LSF na Verificação de Orador

Hugo Cordeiro Carlos Meneses

M2A/ISEL – Grupo de Multimédia e Aprendizagem Automática
Instituto Superior de Engenharia de Lisboa
Rua Conselheiro Emídio Navarro nº1, 1950-062 Lisboa, Portugal
{hcordeiro,cmeneses}@deetc.isel.ipl.pt
+351.218317224

Resumo

Este artigo descreve um sistema de verificação de orador. Pretende-se despertar para possíveis alternativas aos métodos tradicionalmente utilizados no reconhecimento de orador. Os oradores são caracterizados através dos coeficientes LSF e os resultados são comparados como os tradicionais coeficientes MFCC. Nos resultados obtidos verifica-se um desempenho semelhante entre os coeficientes LSF, agora propostos, e os MFCC.

O método de classificação implementado é o SVM, sendo o corpora utilizado o "2002 NIST Speaker Recognition Evaluation Corpus".

1. Introdução

O reconhecimento de orador divide-se em verificação e identificação. A verificação de orador é o processo que determina, perante um sinal de fala produzido por um orador, se este é, de entre um conjunto de oradores conhecidos, quem clama ser. Esta distingue-se da identificação de orador, que determina quem é o orador, de entre um conjunto conhecido de oradores, sem à partida pressupor a sua identidade.

Depara-se com a possibilidade de aplicar a verificação de orador essencialmente no controlo de acessos, seja acesso físico a determinado local, seja a bases de dados ou a aplicações pessoais bancárias via *internet*. É também ajuda preciosa em casos judiciais, em que seja necessário provar a autoria de declarações gravadas.

Este artigo descreve testes de verificação de orador, propondo caracteriza-lo através dos coeficientes LSF (*Line Spectrum Frequencies*)[1], em vez dos tradicionais MFCC (*Mel Frequency Cepstral Coefficients*)[2].

A secção seguinte apresenta as características e o motor de reconhecimento utilizado. Na secção 3 é apresentado o *corpus* e na secção 4 o método de avaliação. Na secção 5 são descritos os testes efectuados e apresentados os resultados obtidos. O artigo termina com a secção dedicada às conclusões e trabalho futuro.

2. Características e motor de reconhecimento

Os MFCC são os coeficientes vulgarmente utilizados como caracterizadores em aplicações de reconhecimento orador e também de reconhecimento de fala. Estes têm a particularidade, pela introdução de um banco de filtros com escala Mel, de ter em conta as propriedades acústicas do ouvido humano.

Por outro lado, os coeficientes LSF foram utilizados com sucesso em adaptação ao orador no contexto da codificação fonética [3], mas embora com informação directa sobre os formantes de um segmento fonético,

praticamente só têm sido utilizados para quantificação dos coeficientes de predição linear em aplicações de codificação de fala. Pretende-se com este estudo verificar até que ponto estes coeficientes terão informação relevante no âmbito do reconhecimento de orador, pelo que será apresentada uma comparação entre os resultados obtidos com coeficientes LSF e MFCC.

Com vista a melhorar a taxa de reconhecimento foram introduzidas os coeficientes Delta-MFCC [2], incluindo agora o termo da energia. À sua semelhança, foram testados coeficientes Delta-LSF, para entender da relevância da dinâmica destes coeficientes.

A utilização de SVM (*Support Vector Machines*) [4] como motor de reconhecimento surge actualmente como alternativa [5] ao tradicional modelo de misturas de Gaussianas (GMM – *Gaussian Mixture Model*), vulgarmente utilizado [6] em sistemas de reconhecimento de orador independente do texto.

3. Corpus

O *corpus* de fala utilizado foi o "2002 NIST Speaker Recognition Evaluation *Corpus*" [7], que se tem tornado uma referência nesta área. Deste *corpus* foi usada a componente celular gravada a partir de sistemas CDMA e GSM, em ambos os casos com diferentes telefones e consequentemente com diversos microfones e auscultadores. Nos primeiros testes apresentados apenas foram utilizados 10 oradores masculinos. Os dados de treino correspondem a uma instância de 2 minutos por orador. Os dados de teste utilizados correspondem a 5 instâncias com uma duração de cerca de 30 segundos cada uma.

