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

R. Coelho, L. Zão e D. Cavalcante

***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.**

***Palavras-Chave*— classificação de sinais de voz, distorções acústicas, fontes de ruídos e emoções não-estacionárias.**

***Abstract*— Speech classification systems showed to be very interesting for many applications with security needs, such as access control, electronic bank transactions authentication, forensics and biomedical signal processing. Several studies and experiments confirmed that these systems achieve high recognition rates when considering clean or neutral speech signals. However, their performance can be severely degraded when the speech signals are affected by environmental noises (plane, train, car, siren, machine gun) or acoustic emotion (anger, fear, sadness). This paper highlights some solutions proposed in the literature that could improve the robustness of the speech classification systems to these acoustic distortions.**

***Keywords*— speech signal classification, acoustic distortion, non-stationary noise and emotion sources.**

I. INTRODUÇÃO

A crescente necessidade de sistemas de segurança com requisitos cada vez mais rígidos, impulsionaram o uso de identificação e autenticação baseada em sinais biométricos. A voz é uma das principais características biométricas dos seres humanos. Por ser o sinal acústico resultante do sistema de produção da fala, a voz possui informações que incluem a identidade, o sexo, o idioma e as condições físico-emocionais do locutor. Além da fala ser um dos mais importantes meios de comunicação do homem, o sinal de voz é de fácil aquisição.

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 [1] [2] [3]. Assim, sistemas de RAL têm ampla aceitação em diversas áreas e aplicações, tais como autenticação de transações eletrônicas, ciência forense, biomedicina, segurança e controle de acesso. No entanto, apesar apresentarem bons resultados para locuções limpas (sem distorção por ruídos ambientais) ou locuções neutras (sem efeito de estado emocional), estes sistemas de classificação podem sofrer severa degradação de desempenho quando o sinal de voz estiver corrompido por estas distorções ou variações acústicas.

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 [4] [5], determina-se se a locução pertence ao locutor declarado. A identificação é atribuída ao modelo do locutor que fornecer o maior valor de estimação da máxima probabilidade a posteriori (MAP). A acurácia da identificação é descrita pela razão entre o número de acertos e o número total de testes realizados (taxa de acertos). Na verificação, estima-se a EER (*equal error rate*) cujo valor é definido pelo ponto onde se equiparam as probabilidades de falta aceitação e falsa rejeição. Nos últimos anos, a curva DET (*Detection Error Trade-Off*) [6] é adotada como principal forma de apresentação e avaliação do desempenho dos sistemas de verificação de locutor.

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. Devido aos movimentos das articulações na produção da fala, o sinal sofre contínuas mudanças. Assim, a análise do sinal é realizada em tempo curto, ou seja, em quadros de cerca de 20ms-30ms de duração. Durante este intervalo, o sinal de voz é considerado estável (processo estacionário). Isso permite que vetores de atributos possam ser extraídos ou estimados, de cada quadro de voz. Este conjunto de vetores formará a matriz de atributos do locutor. Na fase de treinamento, os modelos dos locutores são obtidos das matrizes de atributos e armazenados no sistema. Estes modelos são utilizados na fase de teste para realização de uma das tarefas de classificação dos locutores.

Na literatura, os atributos de voz MFCC (*mel-frequency cepstral coefficients*) [7] [8] e o classificador GMM (*gaussian mixture model*), são considerados como a principal referência de bom desempenho em sistemas de RAL. Apesar de apresentarem bons resultados para locuções limpas [9], com taxas de acertos para identificação que alcançam 98%-99% e EER

R. Coelho está no Departamento de Engenharia Elétrica, Instituto Militar de Engenharia (IME). Os autores L. Zão e D. Cavalcante são doutorandos do Programa de Pós-Graduação em Engenharia de Defesa e orientados pela Prof. R. Coelho. Email: coelho@ime.eb.br.

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. Esta degradação pode acarretar, por exemplo, em redução de até 60% na acurácia da identificação de locutor [10] [11], dependendo da fonte de ruído. Considerando-se uma distorção 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 [12], 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. Nos últimos anos, este desafio têm impulsionado a área de pesquisa de processamento de sinais na busca por métodos que reduzam o descasamento entre as fases de treinamento e teste, provocado pela presença de ruídos sonoros. Os métodos propostos, geralmente, atuam em uma das seguintes etapas: pré-processamento, atributos da voz e modelo de representação do locutor.

