Combinação das respostas de classificadores em sub-bandas para reconhecimento de locutor usando o espaço nulo e treinamento com múltiplas condições

Eduardo Esteves Vale e Abraham Alcaim

***Resumo*—Este artigo apresenta a aplicação do treinamento com múltiplas condições no reconhecimento de locutor empregando múltiplos classificadores em sub-bandas e o espaço nulo na combinação das respostas. Os resultados mostraram que, principalmente para os testes em ruído branco, a proposta apresenta melhores resultados que os de outras técnicas da literatura.**

***Palavras-Chave*—Treinamento com múltiplas condições, múltiplos classificadores em sub-bandas, reconhecimento de locutor.**

***Abstract*—This paper presents the application of multicondition training in the speaker recognition scheme which employs multiple subband-classifiers and the null space to combine the responses. The simulation results show that the proposed strategy outperforms other techniques reported in the literature, mainly in case of tests in white noise.**

***Index Terms*—Multicondition training, multiple subband-classifiers, speaker recognition.**

I. INTRODUÇÃO

Os sistemas de reconhecimento de locutor são utilizados em diversas situações de grande interesse, como por exemplo, em investigações policiais, transações comerciais e bancárias usando comandos de voz, e outras aplicações. O desempenho desses sistemas é severamente afetado por diferentes tipos de ruído ambiente, que frequentemente podem estar presentes na voz. Com o objetivo de superar este problema, foi mostrado em vários trabalhos [1]-[4] que o emprego das técnicas de múltiplos classificadores em sub-bandas nos sistemas de reconhecimento, permite que um melhor desempenho seja alcançado em comparação com os sistemas que utilizam um classificador sobre todo o espectro do sinal de voz. A razão disso é que, a partir da decomposição do sinal em sub-bandas de freqüências, as informações relacionadas à identidade do locutor e às contribuições indesejadas, como os ruídos, têm diferentes distribuições em freqüência. Ou seja, algumas regiões do espectro são mais importantes que outras para o reconhecimento do locutor. Portanto, estas técnicas baseadas em múltiplos classificadores em sub-bandas tendem a tirar mais proveito dessas regiões mais importantes do espectro.

Eduardo Esteves Vale (Bolsista do CNPq-Brasil) e Abraham Alcaim, CETUC, Pontifícia Universidade Católica, Rio de Janeiro, Brasil, E-mails: eevale@globo.com, alcaim@cetuc.puc-rio.br.

Esses esquemas de múltiplos classificadores em sub-bandas decompõem o sinal de voz em *n* sub-bandas de freqüências e extraem atributos de cada sinal passa-banda. Os atributos de uma sub-banda são usados como entrada para um classificador aplicado naquela banda. Na fase de teste, as saídas dos *n* classificadores são combinadas com a finalidade de produzir uma resposta conjunta. A Figura 1 mostra este esquema. Todo locutor é modelado por um esquema como esse. Em [1]-[4] foram usados filtros mel-espaçados operando diretamente no domínio do tempo. É importante ressaltar que a escala mel é baseada no funcionamento do sistema auditivo humano [5]. Nessa escala, a região de baixas freqüências do sinal é representada com maior resolução que a de altas freqüências. Portanto, é possível capturar formantes presentes na região de baixas freqüências e que caracterizam as ressonâncias do aparelho vocal. Assim sendo, a escala mel é mais relacionada com as informações de identidade do locutor [2], [6], que a escala linear.

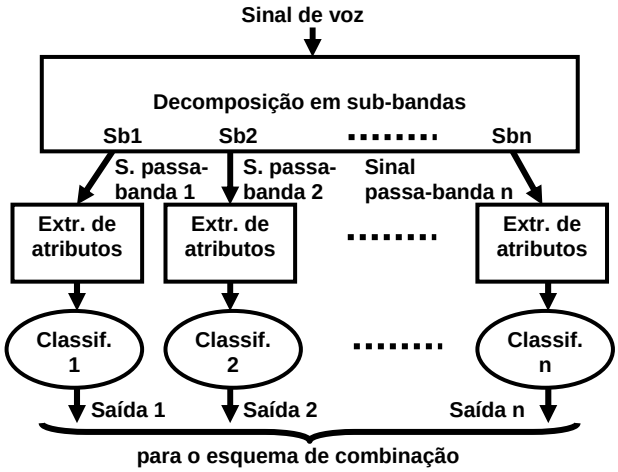

Fig. 1. Sistema de reconhecimento de locutor usando múltiplos classificadores em sub-bandas.

