

UM ESTUDO COMPARATIVO DE CARACTERÍSTICAS DA FALA BASEADAS EM WAVELETS PARA RECONHECIMENTO DE EMOÇÃO

RENAN C. NASCIMENTO*, DANIEL C. CAVALIERI†, LUIZ A. PINTO†, ALBERTO F. DE SOUZA‡, KARIN S. KOMATI*

*Coordenação de Informática - Instituto Federal do Espírito Santo (IFES) campus Serra

†Coordenação de Automação - Instituto Federal do Espírito Santo (IFES) campus Serra
Rodovia ES-010, Km 6,5, Mangunhos, Serra, ES - Brasil. 29173-087

‡Departamento de Informática, Universidade Federal do Espírito Santo (UFES) campus Goiabeiras
Av. Fernando Ferrari, 514, Goiabeiras, Vitória, ES - Brasil. 29075-910

Emails: renan@costanascimento.com.br, daniel.cavalieri@ifes.edu.br,
luiz.pt@ifes.edu.br, alberto@lcad.inf.ufes.br, kkomati@ifes.edu.br

Abstract— Emotional features extracted from a voice signal can improve the performance of speech recognition systems. In this work, a comparative study of speech features based on wavelet families is proposed in order to determine which family has the highest accuracy and sensitivity in recognizing speech emotions. To do so, several wavelets families are used to extract features from a database of emotional speech, EmoDB (*Berlin of Emotional Speech*). The feature extraction experiments cover seven wavelets families in seven different values of decomposition levels, using the Discret Wavelet Transform. Finally, confusion matrices were generated, as well as the measures Precision, Recall and F-measure were calculated in order to analyze the responses of emotion recognition of different families and levels. At the end, the wavelet/level combination that achieved the best performance was Coif1 with 20 levels of decomposition. Other interesting result is that the best performance of the automatic recognition is achieved for the emotion of sadness, regardless of the wavelet family, being much better than the result of human perception. In contrast, for the anger emotion, human perception is much better than the automatic system.

Keywords— Speech Emotion Recognition, Discrete Wavelet Transform, Berlin Emotional Database.

Resumo— Características emocionais extraídas de um sinal de voz podem melhorar o desempenho de sistemas de reconhecimento de fala. Neste trabalho, propõe-se um estudo comparativo de características da fala, extraídas através de famílias de *wavelets*, a fim de determinar qual das famílias possui a maior precisão e sensibilidade no reconhecimento das emoções e, além disso, comparar os resultados do sistema automático com os resultados da percepção humana. Para isso, o trabalho utiliza o banco de dados de vozes, EmoDB (*Berlin of Emotional Speech*), classificadas quanto às seguintes emoções humanas: raiva, desgosto, medo, alegria, tristeza, tédio e neutro. A extração de características dos experimentos compreenderam sete famílias de wavelets em sete diferentes valores de níveis de decomposição, usando a Transformada Wavelet Discreta. Coletou-se os seguintes resultados quantitativos: as matrizes de confusão, as medidas *Precision* (precisão), *Recall* (sensibilidade) e *F-measure* para analisar comparativamente as respostas do reconhecimento de emoções das diferentes famílias e níveis. Ao final, a combinação de wavelet/nível que obteve o maior valor médio de *F-measure* (0,78) foi o Coif1 com 20 níveis de decomposição. Outro resultado interessante é que o reconhecimento automático obteve resultado mais eficiente que o resultado da percepção humana para a emoção de tristeza, independente da família wavelet. Em compensação, a percepção humana reconhece a emoção raiva melhor que o sistema automático.

Palavras-chave— Reconhecimento de Emoção na Fala, Transformada wavelet discreta, Berlin Emotional Database.

1 Introdução

Uma das formas mais naturais de interação entre seres humanos é a fala. De todas as formas de comunicação utilizadas por seres humanos, a fala é uma das mais rápidas, e sua agilidade fez com que esse meio de comunicação fosse bastante utilizado para implementar a interação do ser humano com máquinas.