Este trabalho apresenta algumas das principais soluções propostas na literatura, para tornar os sistemas de classificação de voz robustos a ruídos e emoções acústicas. As técnicas selecionadas estão apresentadas segundo a atuação nas distintas etapas do sistemas. Assim, no restante do trabalho, as seções se referem às soluções propostas para as fases de pré-processamento, atributos da voz e modelos estatísticos de locutor.

II. Pré-Processamento

As técnicas que atuam no pré-processamento têm como principal objetivo o aprimoramento ou compensação da razão sinal-ruído (SNR-*signal-to-noise ratio*) através da supressão ou cancelamento dos ruídos [13]. A maioria destas técnicas foi proposta como solução para o descasamento entre os canais de procedência do sinal de voz. Técnicas de arranjo de microfones com algoritmos de conformação de feixes (*beamforming*) [14] [15] foram empregadas para aprimorar o valor de SNR em sinais de voz capturados em ambientes de grandes dimensões e, também, em ambientes como escritórios, automóveis, entre outros. Estas técnicas não se aplicam aos sinais de voz capturados por canais telefônicos (fixos ou móveis) ou onde não haja conhecimento prévio do tipo de microfone utilizado na gravação do sinal de voz.

Técnicas de pré-filtragem baseada na supressão acústica do ruído (*cepstral mean subtraction*) [16] [17] [18] e filtragem RASTA (*relative spectral*) [19] necessitam da estimação prévia, e apurada, do espectro das fontes de ruído e do sinal de voz, quadro-a-quadro. Estas técnicas utilizam a estimação das estatísticas considerando os valores da densidade espectral de potência obtidos de quadros passados, e considerando que os ruídos são estacionários. A não-estacionaridade, mudanças abruptas, impulsividade, variabilidade e desconhecimento das características das fontes acústicas limitam o desempenho destas técnicas [20] [21] para prover a robustez necessária do sinal de voz a ruídos sonoros.

A. Técnicas de Realce para Ruídos Não-estacionários

Uma solução interessante para realçar os sinais de voz corrompidos por ruídos acústicos não-estacionários é a técnica IMCRA (*improved minima controlled recursive averaging*) [22]. Nesta proposta, a SNR de cada quadro é estimada considerando o menores valores de densidade espectral de potência obtida dos quadros imediatamente anteriores. Em seguida, a potência do ruído em cada componente de frequência é determinada a partir destes valores de SNR e considerando a probabilidade de ocorrência de voz nos quadros. A reconstrução do sinal de voz é realizada utilizando o método OMLSA (*optimally-modified log-spectral amplitude*) [23]. Apesar da técnica IMCRA não considerar a estacionariedade dos ruídos, a estimação torna-se imprecisa quando o ruído apresenta variações muito bruscas, principalmente nas baixas frequências.

Em [24], os autores apresentaram o método EMD (*empirical mode decomposition*) como uma forma empírica de decomposição de sinais não-estacionários. A filtragem EMDF foi proposta em [25] como uma solução de pós-processamento para, utilizando o método EMD, filtrar as componentes de baixas frequências residuais da técnica IMCRA. Embora, consigam aumentar os valores de SNR dos sinais de voz, a adoção das técnicas IMCRA e EMDF não garante, necessariamente, a melhoria na classificação de locutor [26].

III. Atributos da Voz

A literatura de RAL demonstra que os atributos MFCC [7] 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. Coeficientes dinâmicos ou coeficientes delta [27] [28], e suas derivações, são geralmente empregados para captar as variações dinâmicas entre quadros de voz, de forma a atenuar ou compensar condições espectrais que interfiram na integridade dos atributos de voz. Entretanto, os coeficientes delta não produzem melhora significativa para atributos extraídos de sinal de voz corrompido por ruído.

Nos métodos que atuam na matriz dos atributos da voz, pode-se ressaltar os de normalização ou ortogonalização, os de descarte de atributos (*missing feature*) [29] [30], os de moldagem de atributos (*feature warping*) [31] [32] [33] e os de seleção discriminativa de atributos [34].

A análise linear discriminativa (LDA - *linear discriminant analysis*) [35] e a análise de componentes principais (PCA - *principal component analysis*) [36] [37] são algoritmos de ortogonalização clássicos e foram adotados para sinal de voz limpa com uma pequena superioridade de desempenho para o LDA. No entanto, para sinais de voz corrompidos por diferentes fontes de ruído acústico e diversos valores de SNR [38], o desempenho das técnicas LDA e PCA, em média, é semelhante.

