroscience*, vol. 32, no. 39, pp. 13 335–13 338, Sep. 2012.

[2] A. W. Zewoudie, J. Luque, and J. Hernando, "The use of long-term features for GMM- and i-vector-based speaker diarization systems," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2018, no. 1, p. 14, Sep. 2018.