Processamento de Sinais Acústicos: Efeito de Distorções Não-Estacionárias na Classificação de Sinais de Voz

Rosângela Coelho, Leonardo Zão e Dirceu Cavalcante

Laboratório de Processamento de Sinais Acústicos

Instituto Militar de Engenharia (IME), Rio de Janeiro, Brasil

Departamento de Engenharia Elétrica

coelho@ime.eb.br

Resumo

Sistemas que empregam classificação por sinais de voz têm ampla aceitação em diversas áreas e aplicações, tais como a autenticação de transações eletrônicas, ciência forense, segurança e controle de acesso. Contudo, estes sistemas podem sofrer severa degradação de desempenho quando o sinal de voz apresenta distorções acústicas, tais como, ruídos e estados emocionais. Este efeito pode acarretar em redução de até 60% na acurácia da classificação dos sinais de voz. As principais limitações são atribuídas a variabilidade, a não-estacionaridade e ao desconhecimento das características temporais e espectrais das fontes de ruídos ambientais (avião, trem, carro, arma de fogo) e às variações acústicas emocionais (raiva, medo, tristeza), que afetam as locuções. Este trabalho ressalta algumas das principais soluções propostas na literatura para tornar os sistemas de classificação de voz robusto a estes efeitos 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 classificação podem sofrer severa degradação de desempenho quando o sinal de voz é capturado em ambientes acusticamente ruidosos. Esta degradação pode acarretar, por exemplo, em redução de até 60% na acurácia da identificação de locutor, dependendo da fonte de ruído. Considerando-se uma oscilação acústica provocada por um estado emocional, a taxa de classificação de locutor pode ser de 99,66% ou de 12,69% em sinais de voz neutro (sem efeito de emoção) ou sob efeito da emoção raiva, respectivamente. As principais limitações são atribuídas à variabilidade, à não-estacionaridade, ao desconhecimento da origem e das características, temporais e espectrais, das fontes de ruídos ambientais (avião, trem, carro, arma de fogo, fábrica, sirenes) e às variações acústicas emocionais (raiva, medo, tristeza) que afetam os sinais de voz.

Principais desafios:

- Distorções e oscilações acústicas distintas e não-estacionárias;
- Atributos acústicos representativos e robustos a distorções ou oscilações;
- Estimadores, extratores e filtros;
- Novas bases;
- Técnicas de processamento de sinais (realce, filtragem) apropriadas.

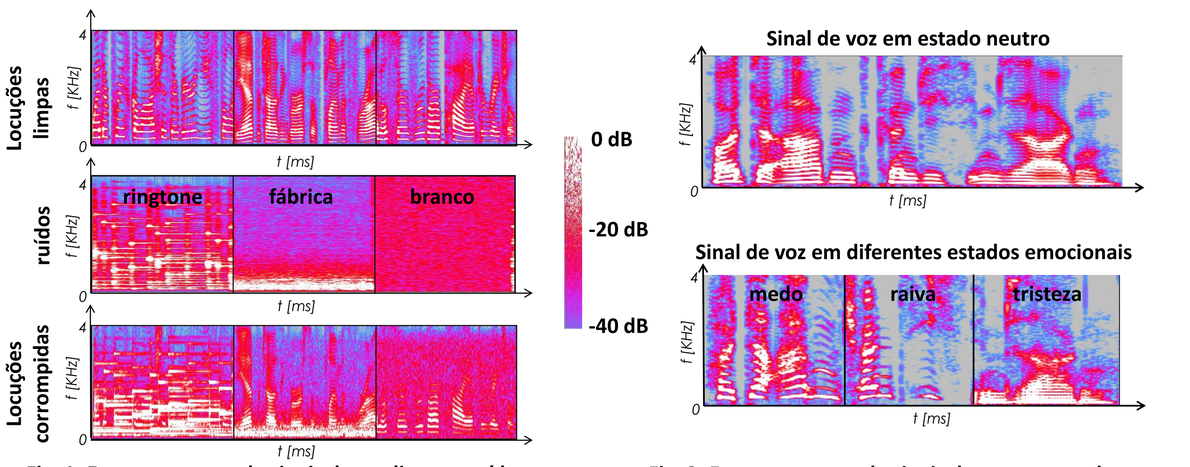

Fig. 1: Espectrogramas de sinais de voz limpos, ruídos, e os mesmos sinais de voz corrompidos.

Fig. 2: Espectrogramas de sinais de voz em estado neutro e distorcido por estados emocionais.