As técnicas de descarte de atributos [29] [30], propõem que os atributos da voz mais afetados pelos ruídos, sejam descartados e desconsiderados pelo sistema de reconhecimento. Estas técnicas consideram uma prévia subtração espectral e o acompanhamento do espectro, quadro-a-quadro,

para detecção de uma região onde haja uma possível alteração da integridade do sinal de voz pelo ruído. A remoção destes quadros corresponderia aos atributos eliminados. As limitações destas técnicas são as mesmas apresentadas pelas técnicas de remoção espectral de ruídos da fase de pré-processamento do sinal de voz. Ou seja, o prévio conhecimento da origem da fonte de ruído.

A principal função das técnicas de moldagem de atributos [31] [32] [33], é aproximar os padrões ou distribuições, representados pelos histogramas das matrizes de atributos, obtidos nas fases de treinamento e teste. E, com isso, diminuir o descasamento de condições entre estas fases. Geralmente, os algoritmos propostos na literatura, moldam uma distribuição genérica para uma distribuição gaussiana. A distinção entre as locuções de treinamento e teste e o não-conhecimento prévio das distribuições do sinal de voz (de cada locutor) e das fontes de ruído, tornam esta solução não-atraente. Para a moldagem é também necessário um pré-processamento do sinal de voz o que torna estas soluções complexas e custosas.

O método de seleção discriminativa de atributos proposto em [34], é realizado nas duas fases do sistema de RAL: treinamento e teste. O algoritmo utiliza a matriz de atributos extraída das locuções de treinamento e realiza testes de máxima verossimilhança ainda na fase de treinamento. Os atributos são classificados como sobrepostos (aqueles em que os atributos de um dado locutor apresentou maior valor de MAP para um outro locutor) e não-sobrepostos (maior valor de MAP para o vetor do próprio locutor). As novas matrizes compostas por vetores sobrepostos e não-sobrepostos, dão origem a dois novos modelos para cada locutor. Na fase de testes, cada vetor é confrontado com os dois modelos. Os vetores que apresentarem maior valor de máxima verossimilhança para o modelo não-sobreposto é selecionado para ser utilizado no reconhecimento propriamente dito. Em [38], este método foi utilizado em sinais de voz corrompidos por diversos ruídos e distintos valores de SNR. Os resultados mostram um significativo aprimoramento da robustez do sistema de RAL a ruídos acústicos. Esta é uma das únicas técnicas propostas exclusivamente como solução voltada para sistemas de RAL.

Há ainda sugestões de atributos de voz mais robustos, dentre as quais pode-se citar os AMFCC (*autocorrelation mel-frequency cepstral coefficients*) [39] onde os termos com correlação cruzada são zerados ou ignorados pelo modelo e os AGFCC (*auditory gammatone frequency cepstral coefficients*) [40] [41] que extraídos com filtros gammatone, captam a faixa de frequência do aparelho auditivo humano (cóclea).

Diferentemente dos atributos espectrais, os vetores pH (parâmetro de Hurst) [42] [43] [44] foram propostos como atributos temporais-estatísticos. Assim, estes são estimados diretamente do sinal de voz e representam a soma dos coeficientes de correlação das amplitudes do sinal de voz. Os atributos pH são obtidos dos coeficientes de detalhes resultante da filtragem wavelets [45] do sinal de voz. Em [46], os vetores pH foram avaliados em diversos experimentos de verificação de locutor submetidos a diversas fontes de ruídos ambientais, estacionárias e não-estacionárias, e diferentes níveis de SNR. Os resultados demonstraram que, a fusão dos atributos pH e MFCC, aprimora a robustez dos sistemas de RAL. A utilização dos vetores pH resultou em uma redução média de cerca de dois pontos percentuais nos valores de EER, e as melhorias mais significativas foram obtidas nas condições mais severas de ruídos.

IV. Modelos de Locutor

Além do clássico GMM, outros modelos foram propostos para serem aplicados na classificação de locutores, tais como o GMM adaptado (A-GMM) [4] [5], o α-GMM [47] e o M-dim-fBm [42]. Os modelos A-GMM, α-GMM e M-dim-fBm podem obter resultados superiores aos do GMM, em situações de descasamento de condições. Estes modelos são também utilizados para a representação dos impostores ou UBM (*Universal Background Model*) na tarefa de verificação.

