3
Identificação de Locutor Usando Técnicas de Múltiplos Classificadores em Sub-bandas Com Pesos Não-Uniformes

Neste capítulo, é apresentada uma nova proposta de combinação de múltiplos classificadores em sub-bandas visando melhorar o desempenho da identificação de locutor independente do texto quando a voz está contaminada por ruído branco ou não-branco.

Um fato importante a ser considerado em relação aos dados usados pelo sistema de reconhecimento de locutor é a distribuição não-uniforme das componentes do sinal de voz na frequência. A informação da identidade do locutor e as contribuições indesejadas, como os ruídos, têm diferentes distribuições em frequência. Como mencionado anteriormente, uma boa estratégia decorrente da observação desse fato é dividir o sinal em sub-bandas, ao invés de usar diretamente toda a faixa de frequências. Por exemplo, as técnicas apresentadas em [38] e [39] dividem o sinal de voz em *n* sub-bandas de frequências e extraem atributos dos sinais passa-banda. Os atributos de uma sub-banda são usados como entradas para um classificador aplicado naquela sub-banda. As técnicas que não se baseiam nesta estratégia de divisão do sinal em bandas aplicam o esquema de reconhecimento nos atributos extraídos considerando globalmente toda a faixa de frequências do sinal.

A técnica de combinação descrita em [38] aplica pesos lineares e iguais nas respostas dos classificadores. O método apresentado em [39] realiza uma operação semelhante. A diferença é que a energia de cada sinal passa-banda de treinamento é utilizada para escolher as sub-bandas a serem usadas no reconhecimento.

Neste capítulo, é proposta uma nova estratégia que utiliza pesos não-uniformes para a combinação das respostas dos classificadores nas sub-bandas. O método consiste em dar mais peso àquelas respostas de classificadores aplicados em sub-bandas mais relacionadas à informação da identidade do locutor, e menos peso às demais, tirando proveito da distribuição não uniforme das componentes do sinal na frequência.

Este capítulo está organizado da seguinte forma: na Seção 3.1 são resumidamente descritos os fundamentos das técnicas de reconhecimento baseadas em classificadores aplicados em sub-bandas, na Seção 3.2 é apresentada a técnica proposta, a Seção 3.3 contém os resultados experimentais, e a Seção 3.4, apresenta as principais conclusões deste trabalho de pesquisa.

3.1
Estratégias Empregadas em Identificação de Locutor Usando Múltiplos Classificadores em Sub-bandas

As técnicas de reconhecimento de locutor baseadas em sub-bandas e apresentadas em [38] e [39] utilizam um banco de filtros para dividir o sinal de voz em bandas. Para tal finalidade foram usados filtros Butterworth espaçados na escala Mel. É importante ressaltar que a escala Mel provê maior resolução às frequências mais baixas, as quais estão mais relacionadas à informação da identidade do locutor. Nessas estratégias, os coeficientes MFCC são extraídos do sinal passa-banda proveniente de cada sub-banda. Esse processo é realizado tanto na fase de projeto quanto na de teste do sistema de reconhecimento. Em [38], o sinal de voz é dividido em *n* sub-bandas e *n* classificadores são aplicados a essas bandas. Isto significa que cada banda tem seu próprio classificador. No teste, as *n* saídas correspondentes aos *n* classificadores são somadas gerando uma medida de verossimilhança conjunta (ou resposta conjunta). Todo o locutor modelado possui um esquema como este. O locutor é identificado se o seu esquema produzir a maior verossimilhança conjunta comparada com a dos outros locutores modelados pelos classificadores. A técnica apresentada em [39] utiliza a energia de cada sinal passa-banda usado no projeto, que é computada no domínio do tempo, para escolher as bandas a serem usadas no reconhecimento. Essa energia é calculada da seguinte forma

$$E_{T_{Sb}} = \sum_{k=1}^{K} s_{Sb}{}^{2}(k) \tag{35}$$

onde $E_{T_{Sb}}$ é a energia total obtida das K amostras do sinal passa-banda utilizado para o treino. Como critério de combinação nos experimentos realizados nesta tese, foram escolhidas as quatro sub-bandas de maior energia e somadas às saídas dos seus respectivos classificadores.

