Reconhecimento Automático de Locutor em Ambientes Acusticamente Ruidosos: Desafios e Tendências

Rosângela Coelho e Leonardo Zão
Instituto Militar de Engenharia (IME), Rio de Janeiro, Brasil
Departamento de Engenharia Elétrica
coelho@ime.eb.br

Resumo

O reconhecimento automático de locutor (RAL) se refere ao processo de automaticamente se determinar ou verificar a identidade de um indivíduo através de sua fala. Sistemas de RAL têm ampla aceitação em diversas áreas e aplicações, tais como a autenticação de transações eletrônicas, ciência forense, segurança e controle de acesso. Um dos principais desafios para tais sistemas é a degradação de desempenho ocorrida quando o sinal de voz é capturado em ambientes acusticamente ruidosos. Este trabalho ressalta algumas das principais soluções propostas na literatura para tornar os sistemas de RAL robusto a ruídos acústicos.

Desafios

Apesar de apresentarem bons resultados para locuções limpas, com taxas de acertos para identificação que alcançam 98%-99% e EER (*Equal Error Rate*) de 1% para a verificação de locutor, os sistemas de RAL podem sofrer severa degradação de desempenho quando o sinal de voz é capturado em ambientes acusticamente ruidosos. As principais limitações são atribuídas à variabilidade, a não-estacionaridade, ao desconhecimento da origem e das características, temporais e espectrais, das fontes de ruídos acústicos ambientais (avião, trem, carro, arma de fogo, fábrica, sirenes) que podem corromper as locuções de voz.

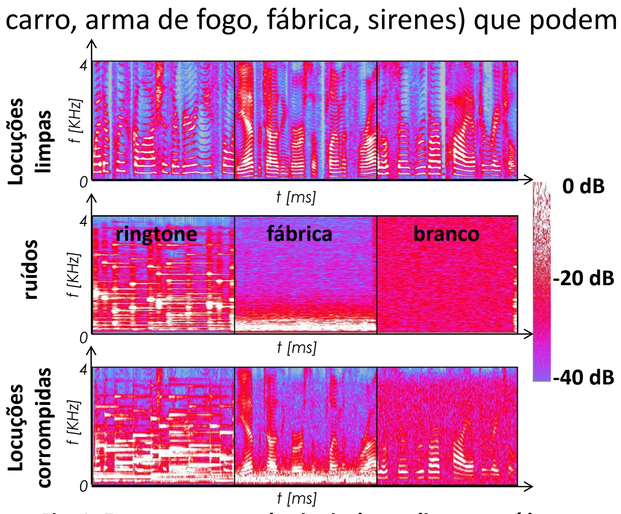

Fig. 1: Espectrogramas de sinais de voz limpos, ruídos, e os mesmos sinais de voz corrompidos.

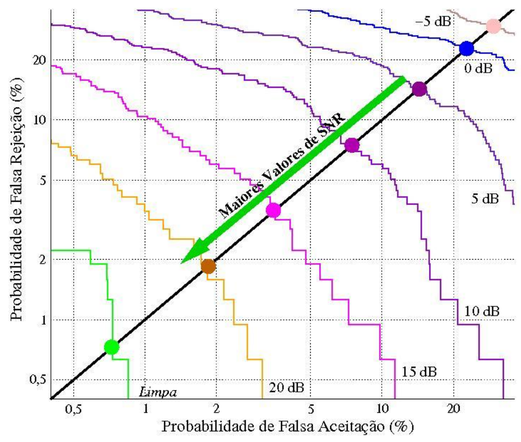

Fig. 2: Curvas DET (*detection error tradeoff*) com valores de EER destacados.

Reconhecimento Automático de Locutor

O reconhecimento automático de locutor engloba, fundamentalmente, duas tarefas ou funções: identificação e verificação. Na identificação, define-se a quem pertence a locução considerando-se um conjunto de modelos de locutores cadastrados no sistema. Na verificação ou autenticação de locutor, determina-se se a locução pertence ao locutor declarado. Um sistema de RAL envolve as fases de treinamento e teste e, geralmente, inclui as etapas de aquisição e pré-processamento do sinal de voz, extração de atributos da voz, modelagem do locutor e decisão ou classificação, propriamente dita.