A situação para a qual foram obtidos os melhores resultados com 10 oradores, foi testada com todo o conjunto de teste do *curpus*, 139 oradores correspondendo a 1398 ficheiros de teste. Cada ficheiro de teste é comparado com 11 oradores, como sugerido pela NIST [8].

4. Avaliação

Ao contrário da identificação de orador, em que é identificado o orador cuja distância for mais próxima da do orador de entrada, na verificação de orador é necessário estimar um limiar a partir da qual se assume que o orador é quem clama ser. Um valor deste limiar demasiado elevado faz com que mais oradores que clamam ser quem de facto são, que se denominam de clientes, sejam rejeitados. À relação entre clientes rejeitados e número total de clientes denomina-se de taxa de rejeição falsa (FRR – *False Rejection Ratio*). De igual modo, com o aumento deste limiar, o número de oradores que clamam ser quem de facto não são, que se denominam de impostores, será também diminuído. À relação entre impostores rejeitados e número total de impostores denomina-se de taxa de aceitações falsas (FAR – False Acceptances Ratio). De modo oposto, uma diminuição do limiar de comparação aumentará a FRR e diminuirá a FAR. Existe assim, claramente, um compromisso entre as taxas FRR e FAR, função do limiar de comparação. À curva FRR em função do FAR função do limiar de comparação dá-se o nome de curva ROC (*Receiver Operating Characteristic*). Dois pontos importantes na curva ROC são o EER (*Equal Error Rate*), em que as taxas FRR e FAR são iguais, e o mínimo do HTER (*Half Total Error Rate*), definido como a média entre o FRR e o FAR para cada limiar de comparação.

Os sistemas de verificação desenvolvidos foram avaliados e comparados com base no valor do EER e do mínimo do HTER, para além de uma avaliação gráfica baseada por um lado na curva DET (*Detection Error Tradeoff*), uma variação da curva ROC, por outro na curva do HTER em função do limiar de comparação.

5. Evolução dos testes e resultados

Os sinais de fala do *corpus* são segmentados de modo a retirar as zonas de silêncio com base num algoritmo de segmentação proposto por Lamel [9]. Após a segmentação são calculados os coeficientes LSF ou MFCC em tramas de 20 ms, embora com um andamento entre tramas de 10 ms.

Os primeiros testes efectuados tiveram como principal objectivo definir um sistema de referência, realizado utilizando um quantificador vectorial. Este quantificador forneceu também a melhor dimensão para o livro de código a vir a ser utilizado nos SVM. Posteriormente foram realizados os restantes testes com SVM e aferidas várias normalizações para as características implementadas.

Quantificação Vectorial

O primeiro método de classificação desenvolvido foi baseado em quantificação vectorial. Neste, foi treinado um quantificador para cada um dos oradores a verificar, após segmentados em tramas os dados de treino. Cada trama do sinal de entrada é quantificada e a distância à palavra de código mais próxima calculada. Se esta distância for menor que determinado limiar, a trama é considerada como sendo do orador em causa. O orador é considerado um cliente se a percentagem de tramas classificadas neste orador for superior a um segundo limiar.

Os melhores resultados foram obtidos com tramas de 20 ms e com um andamento entre tramas de 10 ms, gerando 512 palavras de código. Para coeficientes MFCC, sem o termo de energia, os melhores resultados foram conseguidos com ordem 16. De modo a diminuir o efeito dos diferentes sistemas de comunicação, microfones e auscultadores, foi retirado o valor médio que cada um dos coeficientes cepstrais (equalização cega de canal). Foi obtido um valor de EER de 24% e do mínimo do HTER de 21%. Para coeficientes LSF os melhores resultados foram conseguidos também com ordem 16, tendo sido obtido um valor de EER de 16% e do mínimo do HTER de 14.4%.

