DETECÇÃO DE VOZ CANTADA EM SINAIS DE ÁUDIO POLIFÔNICOS

Aplicações, Abordagens e Desafios

Shayenne Moura

23. April 2018

Instituto de Matemática e Estatística
Universidade de São Paulo

O PROBLEMA

Detecção de voz principal (Voicing Detection)

Instrumentos em geral

Detecção de presença de voz (Voice Activity Detection)

Voz humana

Detecção de voz cantada (Singing Voice Detection)

Voz humana cantada

APLICAÇÕES

APLICAÇÕES

Separação de voz cantada

Identificação de cantor

APLICAÇÕES

Transcrição de melodia cantada

Transcrição de letras

APLICAÇÕES

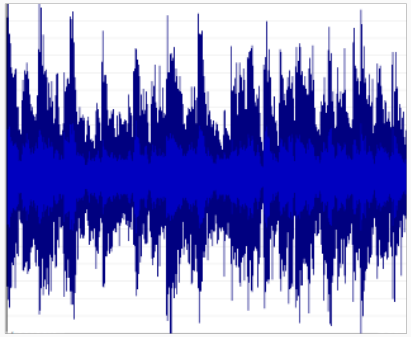
Busca cantarolada

Busca por letra

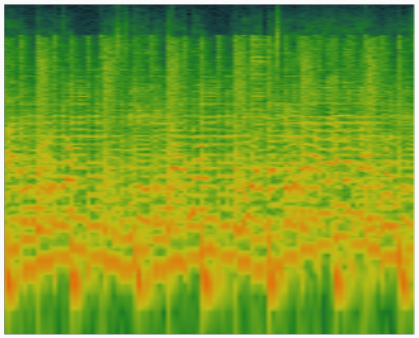
CONCEITOS BÁSICOS

ARQUITETURA BÁSICA

Sinal de áudio

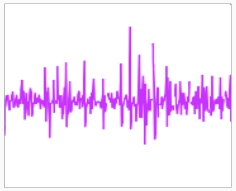


Domínio do tempo

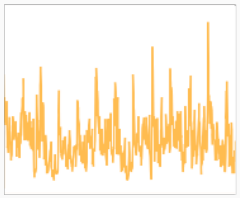


Domínio da frequência

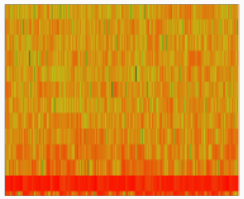
Descritores



Zero Crossing Rate



Fluxo Espectral

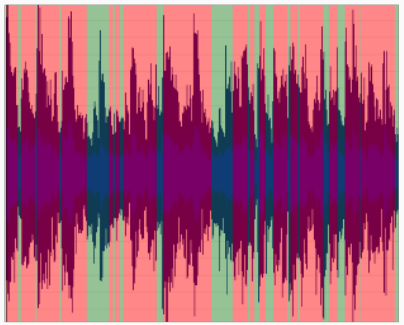


MFCC

Muitos mais...