Para a verificação de locutor, pode-se citar ainda os métodos de normalização ou compensação Hnorm [18], Znorm, Tnorm, Cnorm e Dnorm [5] cujo objetivo é remover uma possível tendência dos valores (*scores*) de máxima verosimilhança obtidos entre os modelos locutores e entre os modelos dos locutores e do intruso. Esta tendência pode ocorrer devido ao uso de um tipo de microfone, um canal ou mesmo de uma fonte de ruído, que são incorporados ao sinal de voz. Isso dificultaria a discriminação entre os locutores. Estes valores são utilizados na definição dos limiares de decisão da tarefa de verificação. Apesar dos resultados de desempenho do RAL não terem sido muito significativos para voz limpa, estes métodos podem contribuir no caso de sinal de voz com presença ou adição de ruído.

A classificação de locutor cuja a geração de modelos é obtida a partir de sub-bandas do sinal da voz [48], foi proposta para ser mais robusta do que a classificação em banda plena. Para isso, as sub-bandas mais afetadas pelo descasamento são desconsideradas ou atenuadas na geração do modelo de locutor e, desta forma, obtém-se o aprimoramento do reconhecimento. Os resultados podem ser promissores mas a principal restrição é semelhante a indicada pelas técnicas de pré-processamento. No entanto, pode ser uma boa alternativa onde tem-se o prévio conhecimento das fontes acústicas.

O método *factor analysis* (FA) [49] propõe uma representação ou modelagem dupla do locutor para a tarefa de verificação utilizando o GMM. Nesta nova representação, é incluída também o modelo das variações do canal. Ou seja, o modelo do locutor é dependente do canal. Os resultados mostraram uma diminuição interessante no valor de EER, quando comparado ao sistema tradicional. O método adota uma prévia ortogonalização da matriz de atributos utilizando o LDA. O estudo do FA com integração da variabilidade das fontes de ruídos, para uma nova representação do locutor em ambientes ruidosos, pode ser interessante para o alcance da robustez.

A. Classificação Acústica de Emoções

A classificação acústica de emoções [50] [51] tem motivado o desenvolvimento de diversas aplicações que dependam da percepção do estado emocional dos indivíduos. Por exemplo, sistemas de emergência, sistemas de segurança, diagnósticos de doenças, pilotos em cabine de avião, entre outras.

O efeito dos estados emocionais induzem alterações nos mecanismos de produção da fala através de mudanças na tensão muscular, fluxo respiratório, batimento cardíaco e pressão arterial que podem ser transientes ou de efeito prolongado. Emoções como raiva, por exemplo, levam a um estado exaltado, onde há um aumento no batimento cardíaco, respiração e energia no sinal de voz, através da excitação do sistema nervoso simpático. Este efeito acústico é conhecido como emoção de alta ativação. Por outro lado, emoções como tristeza levam a um estado reprimido, oposto à raiva, através da excitação do sistema nervoso parassimpático, sendo então chamada de emoções de baixa ativação. Do ponto de vista psicológico, emoções podem ser classificadas quanto à experiência do locutor ou valência. Desta forma, emoções como felicidade proporcionam conforto ao locutor (valência positiva), enquanto medo, o efeito contrário (valência negativa). A classificação acústica de emoções utilizando os conceitos de ativação e valência é uma abordagem frequente na literatura. Com esta abordagem pode-se alcançar taxas de classificação de até 90%. No entanto, a classificação de múltiplas emoções e de sinais de voz sob estes efeitos não-estacionários ainda representa um grande desafio.

A proposta de novos atributos acústicos específicos para a representação destas distorções ou variações emocionais, é fundamental para a pesquisa na área de processamento de sinais acústicos. A Simetria Glotal [52] [12], baseada na fonte de excitação da voz, e o operador Teager [50], baseado no mapeamento do deslocamento em frequência dos harmônicos da *pitch*, foram avaliados para a classificação das alterações no trato vocal induzidas pelas emoções. Outro forte desafio para esta área de pesquisa, corresponde às bases de dados. Além de serem bases prototípicas (emoções obtidas por atores ou por atuação) e em idiomas que não o português, poucas estão disponíveis publicamente. Também não há uma padronização acústica para a obtenção das fontes de emoções, sendo usualmente repetições de frases curtas ou segmentos de trechos de voz. Os modelos estatísticos utilizados na representação das emoções são, geralmente, os mesmos aplicados na classificação dos sinais de voz. Em [12], os resultados mostram que a identificação de locutor em sinal neutro ou sem efeito de emoção acústica, pode alcançar taxas de até 99,66%. Mas, considerando um sinal de voz com variações acústica como, por exemplo, a da emoção raiva, a taxa de identificação é 12,67%.