Suport Vector Machines

O SVM foi treinado com o *kernel* RBF (*radial basis function*) por cada orador a verificar, tendo como contra exemplo os 9 oradores restantes, utilizando os coeficientes correspondentes aos quantificadores vectoriais descritos anteriormente. A utilização das tramas individuais não é possível, pois o número de tramas dos 9 oradores é muito maior que a dos oradores individuais, pelo que foi criado um quantificador misturando os sinais destes oradores. O desempenho para um orador é avaliado como a percentagem de tramas dos dados de entrada quantificados nesse orador. Para coeficientes MFCC foi obtido um valor de EER de 10.8% e do mínimo do HTER de 10%. Para 16 coeficientes LSF foi obtido um valor de EER de 12% e do mínimo do HTER de 11.4%. Este número de coeficientes foi o que maximizou quer o EER quer o HTER.

Uma das dificuldades no treino dos SVM foi o de determinar o valor da variância do *kernel* RBF, pois os dados de entrada apresentam uma grande variabilidade. Para obviar a este problema os dados foram normalizados de modo a que cada coeficiente tivesse média nula e variância unitária. Nestas circunstâncias um valor da variância entre 0.5 e 0.8 não altera significativamente o desempenho do sistema de verificação, tornando-se praticamente independente dos dados de entrada e melhorando o desempenho em relação ao sistema sem normalização. Para coeficientes MFCC foi obtido um valor de EER de 10.5% e do mínimo do HTER de 9.6%. Para coeficientes LSF foi obtido um valor de EER de 8.7%, igual ao do mínimo do HTER.

Os resultados obtidos com coeficientes diferenciais foram calculadas taxas de reconhecimento só com estes parâmetros. As taxas de reconhecimento dos coeficientes diferencias foram pesadas (50%) e somadas com as taxas pesadas (50%) obtidas nos coeficientes originais obtendo-se assim a taxa de reconhecimento total. Verificou-se uma melhoria com a combinação das características com os respectivos coeficientes Delta, quer nos MFCC quer nos LSF.

Os resultados de todos os testes efectuados são apresentados na tabela 1

	EER %	Limiar % de tramas	Mínimo HTER %	Limiar % de tramas
Quantificação vectorial – MFCC	24.0	29	21.4	30
Quantificação vectorial – LSF	16.0	40	14.4	46
SVM sem normalização – MFCC ($\sigma = 2$)	10.4	48	10.0	44
SVM sem normalização – LSF ($\sigma = 20$)	12.0	35	11.4	30
SVM com normalização – MFCC ($\sigma = 0.6$)	10.5	61	9.6	59
SVM com normalização – LSF ($\sigma = 0.6$)	8.7	47	8.7	47
SVM com MFCC + Delta ($\sigma = 0.6$)	10	56	9.1	57
SVM com LSF + Delta ($\sigma = 0.6$)	8	51	7	51

Tabela 1. Resultados obtidos com 10 oradores

As figuras seguintes ilustram as curvas DET e do HTER para todos os testes efectuados com SVM.