A técnica proposta em [1] soma as *n* saídas dos classificadores em sub-bandas para obter uma resposta conjunta. O locutor é identificado se o seu esquema produz a

maior verossimilhança conjunta quando comparada com as dos outros locutores modelados. O esquema mostrado em [2] realiza uma operação semelhante. A diferença é que a energia de cada sinal passa-banda usado no treinamento é utilizada para escolher as bandas a serem empregadas no reconhecimento. Em nossas simulações, como critério de combinação, foram escolhidas as quatro sub-bandas de maior energia (durante o treinamento) e somadas as respectivas saídas dos classificadores associados a essas bandas (no teste). Em nosso esquema proposto em [3] foi empregada uma combinação não uniforme das *n* saídas favorecendo àquelas relacionadas com as baixas freqüências. Os pesos usados para esse propósito são obtidos computando-se, no domínio do tempo, a energia total de cada sinal passa-banda usado no treinamento. A resposta conjunta é a soma ponderada das saídas. Foi mostrado que na maioria dos casos esse esquema apresenta melhores resultados que os demais.

Mais recentemente, nós mostramos em [4] que uma estratégia de combinação das respostas dos classificadores em sub-bandas, empregando o espaço nulo, é capaz de melhorar bastante o desempenho do sistema de reconhecimento em comparação com o das outras técnicas, quando a voz está contaminada por diferentes tipos de ruído. Nesse artigo nós consideraremos essa abordagem combinada com o treinamento usando múltiplas condições.

II. Reconhecimento Usando o Espaço Nulo

A álgebra linear afirma [7] que o espaço nulo de uma matriz A *m* por *n* é um espaço vetorial formado por todas as soluções do sistema homogêneo, Ax=0, onde x representa o vetor solução de dimensão *n*. O cálculo de uma base para o espaço nulo é uma operação que fornece uma coleção de soluções linearmente independentes que é fechada com relação à soma e à multiplicação por escalar. Isto significa que a soma desses vetores ou qualquer múltiplo deles é também uma solução de Ax=0.

A técnica mostrada em [4] usa a energia total de cada sinal passa-banda de treino como os elementos das colunas de A. Portanto, A é definida a partir de importantes informações de identidade do locutor em diferentes sub-bandas. Note-se que, nesse caso, A é uma matriz 1 por *n*, onde *n* é o número de sub-bandas. Portanto, cada coluna de A representa um valor de energia de uma sub-banda do sinal de voz. O cálculo de uma base para o espaço nulo de A fornece um conjunto de *n*-1 vetores solução $\{x_1, ..., x_{n-1}\}$ que retém as informações dependentes do locutor. Esses vetores fornecem *n*-1 modos de representar essas informações sem redundâncias, já que essas soluções são linearmente independentes. Os elementos desses vetores são usados como pesos não-uniformes para serem aplicados nas saídas dos classificadores em sub-bandas, e refletem a dependência do locutor. Durante o teste de um locutor modelado, todos os vetores da base do espaço nulo relacionados àquele locutor são testados com o propósito de saber qual deles melhor contribui para o reconhecimento. É escolhido aquele que forneça a maior verossimilhança. Note-se que as técnicas presentes em [1]-[3] não fornecem esta habilidade, já que possuem apenas uma alternativa de pesos para cada situação de teste. Por outro lado, a estratégia mostrada em [4] fornece um número de alternativas (graus de liberdade) igual ao número de vetores contidos na base do espaço nulo associado ao locutor modelado.

III. Treinamento Com Múltiplas Condições

O treinamento baseado em múltiplas condições consiste em treinar o classificador a partir de voz contaminada por ruído. Essa estratégia permite realizar uma compensação no modelo para o efeito do ruído [8]. Nesse esquema, o sistema de reconhecimento tira proveito das componentes espectrais mais casadas [9] entre o treino e o teste.

Uma técnica recente, apresentada em [8]-[10] emprega o treinamento em múltiplas condições e a exclusão de atributo. Nessa abordagem, um classificador é treinado com voz sem ruído e com voz contaminada por ruído branco em vários valores de Relação Sinal Ruído (RSR: 10, 12, 14, 16, 18, 20dB). Um banco de filtros mel é utilizado na extração dos atributos. As saídas dos filtros mel são descorrelatadas por um filtro passa-altas ($H(z)=1-z^{-1}$), cujas saídas correspondem às contribuições das sub-bandas empregadas nessa abordagem. No teste de cada janela de voz, são feitas todas as combinações (soma duas a duas, três a três, quatro a quatro, etc.) de verossimilhanças associadas a cada uma das sub-bandas e produzidas por um classificador GMM (Gaussian Mixture Model). A resposta do classificador para a janela testada é a verossimilhança resultante da combinação que fornecer o maior resultado. Note-se que as sub-bandas que contribuem menos para o reconhecimento são simplesmente excluídas no teste. Esse método é, entretanto, de alta complexidade computacional.