B. Treinamento em Múltiplas Condições

Uma técnica interessante, onde o modelo do locutor é obtido através de treinamento em múltiplas condições (TMC) ou situações, foi proposta, inicialmente, para prover robustez diante de situações de descasamento para reconhecimento de voz [53] [54]. Em [54], o treinamento de cada locutor é realizado com locuções limpas e corrompidas por fontes de ruídos ambientais reais em diversos valores de SNR. Contudo, esta solução só se aplica quando há conhecimento prévio da fonte de ruído que corromperá as locuções de teste, como, por exemplo, em ambiente fechado e controlado acusticamente.

Recentemente, a técnica TMC foi examinada para tornar os sistemas de RAL robustos a ruídos ambientais [10] quando não é conhecida a fonte de ruído. Nesta nova proposta, as locuções de cada locutor, são replicadas e corrompidas com ruído artifical gaussiano branco (TMCB) em diferentes valores de SNR. Os diversos modelos dos locutores são obtidos das locuções limpas e das resultantes da adição do ruído branco gerado artificialmente. O ruído branco foi adotado como solução para as situações onde não haja informações suficientes sobre a origem e/ou as características estatísticas das fonte dos ruídos presentes nas locuções de testes.

Outra solução [11] [55] para o treinamento em múltiplas condições, baseia-se em amostras de ruídos geradas artificialmente e com espectros coloridos (TMCC). Nesta técnica é considerado um único valor de SNR. Os resultados mostraram que a solução TMCC aprimorou a robustez da identificação de locutor em média de 30% e de 3,75%, quando comparado ao treinamento sem múltiplas condições e ao TMCB, respectivamente. Este desempenho da técnica TMCC foi obtido sem considerar qualquer pré-processamento prévio do sinal ou técnica de descarte de atributos. Logo, espera-se um aprimoramento quando esta for conjugada com as demais técnicas.

V. Bases de Voz, Ruídos Ambientais e Emoções Acústicas

Para a avaliação dos sistemas de RAL em ambientes ruidosos, nos experimentos são adotadas bases de ruídos acústicos ambientais. Os ruídos são utilizados como aditivos ao sinal de voz (adição eletrônica). Há ainda bases de voz cujas locuções são gravadas em ambientes reais, tais como ruas, escritórios ou lojas (adição acústica). A base de ruídos NOISEX-92 [56] é constituída de 15 fontes acústicas de distintas origens, possui longa duração de gravação e está disponível de forma pública e gratuita. Por consequência, esta base é adotada na maioria das pesquisas e experimentos da área. As gravações dos ruídos foram executadas em ambientes sonoros (avião, carro, fábrica, balbúrdia, arma de fogo, entre outras) e possuem 235 segundos de duração com taxa de amostragem de 19,98 KHz.

A base AURORA [54] é composta pelas locuções da base TIDigits [57] adicionadas, eletronicamente, a 8 fontes de ruídos acústicos ambientais reais em 6 diferentes valores de SNR. A base de voz SPINE [58] é composta de conversações livres entre pares de locutores gravadas em ambiente fechado submetido a diversos ruídos acústicos de natureza militar. A base SPINE é dividida em quatro subconjuntos, o maior deles composto de 40 locutores. A base de voz MIT [59] foi criada, especificamente, para verificação de locutor e foi gravada em 3 ambientes acusticamente ruidosos. Esta base possui 48 locutores para serem cadastrados no sistema, e um grupo de 40 impostores. As sessões de conversação possuem 20 minutos de duração por locutor.

Entre as principais bases de voz utilizadas nos experimentos com adição eletrônica de ruídos, pode-se ressaltar: KING, TIMIT, NTIMIT e YOHO. Estas bases, entre outras geradas pelo NIST (*National Institute of Standards and Technology*), são disponibilizadas, publicamente, pelo LDC (*Linguistic Data Consortium*, disponível em: ldc@ldc.upenn.edu) contudo, não são gratuitas. A base KING é composta de 51 locutores do sexo masculino. Cada locutor gravou 10 sessões de