3.2
O Método Proposto que Utiliza uma Combinação de Pesos Não-Uniformes

Uma das novas regras de combinação propostas nesta tese consiste em aplicar pesos não-uniformes nas saídas dos classificadores em sub-bandas. A medida de verossimilhança conjunta é a soma ponderada dessas saídas. Os pesos são obtidos calculando-se, no domínio do tempo, a energia total de cada sinal passa-banda usado na fase de projeto do reconhecimento. A Figura 12 mostra um diagrama de blocos desse esquema de combinação. Nessa estratégia, a sub-banda mais energética de um locutor modelado tem a maior contribuição para a verossimilhança conjunta. As contribuições nas regiões do espectro de baixas frequências tendem a ser maiores do que nas outras, as quais são mais relacionadas, por exemplo, aos ruídos. Portanto, esta proposta aplica mais ênfase nas saídas dos classificadores utilizados nas bandas de frequências mais baixas, as quais estão mais relacionadas à informação da identidade do locutor. Isso tem por vantagem aproveitar melhor as contribuições mais importantes para a identificação do locutor, e menos, àquelas que tendem a ser mais afetadas por ruídos, sem descartá-las totalmente. Essa abordagem não é observada em [38], já que essa estratégia considera igualmente todas as contribuições. No método proposto, o locutor é identificado se o seu sistema de reconhecimento produzir a maior verossimilhança conjunta comparada com a dos outros locutores. Nessa proposta, nada se assume em relação à natureza da degradação, como por exemplo, aditividade, descorrelação, entre outras. Na próxima seção, são apresentados resultados experimentais para diferentes tipos de ruído e RSR (Razão Sinal Ruído), para a aplicação de identificação de locutor independente do texto.