No entanto, essa interação homem-máquina não é tão natural quanto a interação entre duas pessoas. Um dos motivos é que para que o entendimento da fala seja completo é preciso mais do que o significado da palavra falada, pois a forma como ela é expressa contém uma grande quantidade de informações necessárias para o entendimento (Nicholson et al., 1999). Através da expres-

são de emoções na fala é possível distinguir, por exemplo, a expressão “bem feito” com o significado positivo de “trabalho bem realizado” e do significado negativo de “prazer ao ver uma outra pessoa sendo prejudicada”. Esse é apenas um dos inúmeros casos em que emoções são essenciais para o entendimento mútuo de uma conversação.

Aplicações que necessitam de interação natural entre homem e máquina podem ser beneficiadas com o reconhecimento de emoções na fala, principalmente aquelas em que a resposta do sistema depende da emoção expressa pelo usuário. Em (Schuller et al., 2004), um sistema de carro detecta o estado mental do motorista, de maneira que o sistema possa iniciar procedimentos que vissem a segurança do condutor. Já em (France et al., 2000), a tecnologia foi utilizada para de-

terminar casos de depressão e potenciais riscos de suicídio em pacientes. Além disso, no trabalho de (Hansen and Cairns, 1995) apresenta sistemas de reconhecimento de voz para identificação de estresse em cabines de aeronaves. O reconhecimento de emoções da fala também foi utilizado para adaptar a resposta de sistemas de “call center” ao detectar frustração ou aborrecimento da voz do cliente (Ma et al., 2006).

Apesar da grande utilidade existente no reconhecimento de emoções na fala, El Ayadi e colegas (2011) citam diversas questões que tornam a tecnologia bastante desafiadora. Uma delas é que não se conhece exatamente quais são as características da fala que possuem maior impacto para reconhecer determinadas emoções. Além disso, é possível que haja mais de uma emoção em um único sinal de voz, em que cada emoção está presente em um diferente trecho da fala. Essa característica torna trabalhosa a tarefa de identificar qual trecho de voz carrega uma determinada emoção. Por fim, emoções iguais podem ser expressas de maneiras diferentes dependendo do falante, seu sotaque e sua cultura.

Nesse contexto, este trabalho propõe um estudo comparativo de extração de características baseadas em wavelets (Graps, 1995; Mallat, 2008) variando a família e os níveis de decomposição para reconhecimento de emoções da fala, a fim de determinar qual das famílias/níveis são as que possuem melhor desempenho, isto é, apresentam melhor combinação das medidas estatísticas de precisão e sensibilidade. Escolheu-se como base de dados de entrada, o EmoDB - *Berlin of Emotional Speech* (Burkhardt et al., 2005), que possui áudios com sete emoções: Raiva, Desgosto/Nojo, Medo, Alegria, Tristeza, Tédio e Neutro. Cada áudio possui apenas uma única classificação de emoção, logo, não há variação de emoção no mesmo áudio.

2 Metodologia e Referencial Teórico

Sistemas de reconhecimento de emoções comumente possuem 4 (quatro) etapas de processamento: pré-processamento, extração de características, seleção de características e classificação de características. A Figura 1, traduzida de (Bhatti et al., 2004), ilustra a arquitetura genérica deste tipo de sistema.

A entrada dos dados do sistema é uma base de dados composta de diversas gravações da fala humana expressando diferentes emoções, a *Berlin Database of Emotional Speech* (EmoDB) (Burkhardt et al., 2005), que possui 535 áudios. Sua escolha se deu por dois motivos: por ser uma base de dados pública e gratuita, e por ter sido utilizada em diversos trabalhos correlatos, tais como em (Kotti and Paternò, 2012; Pan et al., 2012). A EmoDB foi desenvolvida seguindo rigorosos testes de gravação e validação de seus dados com o

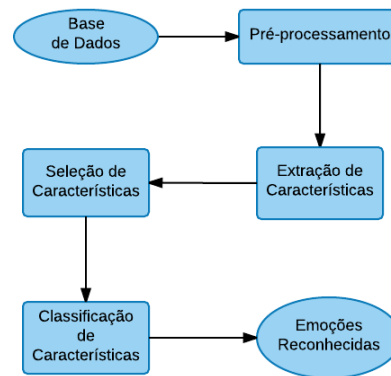


Figura 1: Arquitetura genérica de sistemas de reconhecimento de emoção da fala.

intuito de garantir qualidade das emoções expressas nos áudios verbalizadas por atores profissionais, logo, não há a necessidade da etapa de pré-processamento para retirada de ruídos ou melhoria do áudio.