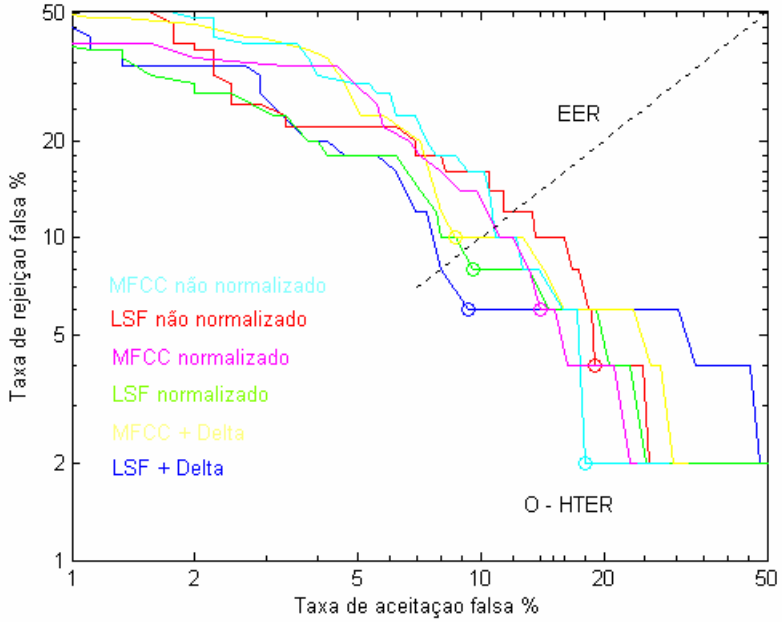

Figura 1. Curvas DET, resultados obtidos com SVM num conjunto de 10 oradores

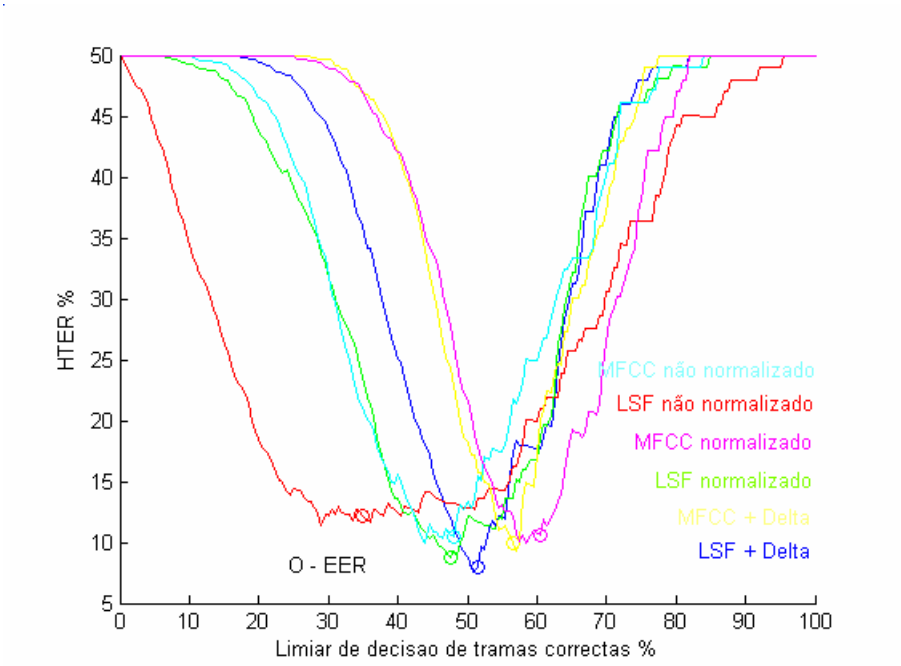

Figura 2. HTER, resultados obtidos com SVM num conjunto de 10 oradores

Por último foram realizados testes para todo o *corpus*, que é constituído por 139 oradores correspondendo a 1398 ficheiros de teste. Cada ficheiro de teste é avaliado contra 11 oradores, podendo ou não um deles ser um cliente. O treino do SVM consistiu na criação de um modelo do mundo com 20 oradores, retirados aleatoriamente do total dos 139. Como os ficheiros de treino têm de ser todos treinados, os 20 oradores que representam o mundo foram treinados no SVM cada um contra apenas os outros 19.

O teste realizou-se a partir dos modelos que obtiveram melhores resultados nos testes anteriores, ou seja, 16xMFCC+16xDela MFCC+Delta Energia, e 16xLSF+16xDelta LSF+Delta Energia. Os resultados obtidos são apresentados na tabela 2.