Classificador

Saída



Resultados

CLASSIFICADOR BINÁRIO

Treinamento

Descritores

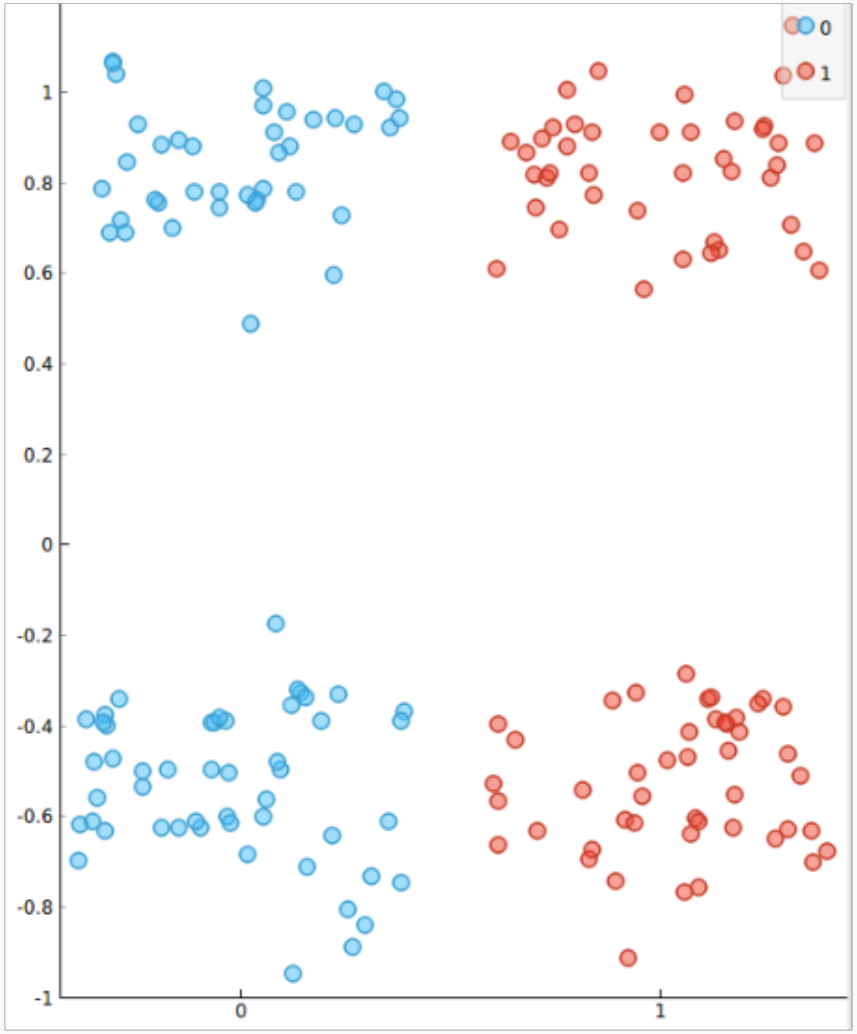
d = □□□□□□□□□□

Rótulos

r = □□□□□□□□□□

$$f : \mathcal{D} \to \{0, 1\}$$

$$g \approx f$$



CLASSIFICADOR BINÁRIO

Treinamento

Descritores

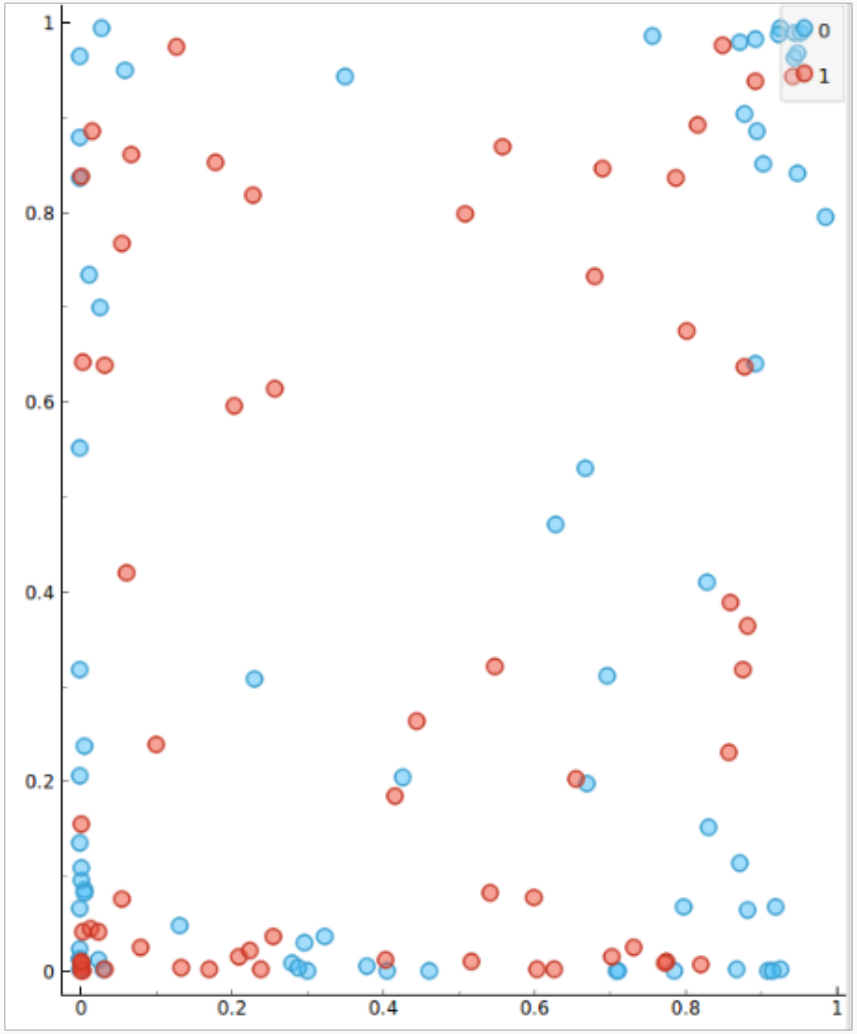
d = □□□□□□□□□□

Rótulos

r = □□□□□□□□□□

$$f : \mathcal{D} \rightarrow \{0, 1\}$$

$$g \approx f$$



ABORDAGENS