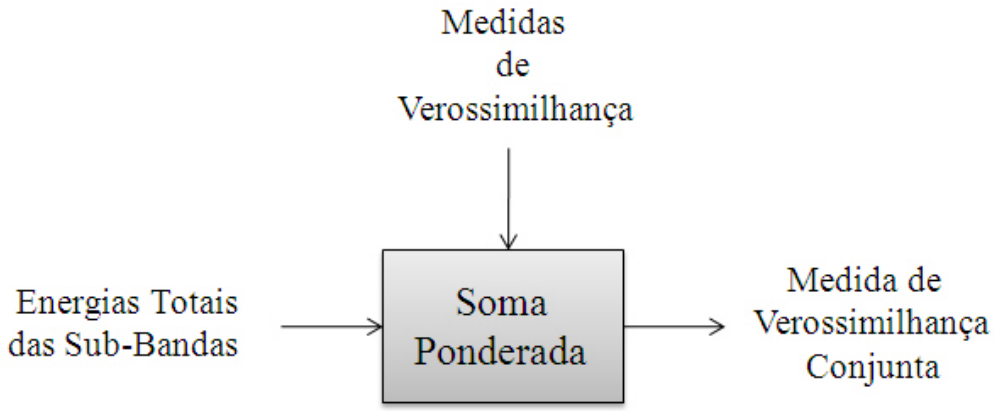

Figura 12 – Esquema de combinação de pesos não-uniformes

3.3
Resultados de Simulação

Os resultados experimentais, em termos da taxa de identificação (número de identificações corretas / número de testes) × 100%, são apresentados visando mostrar o desempenho dessa nova proposta quando comparada com outras técnicas. A base de voz KING [92] foi utilizada nas simulações. Os sinais dessa base foram coletados em 1987 inicialmente para serem usados em um contrato de pesquisa do governo dos Estados Unidos. O conteúdo oficialmente tornado público é fornecido pela LDC (*Linguistic Data Consortium*). Essa base possui vozes gravadas de 51 locutores do sexo masculino em duas versões, que diferem quanto ao processo de aquisição. Uma versão contém as vozes provenientes da telefonia e a outra versão contém as vozes diretamente coletadas por um microfone de alta qualidade (versão sem ruído). Os locutores são subdivididos em dois grupos, um de 25 e o outro de 26 indivíduos, os quais foram gravados em localidades distintas. Para cada locutor e versão existem cerca de 10 gravações, correspondendo cada uma delas a uma sessão de aproximadamente 30 a 60 segundos de duração. O intervalo entre a aquisição de cada sessão varia de aproximadamente uma semana a um mês. As cinco primeiras sessões de cada locutor foram gravadas com uma semana de intervalo entre elas, e as demais sessões foram gravadas com um mês de intervalo. Um total de 51 locutores

produziram cinco gravações da versão sem ruído, enquanto que 49 desses 51 produziram as outras cinco gravações. Na versão de telefonia, cada um dos 51 locutores produziram 10 gravações. O conteúdo das sessões correspondem às falas no idioma inglês resultantes da interpretação de desenhos, resolução de problemas, descrição de uma figura, e outros textos. Os sinais dessa base são amostrados em 8 kHz.

Nos experimentos desta tese, são utilizados sinais de voz da base KING corrompidos pelos ruídos coloridos da base de ruídos NOISEX-92 e por ruído gaussiano branco gerado em Matlab. O tipo de classificador utilizado é o GMM, já que essa é uma eficiente ferramenta estatística extensivamente usada e testada em aplicações de reconhecimento de locutor. Nas simulações são usados 49 locutores e seus correspondentes sinais de voz das sessões 1 a 5 como em [6]. Os sinais sem ruído das sessões 1, 2 e 3, e sem silêncio são utilizados para treinar os classificadores GMM com 90 segundos de voz. Os sinais das duas sessões restantes, compostos por quatro segmentos de 15 segundos (para cada locutor e sem silêncio), são corrompidos pelos ruídos e usados para o teste. São utilizados 20 atributos MFCC extraídos em janelas de 20 ms de voz (usando janela de Hamming com superposição de 50%), no caso do reconhecimento banda larga e com um classificador GMM. As técnicas apresentadas em [38], [39] e a proposta (aqui denominada Prop1) usam sub-bandas produzidas por filtros Butterworth Mel-espaçados e de 6^{a} ordem, aplicados em cada janela de voz e processados diretamente no domínio do tempo. Vetores de 20 atributos MFCC são extraídos em cada sub-banda (com janela de Hamming de 20 ms e superposição de 50%). Todos os classificadores GMM usam 32 gaussianas. O desempenho de identificação, em termos da taxa de acerto, obtido pelo uso de um GMM banda larga é de 96,43%, no teste com voz sem ruído. As tabelas a seguir mostram os resultados obtidos nos experimentos.

NO. DE SB	2	4	6	8	10	12	14	16	18	20
Soma[38]	37,24	34,69	35,71	35,20	28,06	16,84	27,04	27,04	33,16	28,57
Energia[39]	-	-	29,59	26,02	4,08	1,02	2,04	2,04	0,00	0,00
Prop1	37,24	34,69	37,24	38,78	37,24	34,69	39,29	34,69	35,20	31,12

Tabela 3.1 - Desempenho de identificação (%) em 15s de teste para ruído de Fábrica em RSR=10dB; com um GMM: 34,18%

NO. DE SB	2	4	6	8	10	12	14	16	18	20
Soma[38]	6,12	4,59	3,57	3,06	4,59	2,55	3,57	3,06	4,08	4,08
Energia[39]	-	-	5,61	5,10	1,53	1,53	2,55	2,04	0,00	0,00
Prop1	6,12	4,08	3,57	3,06	5,10	3,57	4,59	4,59	4,08	2,55