conversação com duração entre 30 s e 60 s. O intervalo entre gravações do mesmo locutor, varia entre uma semana e um mês. Devido a uma mudança no equipamento durante a gravação, utiliza-se, geralmente, apenas 5 sessões para testes de reconhecimento de locutor. Além da gravação local, a base KING também está disponível após a transmissão de suas locuções por um canal telefônico de longa distância. A base TIMIT [60] é composta de 630 locutores de ambos os sexos. Cada uma das 10 locuções gravadas por cada locutor possui aproximadamente 3 s de duração. As sentenças faladas foram escolhidas de forma a apresentar grande variabilidade fonética. A base NTIMIT [61] possui as mesmas locuções utilizadas nas gravações da base TIMIT, mas captadas após as suas transmissões por canais telefônicos de curtas e longas distâncias. A base YOHO [62] foi obtida, inicialmente, para experimentos de verificação de locutor e é composta por 138 locutores. As 10 sessões de gravação de cada locutor foram captadas em ambiente de escritório por um *handset* telefônico, possuindo banda limitada entre 120 Hz e 3800 Hz.

As principais bases acústicas de emoções utilizadas na literatura são a *Berlin Emotional Speech Database* (EMO-DB) que apresenta gravações de 494 frases em alemão (pública e gratuita) e a *Speech Under Simulated and Actual Stress* (SUSAS) em inglês, disponível comercialmente pelo LDC. Em [63], é apresentada uma descrição detalhada das diferentes bases acústicas de emoções.

VI. Tendências

Este trabalho apresentou algumas das principais técnicas que podem atuar como agentes para tornar os sistemas de classificação de voz robustos a ruídos ambientais e emoções acústicas. Apesar das diversidade das soluções propostas, que atuam nas diferentes fases, etapas e situações, ainda não há uma solução única capaz de atingir taxas de reconhecimento semelhantes às obtidas com sinal de voz limpo ou neutro.

Acredita-se que a melhor solução deste problema complexo e não-universal, seja multimodo devendo, portanto, englobar algumas ou todas as fases e etapas da classificação. Um dos desafios é a definição de quais as melhores técnicas para cada uma das etapas em um problema que não é universal. As aplicações baseadas em reconhecimento de locutor e de emoções apresentam requisitos específicos para a obtenção de bom desempenho [64] e estes devem ser considerados na escolha da melhor solução.

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 [65], inclusão de situações reais no sistemas de reconhecimento, processamento acústico do sinal com foco na aplicação de reconhecimento de locutor, e a caracterização temporal e espectral [66] [11] 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

[1] D. O'Shaughnessy, "Speaker recognition," *IEEE Acoustics, Speech and Signal Processing Magazine*, vol. 3, pp. 4–17, October 1986.

[2] J. Campbell, "Speaker recognition: A tutorial," *Proceedings of the IEEE*, vol. 85, pp. 1437–1461, September 1997.

[3] G. Doddington, "Speaker verification - identifying people by their voices," *Proceedings of the IEEE*, vol. 73, pp. 1651–1664, November 1985.

[4] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, January 2000.

[5] F. Bimbot, J. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacretaz, and D. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Processing*, vol. 4, pp. 430–451, 2004.

[6] A. Martin et al, "The det curve in assessment of detection task performance," *Proceedings of EuroSpeech 97*, pp. 1895–1898, 1997.

[7] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acousticd, Speech, and Signal Processing*, vol. 28, pp. 357–366, July 2007.

[8] S. Imai, "Cepstral analysis synthesis on the mel frequency scale," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1983)*, vol. 8, pp. 93–96, April 1983.

[9] D. A. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," *Speech Communication*, vol. 17, pp. 91–108, August 1995.

[10] J. Ming, T. Hazen, J. Glass, and D. Reynolds, "Robust speaker recognition in noisy conditions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 1711–1723, July 2007.

[11] L. Zão, "Reconhecimento automático de locutor robusto a ruídos acústicos ambientais de espectros coloridos," *Dissertação de Mestrado - Instituto Militar de Engenharia (IME)*, 2010.

[12] D. Cavalcante and R. Coelho, "Atributos acústicos baseados na simetri glotal e no classificador α-gmm para identificação de emoções e locutor," *Anais do XXX Simpósio Brasileiro de Telecomunicações (SBrT 2012)*, pp. 1–5, September 2012.

[13] T. Quatieri, *Discrete-time Speech Signal Processing: Principles and Practice*. Prentice Hall, 2002.

[14] J. Flanagan, J. Johnston, R. Zahn, and G. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," *Journal Acoust. Soc. Amer.*, vol. 78, no. 5, pp. 1508–1518, 1985.

[15] S. Fischer and K. Simmer, "Beamforming microphone arrays for speech acquisition in noisy environments," *Speech Communication*, vol. 20, no. 3-4, pp. 215–227, 1996.