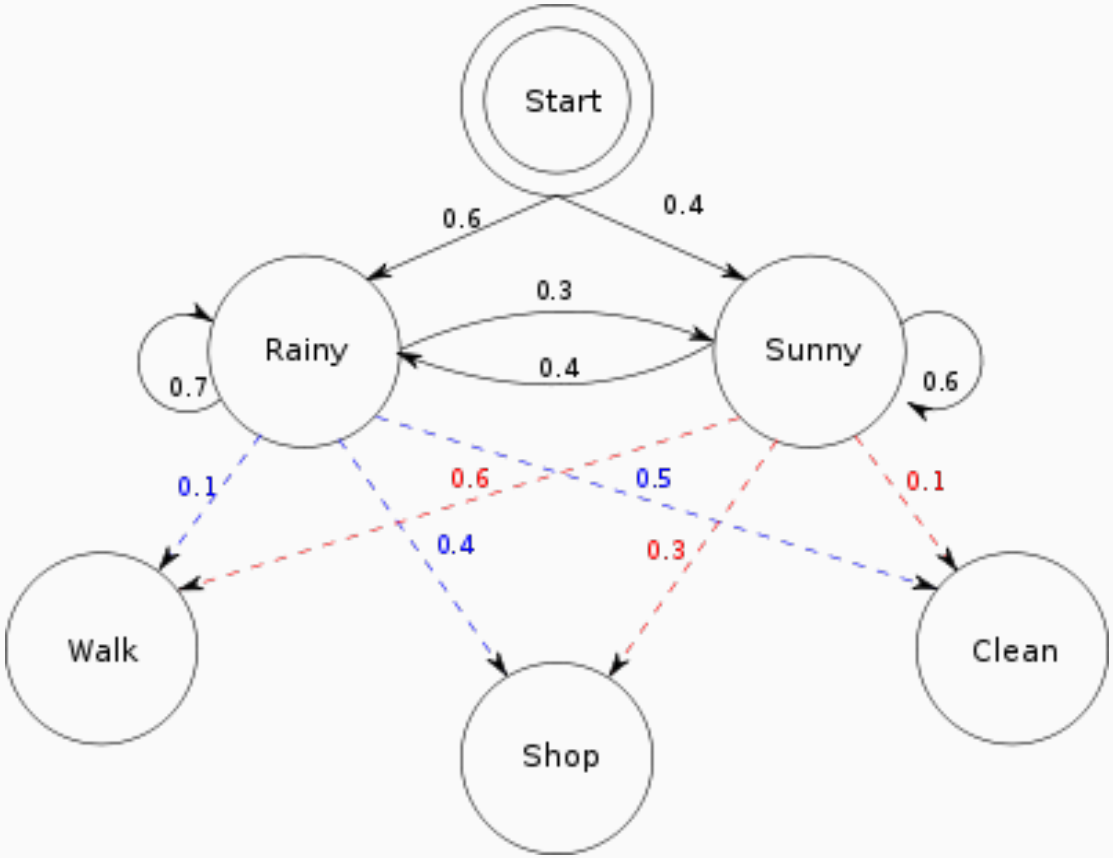
BERENZWEIG E ELLIS, 2001

Locating Singing Voice Segments Within Music Signals

- Usam uma **rede neural artificial** treinada para classificar speech-noise para gerar descritores denominados “posterior probability features”
- Calculam **estatísticas** das features obtidas para alimentar um modelo oculto de Markov (HMM)
- O **HMM** é treinada para identificar os trechos cantados das músicas em “singing” e “not singing”