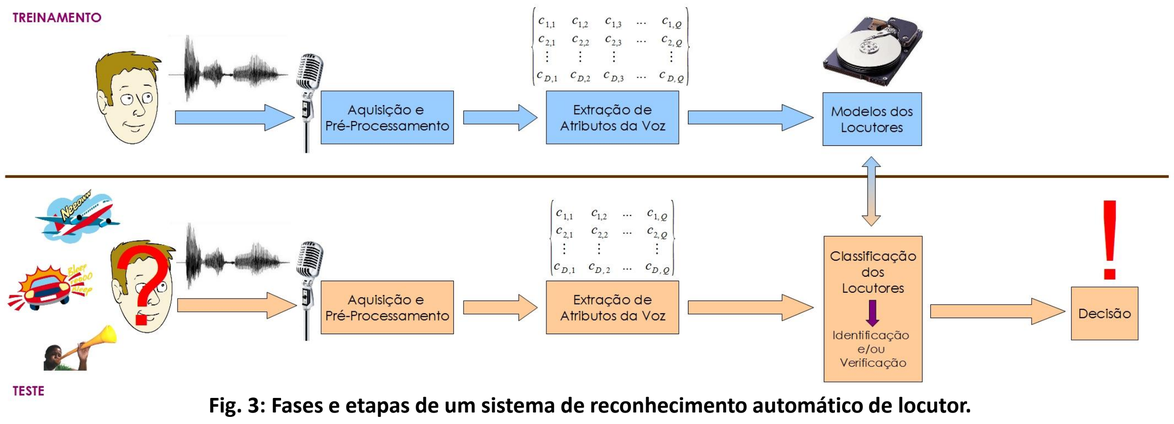

Fig. 3: Fases e etapas de um sistema de reconhecimento automático de locutor.

Pré-Processamento

As técnicas que atuam no pré-processamento têm como principal objetivo o aprimoramento ou compensação da razão sinal-ruído (*signal-to-noise ratio* - SNR) através da supressão ou cancelamento dos ruídos.

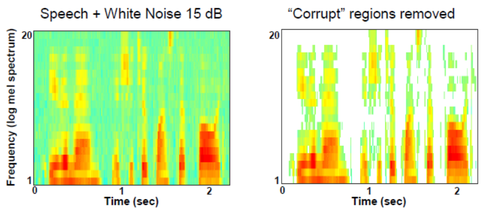

Fig. 4: Supressão acústica do ruído.

Principais técnicas:
- Arranjo de microfones com algoritmos de conformação de feixes (*beamforming*)
- Supressão acústica do ruído (*cepstral mean subtraction*)
- Filtragem RASTA (*relative spectral*)

A não-estacionaridade, mudanças abruptas, impulsividade, variabilidade e desconhecimento das características das fontes acústicas limitam o desempenho destas técnicas para prover a robustez necessária do sinal de voz a ruídos sonoros.

Atributos da Voz

Os atributos **MFCC** (*Mel-frequency cepstral coefficients*) permitem uma boa representação da característica vocal quando estes são extraídos de sinal de voz limpo (sem presença de ruído). No entanto, estes atributos não são robustos a ruídos acústicos.

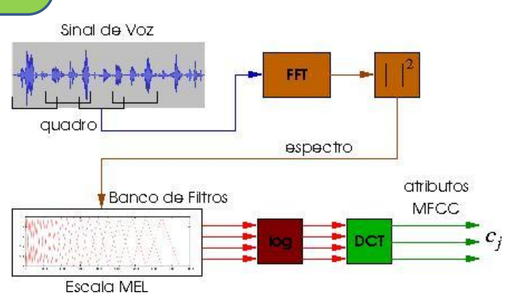

Fig. 5: Extração de atributos MFCC.

Técnicas que atuam nos atributos da voz:
- Análise linear discriminativa
(LDA - *linear discriminant analysis*)
- Análise de componentes principais
(PCA - *principal component analysis*)
- Descarte de atributos (*missing feature*)
- Moldagem de atributos (*feature warping*)

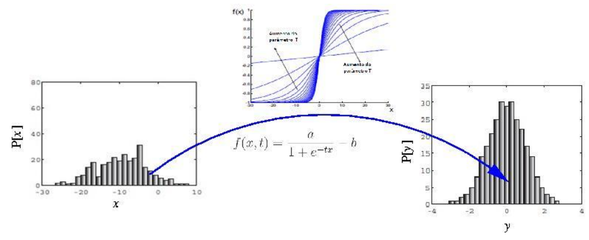

Fig. 6: Moldagem de atributos .

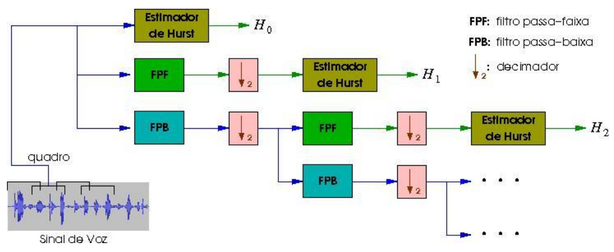

Fig. 7: Extração de vetores pH.