A nossa proposta consiste em empregar o treinamento em múltiplas condições para treinar os classificadores em sub-bandas que utilizam o espaço nulo na combinação das respostas. Nesse esquema, o sinal de treino em cada valor de RSR é usado para treinar um sistema, como o descrito em [4], de classificadores em sub-bandas que empregam o espaço nulo. Além disso, é treinado mais um desses sistemas usando voz sem ruído. Então, considerando os valores de RSR para o treinamento como sendo 10, 12, 14, 16, 18, 20dB, cada locutor será modelado por sete sistemas de classificadores em sub-bandas empregando o espaço nulo. Esse espaço nulo é obtido do sinal de treino sem ruído. A resposta conjunta final é a do sistema que produzir o maior valor de verossimilhança de saída.

Ressalta-se que nessa proposta nenhuma sub-banda é desprezada, e não são feitas combinações duas a duas, três a três, etc. de verossimilhanças, reduzindo bastante a complexidade em relação ao esquema apresentado em [8]-[10].

IV. Resultados de Simulação

Serão apresentados resultados de simulação para a identificação de locutor independente do texto com o objetivo de mostrar o desempenho da nova proposta quando comparada com outras técnicas. Em nossos experimentos, os sinais de voz utilizados no teste foram contaminados por

ruídos coloridos da base NOISEX-92 [11] e por ruído gaussiano branco. Foram usados 49 locutores (masculinos) e seus sinais de voz sem ruído (amostrados em 8 kHz) obtidos das sessões de 1 a 5 da base KING [12]. O classificador em sub-bandas usado foi o GMM, já que é uma poderosa ferramenta estatística extensivamente usada nas aplicações de reconhecimento de locutor [13]. Os sinais sem ruído das sessões 1, 2 e 3, sem silêncio, foram usados para treinar os classificadores (na condição de treino sem ruído) com 90 segundos de voz. Já na condição de treino (em 90s) com ruído, os sinais dessas sessões foram contaminados por ruído gaussiano branco nos valores de RSR: 10, 12, 14, 16, 18 e 20dB. Os sinais das duas sessões restantes, compostas por quatro segmentos de 15 segundos (para cada locutor e sem silêncio), foram corrompidos por ruído em 10dB de RSR e utilizados para o teste. Foram usados 20 atributos MFCC (Mel Frequency Cepstral Coefficients) [6] extraídos em janelas de voz (janela de Hamming e superposição de 50%) de 20ms. As técnicas apresentadas em [1]-[4] e a nova proposta usaram sub-bandas produzidas por filtros de Butterworth de 6ª ordem mel espaçados. Todos os GMM usaram 32 gaussianas. O desempenho da identificação usando apenas um GMM (sem decomposição em sub-bandas) foi de 96,43% para o teste com voz sem ruído. Esse desempenho cai severamente (10,2%) em ambientes ruidosos, como veremos a seguir.

Conforme mostrado na Tabela 1, quando a voz de teste está contaminada por ruído gaussiano branco, a técnica que apresenta os melhores resultados (72,45% usando 6 sub-bandas, 73,47% usando 4 sub-bandas) é a que emprega a combinação do espaço nulo com o treinamento em múltiplas condições. Esses resultados são bem superiores ao obtido com espaço nulo [4] sem múltiplas condições (25,51%). A nova proposta também apresenta os melhores resultados (58,67% usando 6 sub-bandas, 60,71% usando 4 sub-bandas) quando o sinal de teste está contaminado com ruído de carro.

TABELA I
TESTE EM 15S E COM RUÍDOS EM 10DB DE RSR.

Em % acerto	Fábri-ca	Fala-tório	Carro	Branco
Soma(4Sbs)[1]	34,69	69,39	43,88	17,35
Ncomb (4Sbs)[3]	34,69	61,22	54,59	13,27
Espaço nulo (4Sbs) [4]	48,98	80,10	58,16	25,51
CRSR Esp. Nulo (6Sbs)	42,86	67,86	58,67	72,45
CRSR Esp. Nulo (4Sbs)	46,43	63,78	60,71	73,47
1GMM	34,18	63,27	33,67	10,20