Concluem que o classificador para fala não diferencia os termos acústicos de voz cantada em música popular

HIDDEN MARKOV MODELS



ZHANG, 2002

System and Method for Automatic Singer Identification

- Identifica os pontos iniciais dos **trechos cantados** em uma música usando **descritores de energia**, como zero-crossing rate (ZCR) e Fluxo Espectral
- A classificação entre vocal e não vocal é feita com uma série de **thresholds**
- Processa os trechos de áudio entre 10 e 30 segundos a partir dos pontos iniciais encontrados

Conclui que os instrumentos que compõem o acompanhamento atrapalham a identificação dos trechos de voz

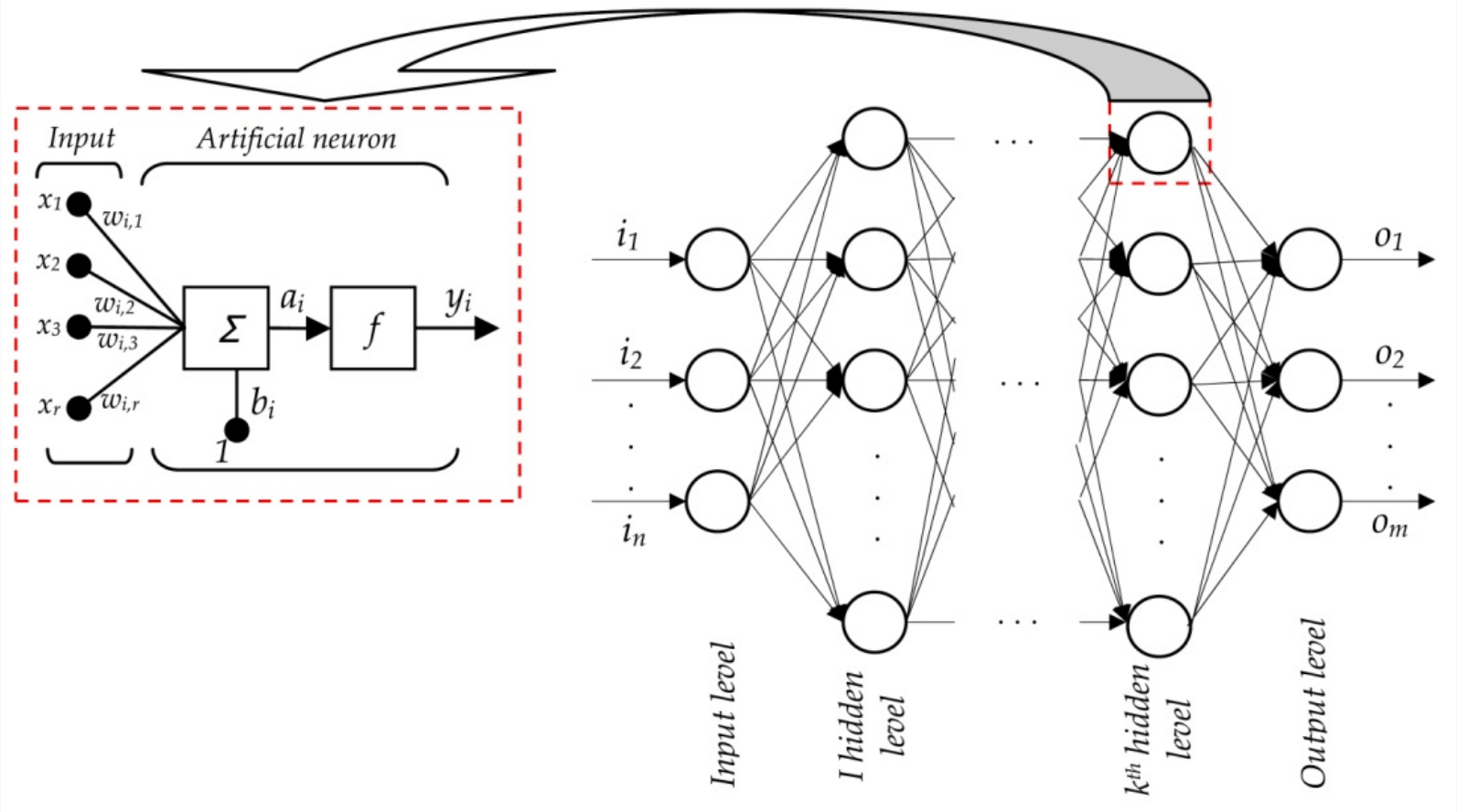
BERENZWEIG ET AL., 2002

Using Voice Segments to Improve Artist Classification of Music