Atributos da voz mais robustos:
- AMFCC (*autocorrelation mel-frequency cepstral coefficients*)
- AGFCC (*auditory gammatone frequency cepstral coefficients*)
- Vetores pH (parâmetro de Hurst)

Modelos de Locutor

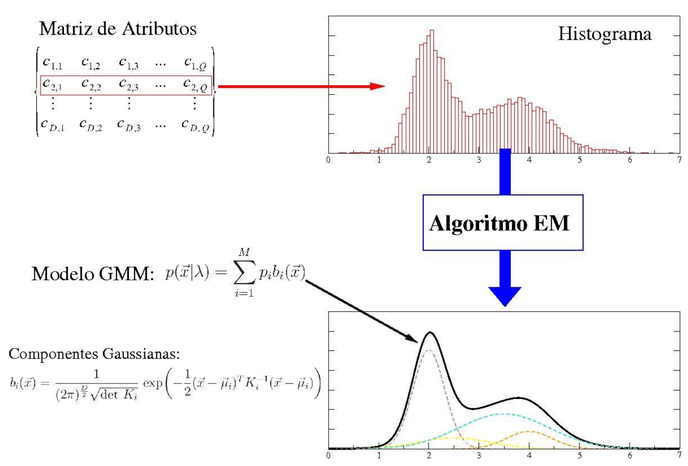

Fig. 8: O modelo de misturas gaussianas (GMM).

Principais classificadores:
- GMM (*gaussian mixture model*)
- GMM adaptado (A-GMM)
- α-GMM
- M-dim-fBm

Técnicas para prover robustez:
- Treinamento em múltiplas condições com ruído branco (TMCB)
- Treinamento em múltiplas condições com ruídos coloridos (TMCC)

Ruído	GMM	TMCB	TMCC
Limpo	**91,58%**	88,01%	88,27%
Avião	25,97%	39,90%	**42,35%**
Branco	25,97%	**57,24%**	55,92%
Carro	76,07%	70,15%	**83,42%**
Fábrica	46,99%	54,80%	**57,50%**
Tanque	52,24%	50,71%	**64,54%**

Tab. 1: Taxas médias de acertos de identificação de locutor (KING + NOISEX-92) para as técnicas TMCB e TMCC, além do GMM.

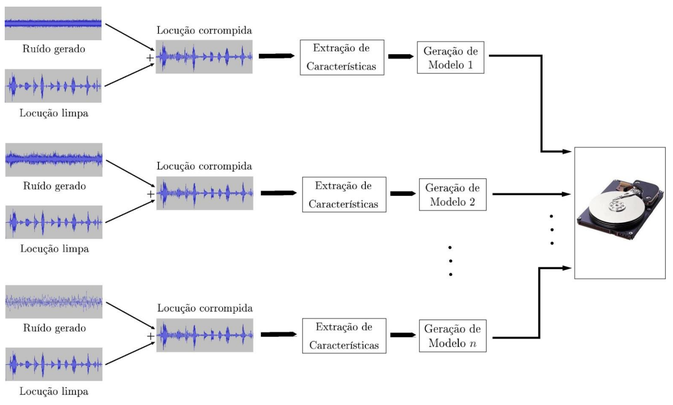

Fig. 9: Treinamento em múltiplas condições com ruídos de espectros coloridos (TMCC).

Bases de Voz e Ruídos

Principais bases de voz:
- KING
- TIMIT
- NTIMIT
- YOHO

Principais bases com ruídos:
- NOISEX-92
- AURORA
- SPINE
- MIT

Tendências

Acredita-se que a melhor solução para tornar sistemas de RAL robustos a ruídos seja multimodo devendo, portanto, englobar algumas ou todas as fases e etapas do reconhecimento de locutor. Um dos desafios é definição de quais as melhores técnicas para cada uma das etapas em um problema que não é universal.

Além disso, novos paradigmas devem ser avaliados, tais como a exploração de informações de alto nível, como, por exemplo, estados emocionais, uso de medidas de prosódica, inclusão de situações reais no sistemas de RAL, processamento do sinal de voz com foco na aplicação em RAL, e a caracterização temporal e espectral de fontes de ruídos acústicos. A proposta de novos atributos da voz (lineares e não-lineares) e classificadores robustos a ruídos acústicos é também um grande desafio e, talvez, uma ótima solução. Os desafios são muitos. O que torna a área de pesquisa bem interessante.

Referências Destaque

- R. Rose, E. Hofstetter e D. Reynolds, "Integrated Models of Signal and Background with Application to Speaker Identification in Noise", IEEE Trans. on Speech and Audio Processing, vol. 2, no. 2, 1994.
- J. Ming, T. Hazen, J. Glass e D. Reynolds, "Robust Speaker Recognition in Unknown Noisy Conditions," IEEE Trans. on Audio, Speech and Language Processing, vol. 15, no. 5, 2007.
- R. Sant'Ana, R. Coelho e A. Alcaim, "Text-Independent Speaker Recognition Based on the Hurst Parameter and the Multidimensional Fractional Brownian Motion Model", IEEE Trans. on Audio, Speech and Language Processing, vol. 14, no. 3, 2006.
- R. Sant'Ana, R. Coelho e A. Alcaim, "Automatic speaker verification based on fractional Brownian motion process", Electronics Letters, v. 40, 2004.
- L. Zão e R. Coelho, "Colored Noise Based Multicondition Training Technique for Robust Speaker Identification", IEEE Signal Processing Letters, vol. 18, no. 11, november 2011.