[16] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, pp. 113–120, April 1979.

[17] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 29, pp. 259–272, April 1981.

[18] R. Mammone, X. Zhang, and R. Ramachandran, "Robust speaker recognition: A feature-based approach," *IEEE Signal Processing Magazine*, vol. 13, pp. 58–71, September 1996.

[19] H. Hernansky et al, "Rasta-plp speech analysis technique," *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP92)*, pp. I.121–I.124, March 2002.

[20] K. Manohar and P. Rao, "Speech enhancement in nonstationary noise environments using noise properties," *Speech Communication*, vol. 48, pp. 96–109, January 2006.

[21] Y. Gong, "Speech recognition in noisy environments: a survey," *Speech Communication*, vol. 16, pp. 261–291, April 1995.

[22] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 466–475, September 2003.

[23] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, no. 11, pp. 2403–2418, 2001.

[24] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N. C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 454, pp. 903–995, March 1998.

[25] N. Chatlani and J. Soraghan, "Emd-based filtering (emdf) of low-frequency noise for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, pp. 1158–1166, May 2012.

[26] L. Zão and R. Coelho, "Realce de sinais de voz em presença de ruídos acústicos não-estacionários utilizando o método emd," *Anais do XXX Simpósio Brasileiro de Telecomunicações (SBrT 2012)*, pp. 1–5, September 2012.

[27] K. Soong and E. Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 36, pp. 871 – 879, June 1988.

[28] D. Reynolds and R. Rose, "Robust text independent speaker identification using gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 72–83, January 1995.

[29] A. Drygajlo and M. El-Maliki, "Speaker verification in noisy environments with combined spectral subtraction and missing feature theory," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1998)*, vol. 1, pp. 121–124, May 1998.

[30] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication*, vol. 34, pp. 267–285, June 2001.

[31] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," *2001: A Speaker Odyssey - The Speaker Recognition Workshop*, pp. 1–6, June 2001.

[32] B. Xiang, U. Chaudhari, J. Navratil, G. Ramaswamy, and R. Gopinath, "Short-time gaussianization for robust speaker verification," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2002)*, pp. I: 681–684, May 2002.

[33] Y. Xie, B. Dai, and J. Sun, "Kurtosis normalization after short-time gaussianization for robust speaker verification," *Proceedings of the 6th World Congress on Intelligent Control and Automation (WCICA 2006)*, vol. 2, pp. 9463–9467, June 2006.

[34] S. Kwon and S. Narayanan, "Robust speaker identification based on selective use of feature vectors," *Pattern Recognition Letters*, vol. 28, pp. 85–89, January 2007.

[35] Q. Jin and A. Waibel, "Application of lda to speaker recognition," *Proceedings of the International Conference on Spoken Language Processing (ICSLP 2000)*, pp. 250–253, October 2000.

[36] I. Magrin-Chagnolleau, G. Durou, and F. Bimbot, "Application of time-frequency principal component analysis to text-independent speaker identification," *IEEE Transactions on Speech and Audio Processing*, vol. 10, pp. 371–378, September 2002.

[37] C. Seo, K. Y. Lee, and J. Lee, "GMM based on local PCA for speaker identification," *Electronics Letters*, vol. 37, pp. 1486–1488, November 2001.

[38] R. Santana, "Identificação automática e robusta de locutor utilizando seleção discriminativa de atributos da voz," *Dissertação de Mestrado - Instituto Militar de Engenharia (IME)*, 2011.

[39] B. Shannon and K. Paliwal, "Mfcc computation from magnitude spectrum of higher lag autocorrelation coefficients for robust speech recognition," *Proceedings of the International Conference on Spoken Language Processing (ICSLP 2004)*, pp. 129–132, October 2004.

[40] Y. Shao and D. Wang, "Robust speaker identification using auditory features and computational auditory scene analysis," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2008)*, pp. 1589–1592, April 2008.

[41] Y. Zhang and W. H. Abdulla, "Robust speaker identification in noisy environment using cross diagonal GTF-ICA feature," *Proceedings of the 6th International Conference on Information, Communications and Signal Processing (ICICS 2007)*, pp. 1–4, December 2007.

[42] R. Sant'Ana, R. Coelho, and A. Alcaim, "Text-independent speaker recognition based on the hurst parameter and the multidimensional fractional brownian motion model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 931–940, May 2006.