- Usam Perceptual Linear Prediction (PLP), deltas e double deltas como descritores para acrescentar informação do contexto temporal
- Treinam uma **rede neural artificial** (ANN), cuja saída era um vetor de probabilidades posteriores de duas classes.
- Realizam uma **suavização** com filtro de média e comparam a um threshold

Concluem que a segmentação de trechos vozeados é importante para uma melhor performance na identificação de artista

ARTIFICIAL NEURAL NETWORKS



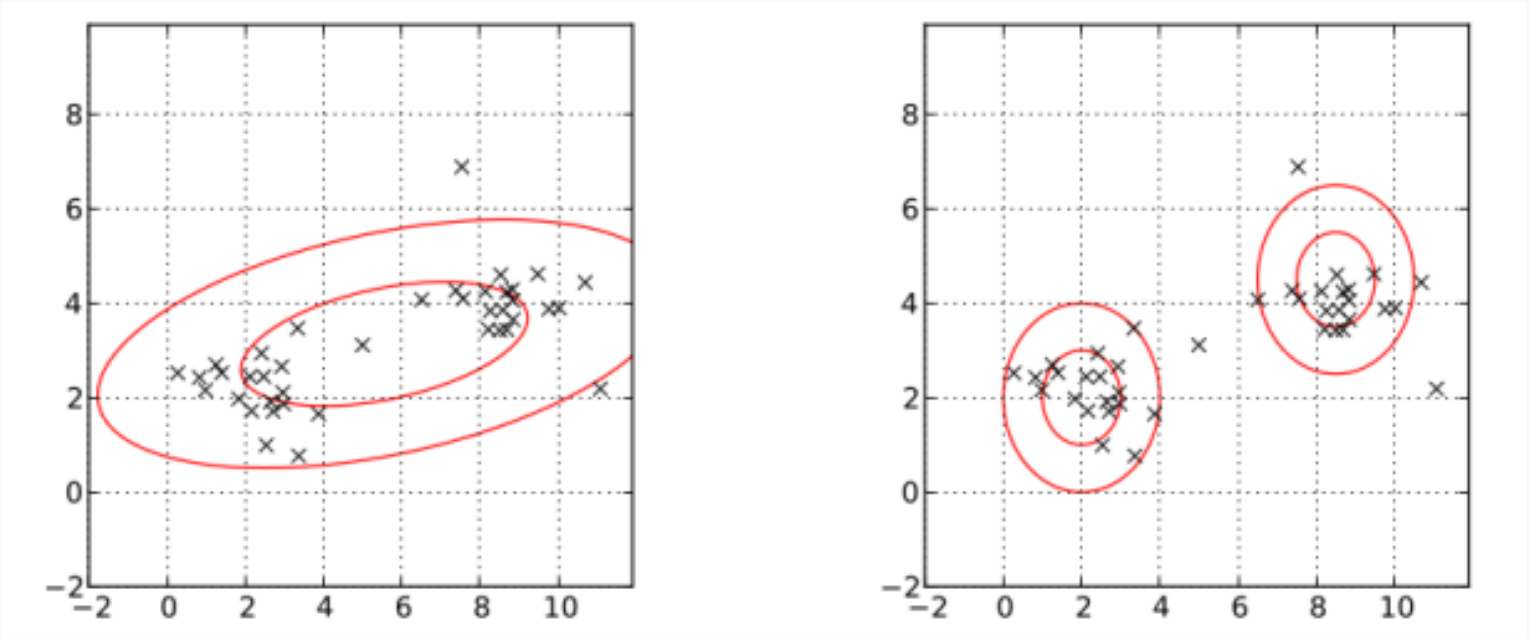
LI E WANG, 2007

Separation of Singing Voice from Music Accompaniment for Monaural Recordings

- Usam Mel-Frequency Cepstral Coefficients (**MFCCs**)
- Usam modelos de mistura gaussianos (**GMM**) como classificador
- Classificam os trechos de áudio e ajustam os resultados baseados na detecção de mudança espectral