Este trabalho foca apenas na questão de qual é o melhor conjunto de características baseada em Wavelets, mantendo todas as demais etapas idênticas. Decidiu-se também que não haveria a seleção (diminuição) de características e se manteria sempre o mesmo classificador, isto é, mantendo todo o sistema inalterado variando apenas a etapa de extração de características baseada em Wavelets. A classificação ocorreu através do algoritmo LDA (*Linear Discriminant Analysis*) (Balakrishnama et al., 1999). A arquitetura do sistema de testes deste trabalho é apresentada na Figura 2.

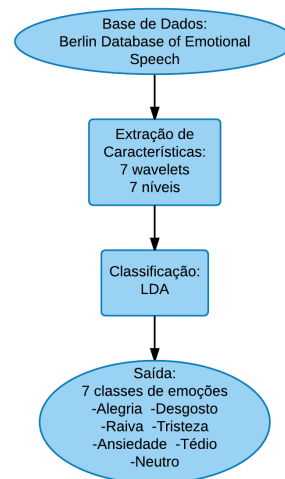


Figura 2: Arquitetura do sistema de testes.

2.1 Extração de Características

A extração de características é a fase em que os áudios são analisados e um conjunto de descritores é extraído, de modo que possa ser posteriormente

utilizado para classificar as emoções. Há diversas técnicas e algoritmos que possibilitam a extração de características, como a Fourier Transform e Wavelet Transform.

Wavelets são funções que satisfazem certos requisitos matemáticos e são usadas para representar dados e outras funções, semelhante a Fourier. O processamento de wavelets processa dados em diferentes escalas e resoluções, permitindo que sejam vistos tanto as informações globais quanto os detalhes de um sinal. Uma wavelet, diferentemente de senos e cossenos, base da análise de Fourier, tem duração limitada. Além disso, enquanto senóides são suaves e previsíveis, wavelets tendem a ser irregulares e assimétricas (Graps, 1995).

O procedimento da análise wavelet é realizada através da adoção de uma função base, ou ainda uma *mother wavelet*. Os coeficientes são ordenados em dois padrões dominantes, implementados por bancos de filtros. O primeiro padrão suaviza o sinal, enquanto o segundo mostra os detalhes do mesmo. A decomposição do sinal é feita em duas partes iguais, uma contendo os coeficientes do detalhe, e a outra os coeficientes de aproximação. A decomposição é então reaplicada na parte suavizada, dividindo o sinal em outras duas partes iguais, detalhe e suavização. Cada iteração do processo representa um nível, ao final do processo teremos decomposto o sinal em $J+1$ caminhos diferentes, em que J é a quantidade de níveis escolhido para executar o processo.

Foram realizadas extrações utilizando sete famílias de wavelets e sete níveis de decomposição diferentes (5, 10, 15, 20, 25, 30 e 35) - totalizando 49 (quarenta e nove) diferentes conjuntos de características para cada um dos 535 áudios da base de entrada. Para cada nível, calcula-se quatro características - energia, entropia, variância e comprimento de onda. As seguintes famílias de wavelets foram escolhidas para a extração de características:

- Biorthogonal: Bior2.8 e Bior3.9
- Coiflets: Coif1 e Coif5
- Daubechies: Db2, Db6 e Db8

2.2 Medidas Estatísticas

Para cada um dos 49 (quarenta e nove) conjuntos de características, coleta-se o resultado do classificador sob a forma de matrizes de confusão. Um exemplo da matriz de confusão gerada para cada experimento é mostrada a seguir:

Cada linha da matriz representa a emoção real e cada coluna a emoção reconhecida pelo sistema. Assim, quanto maior o valor na diagonal principal, melhor foi o resultado.