[43] R. Sant'Ana, A. Alcaim, and R. Coelho, "Automatic speaker verification based on fractional brownian motion process," *Electronics Letters*, vol. 40, pp. 1232–1233, September 2004.

[44] L. Zao, A. Alcaim, and R. Coelho, "Robust access based on speaker identification for optical communications security," *Proceedings of the 16th International Conference on Digital Signal Processing (DSP 2009)*, pp. 1–5, July 2009.

[45] P. Flandrin, "Wavelet analysis and synthesis of fractional brownian motion," *IEEE Trans. on Information Theory*, vol. 38, pp. 910–917, March 1992.

[46] L. Zão and R. Coelho, "Noise Robust Speaker Verification based on the MFCC and pH Features Fusion and Multicondition Training," *Proceedings of the International Conference on Bio-inspired Systems and Signal Processing (BIOSIGNALS 2012)*, pp. 137–143, February 2012.

[47] D. Wu, J. Li, and H. Wu, "α-gaussian mixture modelling for speaker recognition," *Pattern Recognition Letters*, vol. 30, pp. 589–594, April 2009.

[48] H. Bourlard and S. Dupont, "A new asr approach based on independent processing and recombination of partial frequency bands," *Proceedings of the International Conference on Spoken Language Processing (ICSLP 1996)*, vol. 1, pp. 426–429, October 1996.

[49] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Font-end factor analysis for speaker verification," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, pp. 788–798, May 2011.

[50] G. Zhou, J. Hansen, and J. Kaiser, "Nonlinear feature based classification of speech under stress," *IEEE Transactions on Speech and Audio Processing*, vol. 9, pp. 201–216, March 2001.

[51] M. El Ayadi, M. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, pp. 572–587, March 2011.

[52] A. Iliev and M. Scordilis, "Spoken emotion recognition using glottal symmetry," *EURASIP Journal on Advances in Signal Processing*, vol. 2011, pp. 1–11, January 2011.

[53] R. Lippmann, E. Martin, and D. Paul, "Multi-style training for robust isolated-word speech recognition," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP87)*, vol. 12, pp. 705–708, April 1987.

[54] D. Pearce and H. Hirsch, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," *Proceedings of the International Conference on Spoken Language Processing (ICSLP 2000)*, vol. 4, pp. 29–32, October 2000.

[55] L. Zão and R. Coelho, "Colored noise based multicondition training technique for robust speaker identification," *IEEE Signal Processing Letters*, vol. 18, pp. 675–678, November 2011.

[56] A. Varga and H. Steeneken, "Assessment for automatic speech recognition ii: NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, pp. 247–251, July 1993.

[57] R. Leonard, "A database for speaker-independent digit recognition," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP84)*, vol. 9, pp. 328–331, March 1984.

[58] B. Kingsbury, G. Saon, L. Mangu, M. Padmanabhan, and R. Sarikaya, "Robust speech recognition in noisy environments: The 2001 ibm spine evaluation system," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2002)*, pp. I: 53–56, May 2002.

[59] R. Woo, A. Park, and T. Hazen, "The MIT mobile device speaker verification corpus: Data collection and preliminary experiments," *Proceedings of Odyssey 2006, The Speaker and Language Recognition Workshop*, pp. 1–6, June 2006.

[60] W. Fisher, R. Doddington, and M. Goudie-Marshall, "The darpa speech recognition research database: Specifications and status," *Proceedings of the DARPA Workshop on Speech Recognition*, pp. 93–99, February 1986.

[61] C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz, "Ntimit: a phonetically balanced, continuous speech, telephone bandwidth speech database," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1990)*, vol. 1, pp. 109–112, April 1990.

[62] A. Higgins, L. Bahler, G. Vensko, J. Porter, and D. Vermilyea, "YOHO speaker authentication final report," *ITT Defense Communications Division*, 1989.

[63] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth, "Acoustic emotion recognition: A benchmark comparison of performances," *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2009)*, pp. 552–557, December 2009.

[64] J. Campbell, W. Shen, W. Campbell, R. Schwartz, J. Bonastre, and D. Matrouf, "Forensic speaker recognition," *IEEE Signal Processing Magazine*, vol. 26, pp. 95–103, March 2009.

[65] D. A. Reynolds, "An overview of automatic speaker recognition technology," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2002)*, pp. IV: 4072–4075, May 2002.

[66] J. Webster, "Ambient noise statistics," *IEEE Transactions on Signal Processing*, vol. 41, pp. 2249–2253, June 1993.