Tabela 3.2 - Desempenho de identificação (%) em 15s de teste para ruído de Fábrica em RSR=0dB; com um GMM: 3,57%

NO. DE SB	2	4	6	8	10	12	14	16	18	20
Soma[38]	70,41	69,39	64,29	60,20	43,88	30,10	43,88	43,37	37,24	35,20
Energia[39]	-	-	64,29	49,49	3,57	1,53	2,55	2,55	0,00	0,00
Prop1	65,82	61,22	62,24	58,67	52,04	50,51	52,55	52,55	42,35	39,80

Tabela 3.3 - Desempenho de identificação (%) em 15s de teste para ruído de Falatório em RSR=10dB; com um GMM: 63,27%

NO. DE SB	2	4	6	8	10	12	14	16	18	20
Soma[38]	10,71	11,22	13,78	11,73	5,61	14,80	18,37	8,67	6,63	16,84
Energia[39]	-	-	7,14	5,61	3,57	2,55	2,04	2,04	0,00	0,00
Prop1	10,71	12,24	13,78	14,29	9,18	23,47	20,92	11,22	6,63	21,94

Tabela 3.4 - Desempenho de identificação (%) em 15s de teste para ruído de Falatório em RSR=0dB; com um GMM: 10,71%

NO. DE SB	2	4	6	8	10	12	14	16	18	20
Soma[38]	43,37	43,88	40,31	37,76	33,16	19,39	29,59	29,08	33,16	31,63
Energia[39]	-	-	28,06	22,96	7,14	1,53	2,04	2,55	0,00	0,00
Prop1	51,53	54,59	57,65	53,57	43,88	42,86	42,35	39,80	42,35	41,33

Tabela 3.5 - Desempenho de identificação (%) em 15s de teste para ruído de Carro em RSR=10dB; com um GMM: 33,67%

NO. DE SB	2	4	6	8	10	12	14	16	18	20
Soma[38]	14,80	13,78	13,27	11,22	9,69	5,10	7,14	9,69	9,18	9,18
Energia[39]	-	-	7,65	9,69	2,04	1,02	2,04	2,55	0,00	0,00
Prop1	23,47	21,94	24,49	27,04	20,92	22,96	16,84	14,29	11,73	14,80

Tabela 3.6 - Desempenho de identificação (%) em 15s de teste para ruído de Carro em RSR=0dB; com um GMM: 13,78%

NO. DE SB	2	4	6	8	10	12	14	16	18	20
Soma[38]	18,37	17,35	18,37	17,86	18,88	13,27	13,27	17,86	18,88	16,33
Energia[39]	-	-	22,96	18,88	1,53	2,04	1,53	2,04	0,00	0,00
Prop1	14,29	13,27	13,27	14,29	13,27	15,31	15,31	15,82	15,82	13,78

Tabela 3.7 - Desempenho de identificação (%) em 15s de teste para ruído Branco em RSR=10dB; com um GMM: 10,20%

NO. DE SB	2	4	6	8	10	12	14	16	18	20
Soma[38]	6,12	5,10	6,12	9,69	8,16	6,12	7,14	3,57	4,59	5,10
Energia[39]	-	-	4,59	5,61	2,04	2,04	1,53	2,55	0,00	0,00
Prop1	4,59	4,08	3,57	7,14	4,08	8,67	6,63	3,57	7,65	4,08

Tabela 3.8 - Desempenho de identificação (%) em 15s de teste para ruído Branco em RSR=0dB; com um GMM: 2,55%