Com base nas matrizes de confusão, é possível calcular as medidas: *precision*, *recall* e *F-measure*.

Bior3.9 - 30 Níveis

Emoções	Classificadas						
	Alegria	Neutro	Raiva	Tristeza	Ansiedade / Medo	Tédio	Desgosto
Alegria	48	2	11	0	2	7	1
Neutro	2	63	0	0	4	10	0
Raiva	10	2	107	2	1	4	1
Tristeza	2	4	0	47	0	7	2
Ansiedade / Medo	4	6	4	1	45	9	0
Tédio	0	17	0	2	3	59	0
Desgosto	6	1	1	2	2	7	27

Figura 3: Exemplo de uma matriz de confusão: Bior3.9 com 30 níveis de decomposição.

Por fim, gráficos foram gerados com o intuito de melhor visualizar e comparar o desempenho das wavelets/níveis, estes gráficos serão detalhados na seção seguinte.

3 Experimentos, Resultados e Discussão

Todo o sistema para a execução dos testes foi desenvolvido usando a ferramenta Matlab e o código de extração de características DWT foi baseado no sistema de Kushaba (2013).

Foram feitas três análises: avaliação de qual(is) é(são) o(s) melhor(es) nível(is) de decomposição, avaliando qual a melhor combinação família/nível e comparando-se os resultados do sistema automático com os resultados da percepção humana.

3.1 Avaliação por nível de decomposição

Um dos gráficos gerados para cada experimento foi o gráfico de linha, um exemplo é mostrado na Figura 4. O gráfico apresentado mostra o resultado de *F-measure* para a família Biorthogonal (Bior2.8 em azul e Bior3.0 em vermelho), e nesse caso para a emoção da tédio. Seu eixo x representa o nível de decomposição utilizado no algoritmo de extração, já o eixo y indica o *F-measure* resultante desse nível. Por fim, são traçadas um número de linhas igual à quantidade de wavelets utilizadas da família analisada.

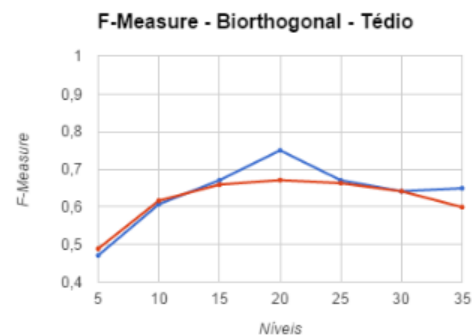


Figura 4: Exemplo de um gráfico de linha para família Biorthogonal na emoção Tédio.

Independente da família e emoção, os gráficos apresentaram os melhores resultados para os níveis 20 e 25. Em quase todos os gráficos, as curvas começam com valores menores, passam por um pico e depois tem a tendência a diminuir o *F-measure*. Logo, para as seções posteriores, manteremos as discussões focadas apenas nestes dois níveis.

3.2 Em busca da melhor família

O objetivo dos gráficos de radar da Figura 5 é mostrar os valores de *F-measure* alcançados pelos melhores níveis de decomposição (20 e 25) em cada uma das emoções estudadas. Cada radar representa uma emoção e possui sete eixos, duas linhas e catorze pontos. Os eixos representam as wavelets utilizadas na etapa de extração de características, já as duas linhas indicam os níveis que foram utilizados nesse experimento, por fim, os pontos representam o *F-measure* alcançado pela combinação wavelet/nível.

Começando pelo gráfico de alegria (Figura 5(a)), os melhores resultados são obtidos pelo nível 25, pois os pontos do nível 25 estão mais afastados do centro do que os do nível 20. Há uma exceção, para a wavelet DB2 que é melhor no nível 20. A disparidade entre os dois níveis é melhor observada para a wavelet Coif5. Essa mesma wavelet, no nível 25, é responsável pelo melhor resultado, junto a Bior2.8 nos dois níveis. Coif1 e Bior3.9, ambas no nível 20, alcançam o pior desempenho para a emoção.