Os resultados obtidos indicam que acrescentar no esquema de combinação das respostas dos classificadores e na geração de modelos, as informações relacionadas ao comportamento espectral do sinal ruidoso, ou seja, a distribuição do ruído (treinamento com múltiplas condições) e das contribuições do sinal de voz na freqüência (espaço nulo), pode melhorar muito o desempenho do reconhecimento para ruídos de teste semelhantes aos usados no treino. Na abordagem desse artigo, isso foi feito através da utilização de uma base para o espaço nulo obtido das energias das sub-bandas dos sinais sem ruído, e pelo treinamento em múltiplas condições usando ruído branco. No caso do ruído colorido é necessário que seja feita uma melhor modelagem de sua distribuição em freqüência para que o treinamento em múltiplas condições ajude a melhorar significativamente o desempenho. Ou seja, o ruído a ser usado no treino deve ter comportamento espectral o mais semelhante possível do ruído presente no sinal que testará o sistema, tal que melhore o casamento entre o modelo treinado e o sinal de teste. O ruído branco utilizado no treino não demonstrou ter um comportamento espectral semelhante aos ruídos coloridos presentes nos sinais usados para testar o sistema, exceto para o caso do ruído de carro. Apesar do ruído branco não ser semelhante ao de carro, ele contribui favoravelmente dentro das sub-bandas para melhorar o reconhecimento.

V. CONCLUSÕES

Foi apresentada neste artigo uma abordagem para reconhecimento de locutor independente do texto que emprega o espaço nulo na combinação das respostas dos classificadores em sub-bandas e o treinamento em múltiplas condições. Mostrou-se que a técnica baseada no espaço nulo pode tirar proveito da compensação realizada pelo treinamento em múltiplas condições, no sentido de melhorar o desempenho do reconhecimento quando os sinais de treino estão contaminados com ruído branco e os de teste com ruído branco ou de carro. Os resultados reforçam a idéia de que considerar no esquema de combinação das respostas dos classificadores e na geração de modelos, as informações relacionadas ao comportamento espectral do sinal, pode contribuir significativamente para melhorar o desempenho do sistema de reconhecimento. Para os testes com os demais tipos de ruído colorido, necessita-se que seja feita uma melhor escolha do tipo (ou tipos) de ruído usado para treinar o sistema.

REFERÊNCIAS

[1] J. E. Higgins, R. I. Damper e T. J. Dodd, “Information fusion for subband-HMM speaker recognition,” in *Proc. IEEE Int. Joint Conf. on Neural Networks*, v. 2, pp. 1504-1509, Washington DC, Julho 2001.

[2] Z. Sakka, A. Kachouri, A. Mezghani e M. Samet, “A new method for speech denoising and speaker verification using subband architecture,” *First Int. Symp. Control, Communications and Signal Processing*, pp. 37-40, Hammamet, Março 2004.

[3] E. E. Vale, A. A. Cunha e A. Alcaim, “Robust text-independent speaker identification using multiple subband-classifiers in colored noise environment,” *Int. Conf. on Systems, Signals and Image Processing*, pp. 275-277, Bratislava, Junho 2008.

[4] E. E. Vale e A. Alcaim, “Adaptive weighting of subband-classifier responses for robust text-independent speaker recognition,” *Electronics Letters*, v. 44, issue 21, pp. 1280-1282, Outubro 2008.

[5] A. M. L. Araújo and F. Violaro, “Formant frequency estimation using a mel scale LPC algorithm,” *Proc. SBT/IEEE Int. Telecommunication Symposium*, v. 1, pp. 207-212, São Paulo, Agosto 1998.

[6] S. Chakroborty, A. Roy, S. Majumdar and G. Saha, "Capturing complementary information via reversed filter bank and parallel implementation with MFCC for improved text-independent speaker identification," *Proc Int. Conf Computing: Theory and Applications*, pp. 463-467, Kolkata, Março 2007.

[7] G. Strang, *Linear Algebra and Its Applications*,. Thomson Learning, 1988.

[8] J. Ming, T. J. Hazen, J. R. Glass and D. A. Reynolds, "Robust speaker recognition in noisy conditions," *IEEE Trans. On Audio, Speech and Language Processing*, v. 15, no. 5, pp. 1711-1723, Julho 2007.

[9] J. Ming, D. Stewart and S. Vaseghi, "Speaker identification in unknown noisy conditions-a universal compensation approach," *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, v.1, pp. 617-620, Philadelphia-US, Março 2005.

[10] J. Ming, J. Lin and F. J. Smith, "A posterior union model with applications to robust speech and speaker recognition," *EURASIP Journal on Applied Signal Processing*, v. 2006, Article ID 75390, pp. 1-12.

[11] A. Varga, H. J. M. Steeneken, M. Tomlinson and D. Jones, "The NOISEX-92 study on the effect of additive noise on automatic speech recognition," *Technical Report*, DRA Speech Research Unit, 1992.

[12] J. Godfrey, D. Graff and A. Martin, "Public databases for speaker recognition and verification," *Proc. ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, pp. 39-42, Switzerland, Abril 1994.

[13] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. on Speech and Audio Processing*, v. 3, no. 2, pp. 72-83, Janeiro 1995.