Não há uma conclusão explícita sobre detecção de voz cantada, mas comparam com outros métodos de classificação e consideram melhor usar GMM

GAUSSIAN MIXTURE MODELS



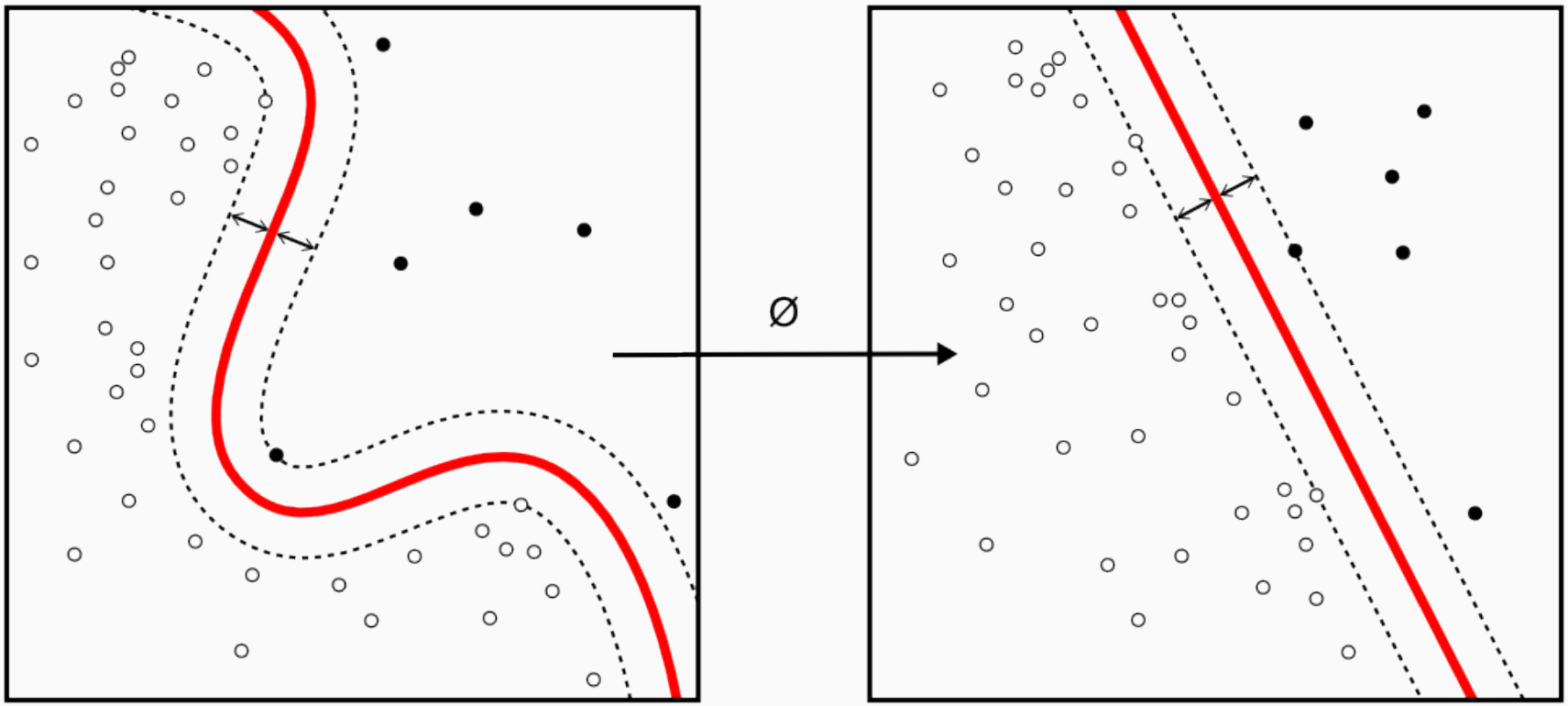
ROCAMORA E HERRERA, 2007

Comparing audio descriptors for singing voice detection in music audio files

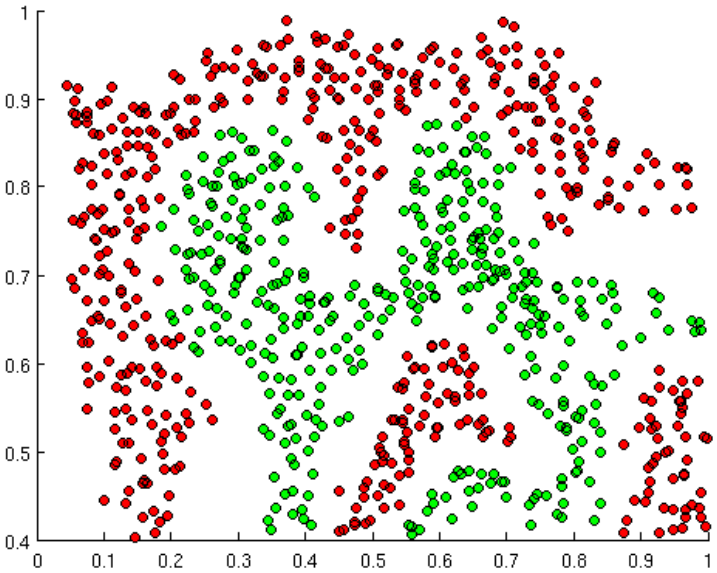
- Comparam **diversos descritores** de áudio utilizados para resolver esse problema
- Comparam **diferentes algoritmos** de classificação (SVM, KNN, ANN)
- Retreinam segmentos pouco prováveis, dividindo-os ao meio