Contrariamente ao ocorrido no radar de alegria, os valores de *F-measure* alcançados pelo nível 20 no radar de desgosto são maiores do que aqueles alcançados pelo nível 25, independente da wavelet. Há uma exceção, no entanto, a Bior3.9 com 25 níveis é levemente superior a Bior3.9 com 20 níveis. Não só isso, como também é essa a combinação de wavelet/nível que possui o melhor *F-measure* para a emoção de desgosto. Quanto à pior combinação, é difícil precisar, uma vez que, de modo geral, as wavelets alcançaram desempenhos similares.

As emoções de raiva e tristeza foram aquelas que obtiveram os melhores resultados, independente de wavelet ou nível. Também foi visto que, para raiva, a quantidade de níveis pouco influenciava no desempenho final do classificador. Observando a Figura 5(c), o resultado independe da wavelet utilizada para raiva.

Na Figura 5(d), de tristeza, o nível 20 se destaca sobre os resultados do nível 25, na maioria das wavelets. A diferença entre os dois níveis pode ser observada mais facilmente para as wavelets Bior2.8, Coif5 e Db8. Com relação a wavelet que melhor representa a emoção, o melhor resultado é dada pela Coif1 com 20 níveis, seguido da Db6 com 25 níveis e Bior2.8 com 20 níveis. Já a pior classificação fica por conta da Coif5 com 25

níveis.

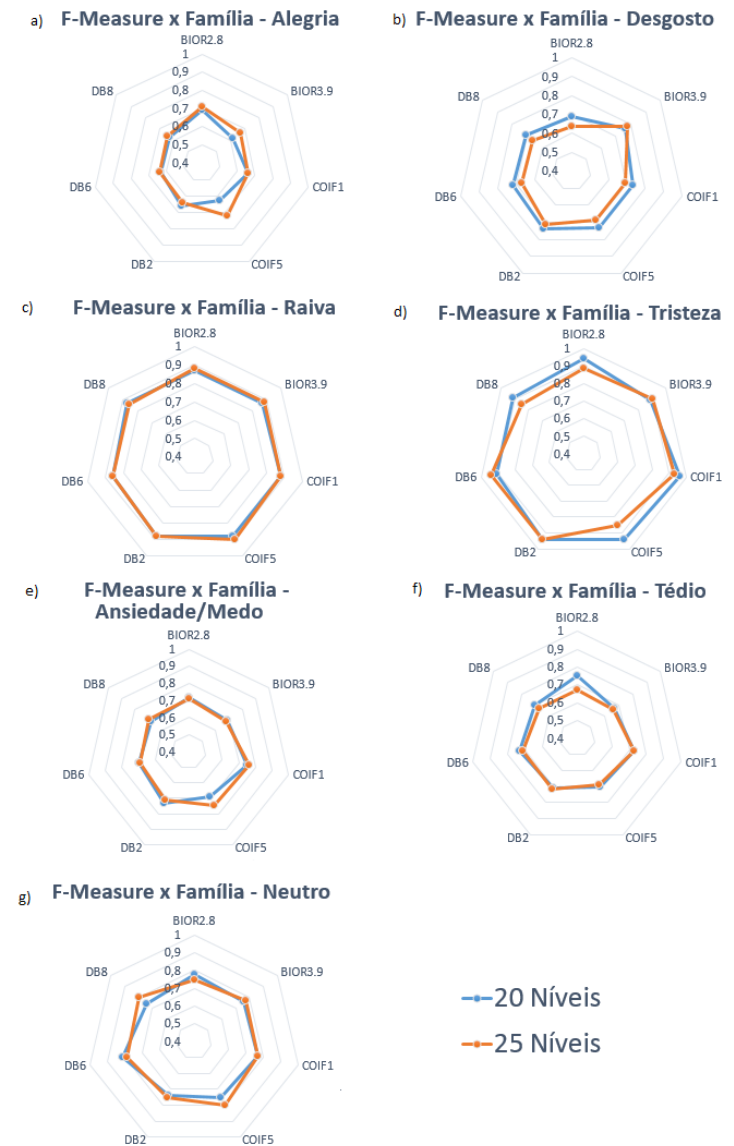


Figura 5: Gráficos de radar que exibem o *F-measure* alcançado por todas as famílias, com os níveis 20 e 25, para todas as emoções.