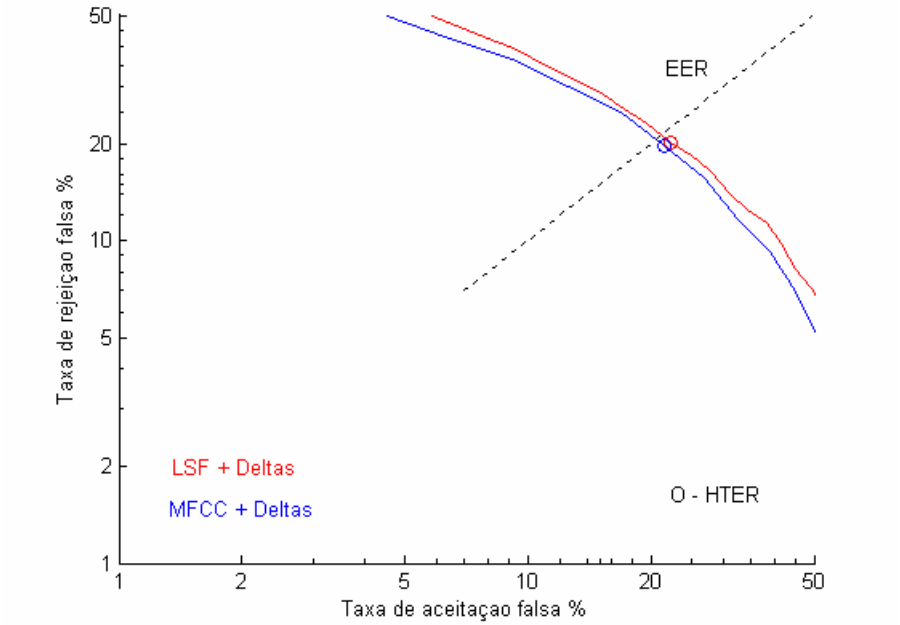

Figura 3. DET, resultados obtidos com 139 oradores e 1398 ficheiros de teste

	EER %	Limiar % de tramas	Mínimo HTER %	Limiar % de tramas
SMV com MFCC + Delta	20,6	73.1	20,5	73
SMV com LSF + Delta	21,3	58,3	21,1	60,2

Tabela 2. Resultados obtidos com 139 oradores e 1398 ficheiros de teste

6. Conclusões e trabalho futuro

Analisando os resultados obtidos verifica-se que o reconhecimento utilizando os coeficientes LSF têm um desempenho semelhante aos coeficientes MFFC. No caso particular com 10 oradores pode ser considerado um universo reduzido, assim foram efectuados testes com 139 oradores de modo observar o comportamento dos LSF num maior universo de oradores. No caso de 10 oradores os coeficientes LSF, mostraram um melhor desempenho que os MFCC, embora esta tendência se tenha invertido quando o teste compreendeu 139 oradores.

O trabalho futuro vai passar por comparar os resultados obtidos com SVM com GMM e perceber os comportamento dos coeficientes LSF e MFCC com esse motor de reconhecimento. Também vão ser testadas técnicas de redução de ruído, uma vez que os ficheiros de fala são gravados em ambiente telefónico.

Referências

[1] - F. Soong, B. Juang, "Line Spectrum Pair (LSP) and Speech Data Compression", *Proc. of the Int. Conf. Acoust., Speech and Signal Processing*, 1.10.1-1.10.4, 1984.

[2] T. Thrasyvoulou e S. Benton, "Speech parameterization using the Mel scale Part II", 2003.

[3] C. Meneses Ribeiro, I M. Trancoso, "Speaker Adaptation in a Phonetic Vocoding Environment", 1999 IEEE Workshop on Speech Coding, Haikko Manor, Porvoo, Finland,1999.

[4] Andrew Ng CS229 Lecture notes, 2004

[5] Vincent Wan e William M. Campbell, "Support Vector Machines For Speaker Verification And Identification". *IEEE Proceeding*, 2000

[6] Douglas A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models", 1995

[7] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2004S04

[8] The NIST Year 2002 Speaker Recognition Evaluation Plan, 2002

[9] Lori F. Famel, Lawrence R. Rabiner, Aarong E. Rosenberg e Jay G. Wilpon, "An Improved Endpoint Detector for Isolated Word Recognition", 1981