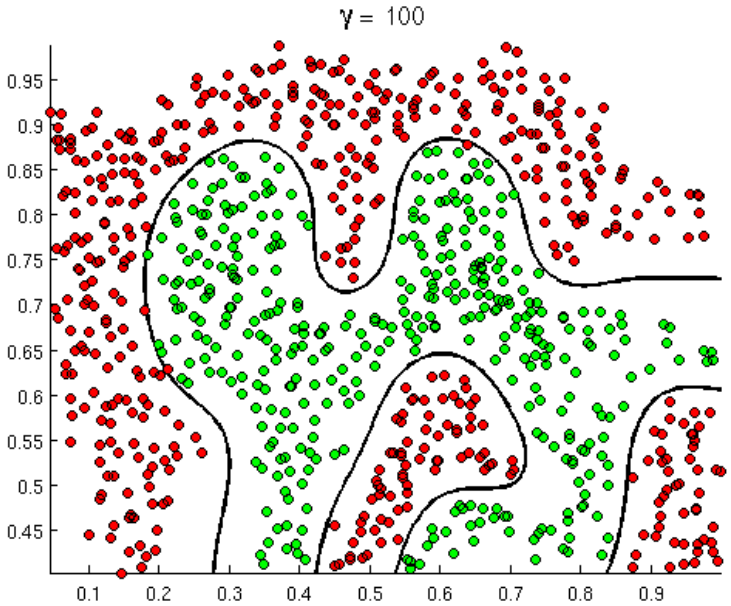
Concluem que entre os descritores mais representativos, o **MFCC** se destaca. Dentre os classificadores, o **SVM**.

SUPORT VECTOR MACHINES



SUPORT VECTOR MACHINES





REGNIER E PEETERS, 2009

Singing Voice Detection in Music Tracks using Direct Voice Vibrato Detection

- Usam características do timbre humano (**vibrato** e **tremolo**)
- Calculam um threshold para classificar os trechos de áudio
- Fazem um pós processamento removendo todos os trechos menores que 1 segundo