As duas emoções seguintes, ansiedade e tédio, possuem gráficos semelhantes em alguns pontos e diferentes em outros. Sua semelhança se dá na proximidade dos valores de *F-measure* alcançados pelos níveis 20 e 25 para cada wavelet. No entanto essa proximidade não ocorre em todos os casos. Para a emoção de ansiedade/medo a disparidade ocorre na Coif5, em que a linha dos 25 níveis é superior à dos 20 níveis. Já para o tédio essa diferença ocorre principalmente na Bior2.8, mas dessa vez é a linha dos 20 níveis que apresenta melhores resultados que 25 níveis. Com relação as wavelets que obtém os melhores resultados para as emoções, a família Coiflets merece destaque, uma vez que o sentimento de ansiedade foi melhor representado por seus integrantes, Coif1 e Coif5, ambas com 25 níveis. Já para o tédio, a wavelet que

se destaca é a Bior2.8 com 20 níveis. De maneira geral o nível que melhor representa a emoção de ansiedade/medo é o 25, enquanto que os melhores resultados do tédio são alcançados pelo nível 20.

A última emoção a ser analisada, neutro, não apresenta superioridade clara de um nível para outro conforme vem acontecendo com os outros sentimentos. Em alguns momentos, o nível 20 obtém os melhores resultados para determinada wavelet, como para as Bior2.8 e Db6, e em outros casos é o nível 25 que é superior, como para Coif5, Db2 e Db8. De maneira geral vê-se que todas as wavelets obtiveram resultados semelhantes, mas é necessário dar destaque para a família Daubechies, que com duas representantes, Db6 com 20 níveis e Db8 com 25 níveis, conseguiu o melhor *F-measure* emparelhada com a Coif5 com 25 níveis. Curiosamente, a família Daubechies também conseguiu as piores classificações através de Db2 e Db8 com 20 níveis em ambas.

Para analisar os *F-measures* obtidos por todas as wavelets em seus melhores níveis e para todas as emoções, foram somados todos os valores de *F-measures*, denominado de *F-measure* cumulativo. A combinação de maior *F-measures* acumulado foi a wavelet Coif1 com 20 níveis com *F-measure* acumulado de 5,48.

3.3 Comparação com a Percepção Humana

A EmoDB fornece os resultados de experimentos (via medida *Precision*) feitos com seres humanos. Essa seção visa comparar os resultados da percepção humana com os resultados obtidos por cada uma das wavelets/nível deste trabalho.

Para isso, foram criados sete gráficos de radar (Figura 6), em que cada radar representa uma wavelet. Além disso, cada gráfico possui 7 (sete) eixos, 3 (três) linhas e 21 (vinte e um) pontos. Os eixos representam as emoções estudadas, já as linhas indicam os níveis que foram utilizados nesse experimento (20 - laranja e 25 - cinza) e a taxa de percepção humana (azul), por fim, os pontos representam o precision alcançado por um nível em uma determinada emoção, ou ainda a taxa de percepção humana alcançada em uma determinada emoção. Note que a curva da percepção humana é a mesma em todos os gráficos.