Pré-Processamento e Realce de Sinais

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.

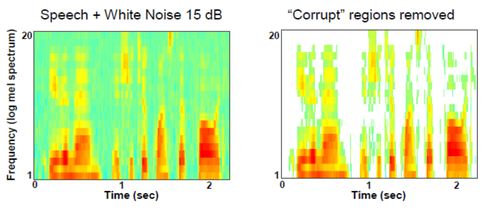

Fig. 3: Supressão acústica do ruído.

Técnicas clássicas:

- 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.

Soluções de realce de sinais para distorções não-estacionárias: tempo-frequência (TF)

- Estimação do espectro dos ruídos: IMCRA (*improved minima controlled recursive averaging*);
- Reconstrução do sinal de voz: OMLSA (*optimally-modified log-spectral amplitude*);
- Solução de pós-processamento utilizando o método EMD (*empirical mode decomposition*) para filtrar as componentes em baixas frequências.

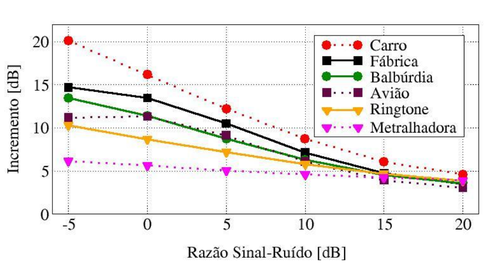

Fig. 4: Incrementos de razão sinal-ruído segmental obtidos pela técnica de realce com IMCRA, OMLSA e EMD, para diferentes fontes de ruídos acústicos.

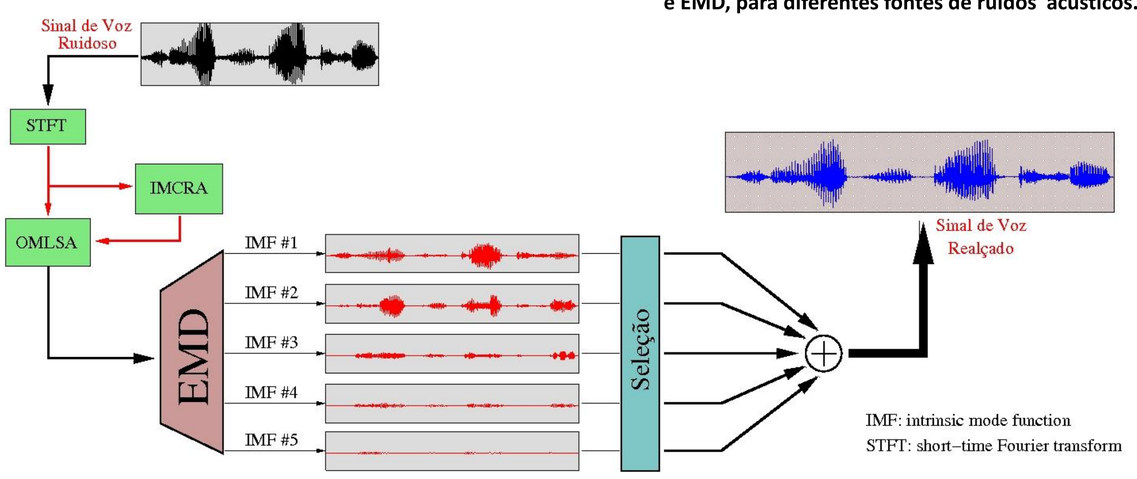

Fig. 5: Exemplo: Técnica de realce baseada no método EMD.

Atributos Acústicos

Atributos acústicos lineares e não-lineares:

- Vetores pH (parâmetro de Hurst)
- MFCC (*Mel-frequency cepstral coefficients*)
- TEO (*Teager Energy Operator*)
- GS (*Glottal Symmetry*)

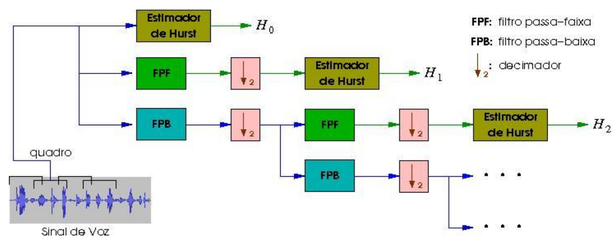

Fig. 6: Extração de vetores pH.