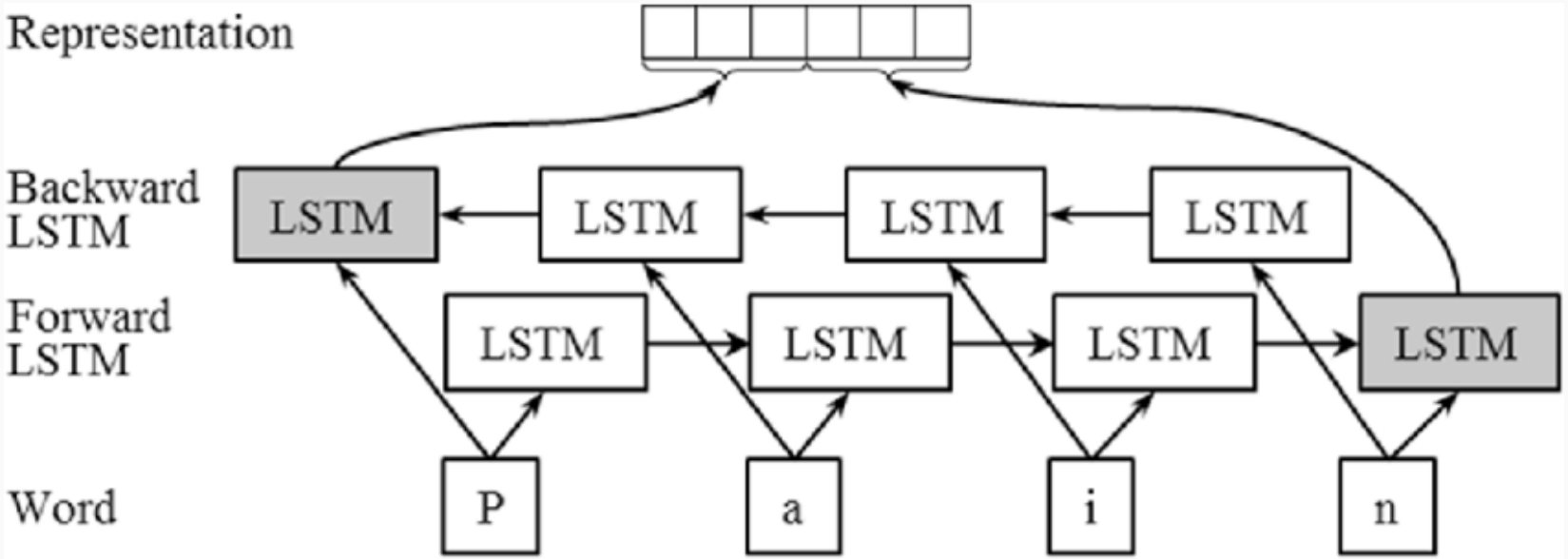
Concluem que o método possui resultados comparáveis com abordagens de aprendizado de máquina

LEGLAIVE ET AL., 2015

Singing voice detection with deep recurrent neural networks

- Usam descritores **MFCC** aplicando filtros para destacar partes harmônicas e percussivas
- Usam **Deep Recurrent Neural Network** com unidades **Long-Short Term Memory** para classificar os trechos cantados baseado em informações de contexto temporal

Concluem que conseguem classificar os trechos melhor que o estado da arte sem necessidade de processamento temporal posterior



Bidirectional LSTM.

HUMPHREY ET AL., 2017

Mining Labeled Data from Web-Scale Collections for Vocal Activity Detection in Music

- Buscam músicas com suas respectivas versões instrumentais para criar conjuntos de dados anotados automaticamente
- Classificam usando Redes Neurais Convolucionais

Ibrahim et al., 2017

Intelligibility of Sung Lyrics: A Pilot Study

- Tentam classificar automaticamente a inteligibilidade de uma música para ouvintes não nativos
- Classificam usando Random Forest

DESAFIOS

DESAFIOS EM DETECÇÃO DE VOZ CANTADA

Modelar classificadores utilizando redes neurais profundas

Aumentar o tamanho dos conjuntos de dados (Data Augmented)

Descobrir novos descritores discriminativos e significativos

Unificar os resultados das diferentes estratégias utilizadas

CONSIDERAÇÕES FINAIS

Os artigos tentam trazer métodos novos e melhores, porém não possuem uma base comum para comparação

A maioria das soluções apresentadas não são reprodutíveis e não estão disponíveis para teste e utilização.

Há uma tendência em usar algoritmos de aprendizagem sem se preocupar com o significado musical do que foi aprendido

Perguntas?