Para a maioria das emoções, a taxa de reconhecimento humana supera, em muito, o desempenho do sistema de reconhecimento. No entanto há duas exceções relevantes que precisam ser destacadas: tristeza e desgosto. A primeira delas pode ser vista em todos os gráficos, independente da wavelet ou nível sob análise: a superioridade do sistema de reconhecimento para a emoção de tristeza. A disparidade entre o sistema e a percepção humana é bastante considerável em todos os casos, mas em especial para Coif1. Essa dificuldade do ser humano em perceber a emoção

de tristeza não era esperada, já que os resultados obtidos em todas as outras emoções foram bons. Como a percepção humana para a emoção de desgosto apresenta baixo desempenho, a disparidade homem-máquina foi pequena para as combinações Bior3.9 com 25 níveis, e Db2 com 20 níveis. Ambas aproximaram-se bastante do desempenho humano, mas a Bior3.9 apresentou resultados melhores, similar ao resultado obtido pelos humanos.

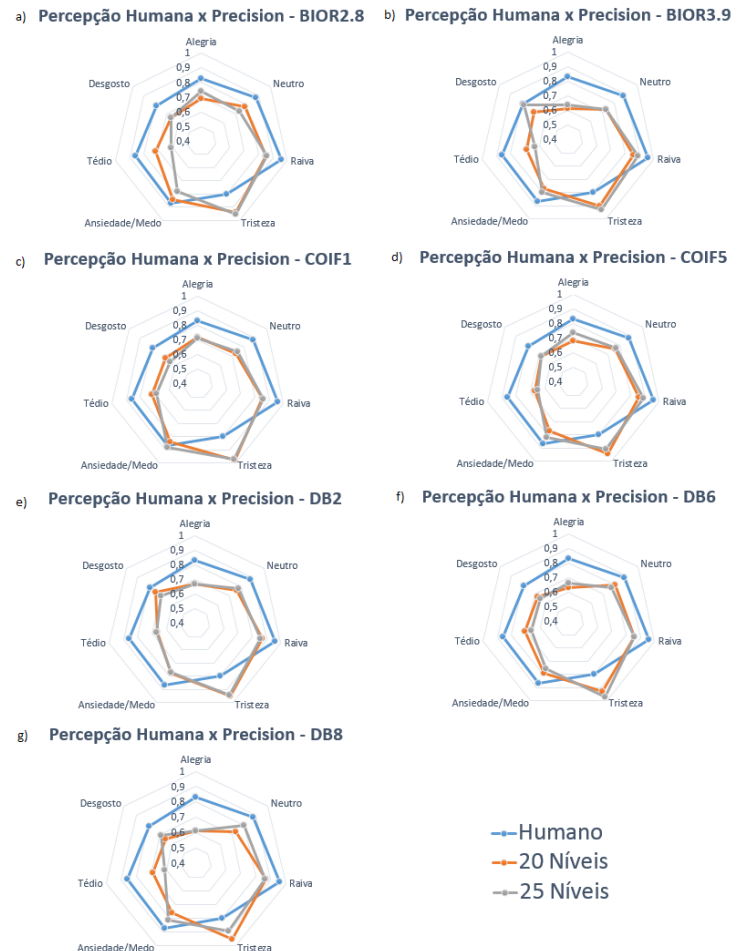


Figura 6: Gráficos de radar que exibem o precision alcançado por todas as famílias, com os níveis 20 e 25, assim como o grau de percepção humana para cada emoção.

Outra emoção em que o sistema conseguiu bons resultados, em algumas combinações, em relação ao reconhecimento humano foi a ansiedade. As wavelets Bior2.8 e Coif1 com 20 níveis, assim como a Coif5 com 25 níveis conseguiram resultados similares ao da percepção humana. No entanto, a Coif1 com 25 níveis superou, por pouco, o desempenho humano.

A percepção humana conseguiu resultados muito superiores para a raiva, e também ocorre nos três sentimentos restantes: alegria, neutro e tédio. A grande diferença para a alegria já era esperada, uma vez que o sistema automático obteve seus piores resultados para esse sentimento.

O maior valor de *Precision* acumulado foi alcançado por Coif5 com 25 níveis (4,55), mas é menor que o desempenho da percepção humana, *Precision* de 5,19.