Fig. 7: 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*)

Modelos para Classificação

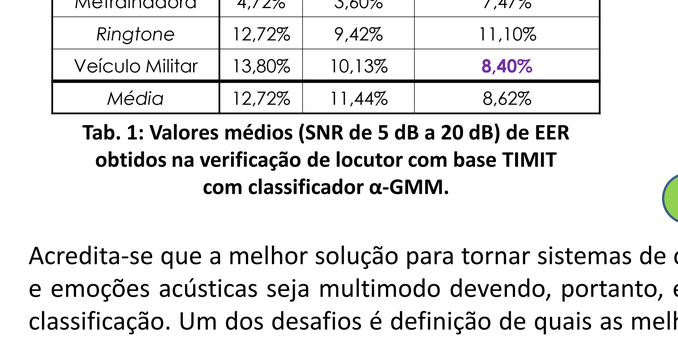

Fig. 8: Treinamento em múltiplas condições com ruídos de espectros coloridos (TMCC).

Principais classificadores de sinais acústicos para reconhecimento de locutor e emoções:

- GMM (*gaussian mixture model*)
- GMM adaptado (A-GMM)
- α-GMM
- M-dim-fBm
- SVM (*support vector machine*)

Técnicas para prover robustez:

- Treinamento em múltiplas condições com ruído branco (TMCB)
- Treinamento em múltiplas condições com ruídos coloridos (TMCC)

Bases: Voz, Ruídos e Emoções

Principais bases de voz:

- KING
- TIMIT
- NTIMIT
- YOHO

Principais bases com ruídos:

- NOISEX-92
- AURORA
- SPINE
- MIT

Principais bases com emoções:

- EMO-DB*
- SUSAS
- DES
- AVIC

Resultados

Ruído	MFCC	MFCC+pH	MFCC+pH c/ IMCRA e TMCC
Balbúrdia	10,63%	9,30%	9,93%
Fábrica	16,12%	12,25%	**6,84%**
Metralhadora	4,72%	3,60%	7,47%
Ringtone	12,72%	9,42%	11,10%
Veículo Militar	13,80%	10,13%	**8,40%**
Média	12,72%	11,44%	8,62%

Tab. 1: Valores médios (SNR de 5 dB a 20 dB) de EER obtidos na verificação de locutor com base TIMIT com classificador α-GMM.

Emoção	GS	pH
Raiva	51,63%	**60,80%**
Felicidade	51,42%	**83,20%**
Neutro	60,87%	**75,00%**
Tristeza	47,50%	**57,00%**
Média	52,86%	**69,00%**

Tab. 2: Acurácia de identificação de emoções utilizando base EMO-DB e classificador α-GMM.

Tendências

Acredita-se que a melhor solução para tornar sistemas de classificação de voz robustos a ruídos ambientais e emoções acústicas seja multimodo devendo, portanto, englobar algumas ou todas as fases e etapas de classificação. 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 sistema de reconhecimento, processamento do sinal de voz com foco na aplicação de reconhecimento de locutor, e a caracterização temporal e espectral de fontes de ruídos e emoções acústicas. Propostas de novos atributos (lineares e não-lineares, estacionários e não-estacionários) com seus extratores correspondentes e classificadores robustos às distorções acústicas, são também um grande desafio. Os desafios são muitos. O que torna a área de pesquisa bem interessante.

Referências Destaque

- ❑ R. Santana e R. Coelho, "Low-Frequency Ambient Noise Generator with Application to Automatic Speaker Classification", EURASIP Journal on Advances in Signal Processing, 2012.
- ❑ L. Zão e R. Coelho, "Colored Noise Based Multicondition Training Technique for Robust Speaker Identification", IEEE Signal Processing Letters, vol. 18, no. 11, 2011.
- ❑ 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.
- ❑ A. Iliev e M. Scordilis, "Spoken Emotion Recognition Using Glottal Symmetry", EURASIP Journal on Advances in Signal Processing, vol. 2011, 2011.
- ❑ G. Zhou, J. Hazen e J. Kaiser, "Nonlinear Feature Based Classification of Speech Under Stress", IEEE Trans. on Speech and Audio Processing, vol. 9, no. 3, 2011.
- ❑ 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.
- ❑ P. Flandrin, G. Rilling e P. Gonçalves, "Empirical Mode Decomposition as a Filter Bank", IEEE Signal Processing Letters, vol. 11, no. 2, 2004.
- ❑ I. Cohen, "Noise Spectrum Estimation in Adverse Environments: Improved Minima Controlled Recursive Averaging", IEEE Trans. On Speech and Audio Processing, vol. 11, no. 5, 2003.