A Tab. 3.1 mostra os resultados de simulação da técnica apresentada em [38] designada por “Soma”, da apresentada em [39] designada por “Energia” e da proposta “Prop1”, usando 2, 4, 6, 8, 10, 12, 14, 16, 18 ou 20 sub-bandas e voz

corrompida por ruído de Fábrica (*factory1* da base NOISEX) em 10dB de RSR. A Tab. 3.2 mostra os resultados de identificação das técnicas para a voz corrompida por ruído de Fábrica em 0 dB. As Tabelas 3.3 e 3.4 apresentam os resultados dos métodos para a voz corrompida por ruído Falatório (*babble* da base NOISEX) em 10dB e 0dB respectivamente. As Tabelas 3.5 e 3.6 mostram os resultados para a voz corrompida por ruído de Carro (*Volvo* da NOISEX) em 10dB e 0dB. Finalmente, as Tabelas 3.7 e 3.8 apresentam os resultados para a voz corrompida por ruído gaussiano branco em 10dB e 0dB, respectivamente.

De acordo com a Tab. 3.1 (RSR=10dB, ruído de Fábrica) o melhor desempenho de identificação (39,29%) é obtido pela técnica proposta usando 14 sub-bandas. Na maioria dos casos, o esquema proposto teve um desempenho melhor do que os demais para o mesmo número de sub-bandas usadas na decomposição do sinal de voz. Quando a RSR é reduzida a 0dB, como pode ser visto na Tab. 3.2, o maior desempenho (6,12%) é obtido usando 2 sub-bandas para os esquemas Soma e proposto, enquanto o esquema com um GMM alcançou 3,57%. Quando o sinal de voz está corrompido com ruído Falatório, o esquema com um GMM gerou um desempenho de 63,27% em 10dB, e todas as técnicas multibandas alcançaram, para um determinado número de sub-bandas, um resultado melhor como podemos verificar da Tab. 3.3. Adicionalmente, quando a RSR é igual a 0dB, o melhor desempenho é obtida pelo método proposto (23,47%). Para esse valor de RSR, a técnica proposta é geralmente melhor do que as demais para o mesmo número de sub-bandas como visto na Tab. 3.4. No caso do ruído de Carro (10dB), mostrado na Tab. 3.5, mais uma vez o melhor resultado (57,65%) é obtido pelo esquema proposto. Adicionalmente, essa técnica supera em desempenho as demais para o mesmo número de sub-bandas na decomposição. O esquema com um GMM alcançou 33,67%. O mesmo comportamento em desempenho pode ser visto na Tab. 3.6 para 0dB, onde a técnica proposta forneceu o melhor resultado (27,04%).

Quando é considerado ruído Branco (10dB), mostrado na Tab. 3.7, o melhor desempenho é obtido pelo Energia (22,96%). Em 0dB, mostrado na Tab. 3.8, o melhor resultado é obtido pelo Soma (9,69%). Portanto, esse novo esquema proposto não é a melhor estratégia para ruído Branco. Note-se que a densidade espectral de potência plana desse tipo de ruído afeta igualmente todas as sub-bandas. Consequentemente, quando diferentes valores de pesos são aplicados a

todas as sub-bandas, em algumas situações, isso pode dar ênfase aos efeitos de ruídos nas sub-bandas menos significativas para a identificação do locutor. Isso não acontece nos outros dois métodos. Por outro lado, para o ruído colorido, em que a densidade espectral de potência não é constante, é mais adequado o uso de pesos diferentes para sub-bandas distintas. Portanto, essa nova técnica proposta fornece melhores resultados do que as demais em face de ruído colorido. Na presença de ruído Branco ela somente melhora mais a taxa de acerto que o esquema com um classificador GMM.

3.4
Conclusões

Neste capítulo, foi apresentada uma nova proposta que usa pesos não-uniformes para combinar as respostas dos classificadores em sub-bandas visando aumentar o desempenho dos sistemas de reconhecimento de locutor em condições de sinais ruidosos. Vários experimentos de identificação de locutor independente do texto usando um amplo material de voz e ruídos ambientes foram desenvolvidos com o propósito de deixar evidente o comportamento dessa nova proposta quando comparada com as demais técnicas. Os resultados mostraram que, para ruído colorido, na maioria dos casos o melhor desempenho foi alcançado pelo método proposto.