4 Considerações Finais

O objetivo deste trabalho foi comparar conjuntos de características baseadas em wavelets com o intuito de verificar o comportamento destas no reconhecimento de sete emoções expressas na fala. Os resultados obtidos mostram que no geral, independente da wavelet ou nível utilizado, o melhor desempenho do reconhecimento é alcançado para as emoções de tristeza, seguida pela raiva. Por outro lado, as emoções que tem a pior taxa de reconhecimento são a alegria, seguida pelo tédio.

Com relação aos níveis de decomposição, valores dos extremos da faixa analisada, os mais baixos ou os mais altos, não obtém bom desempenho, sendo 20 e 25 os níveis que apresentaram os melhores resultados, independente da família. Em termos de família, os melhores resultados foram obtidos pela família Coiflets. Ao final, a combinação de wavelet/nível que alcançou o melhor desempenho foi Coif1 com 20 níveis de decomposição.

Comparando-se com a percepção humana, o sistema automático é bem melhor para reconhecimento de tristeza, enquanto o ser humano é melhor para o reconhecimento da raiva.

Como trabalhos futuros, pretende-se avaliar os resultados para bases de dados em outros idiomas, comparar outras técnicas de extração de características, implementar a etapa de seleção de características, utilizar outros classificadores e implementar a validação cruzada.

Referências

- Balakrishnama, S., Ganapathiraju, A. and Picone, J. (1999). Linear discriminant analysis for signal processing problems, *Southeastcon'99. Proceedings. IEEE*, IEEE, pp. 78–81.
- Bhatti, M. W., Wang, Y. and Guan, L. (2004). A neural network approach for human emotion recognition in speech, *Circuits and Systems, 2004. ISCAS '04. Proceedings of the 2004 International Symposium on*, Vol. 2, pp. II–181–4 Vol.2.
- Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. and Weiss, B. (2005). A database of german emotional speech, in *Proceedings of Interspeech, Lissabon*, pp. 1517–1520.
- El Ayadi, M., Kamel, M. S. and Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases, *Pattern Recogn.* **44**(3): 572–587.
- France, D. J., Shiavi, R. G., Silverman, S., Silverman, M. and Wilkes, M. (2000). Acoustical properties of speech as indicators of depression and suicidal risk, *IEEE Transactions on Biomedical Engineering* **47**(7): 829–837.
- Graps, A. (1995). An introduction to wavelets, *IEEE Comput. Sci. Eng.* **2**(2): 50–61.
- Hansen, J. H. L. and Cairns, D. A. (1995). Icarus: Source generator based real-time recognition of speech in noisy stressful and lombard effect environments, *Speech Commun.* **16**(4): 391–422.
- Khushaba, R. (2013). Feature extraction using multisignal wavelet transform decomposition. URL: <http://www.mathworks.com/matlab-central/fileexchange/37950-feature-extraction-using-multisignal-wavelet-transform-decomposition>
- Kotti, M. and Paternò, F. (2012). Speaker-independent emotion recognition exploiting a psychologically-inspired binary cascade classification schema, *International Journal of Speech Technology* **15**(2): 131–150.
- Ma, J., Jin, H., Yang, L. T. and Tsai, J. J. P. (eds) (2006). *Ubiquitous Intelligence and Computing, Third International Conference, UIC 2006, Wuhan, China, September 3-6, 2006, Proceedings*, Vol. 4159 of *Lecture Notes in Computer Science*, Springer.
- Mallat, S. (2008). *A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way*, 3rd edn, Academic Press.
- Nicholson, J., Takahashi, K. and Nakatsu, R. (1999). Emotion recognition in speech using neural networks, *Neural Information Processing, 1999. Proceedings. ICONIP '99. 6th International Conference on*, Vol. 2, pp. 495–501 vol.2.
- Pan, Y., Shen, P. and Shen, L. (2012). Speech emotion recognition using support vector machine, *International Journal of Smart Home* **6**(2): 101–108.
- Schuller, B., Rigoll, G. and Lang, M. (2004). Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture, *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, Vol. 1, pp. I–577–80 vol.1.