Universidade do Porto
FEUP Faculdade de Engenharia

Dissertação: **Reconhecimento de Orador em Dois Segundos**

Mestrado Integrado em Engenharia Electrotécnica e de Computadores, ramo Telecomunicações

Orientador: Prof. Aníbal Ferreira

Relatório realizador por: Diana Rocha Mendes

Relatório Nº 16/17

6 Junho 2011 a 19 Junho 2011

Tarefas Realizadas:

- Extensão dos testes indicados no último relatório. Extensão da parte vozeada considerada nos testes anteriores: foram incluídas para além de vogais todos os outros fonemas anotados vozeados. Execução dos testes com diferente distribuição dos tempos de treino – num primeiro foi usado o mesmo tempo de treino para a parte vozeada e para a voz completa, num segundo teste foi adaptada a proporção dos tempos de treino de forma a essa proporção corresponder à percentagem de partes vozeadas média nos segmentos TIMIT (cerca de 70%).
- Repetição dos testes para parte não vozeada – o que levou a um redimensionamento dos tempos de treino pelo facto de a quantidade de voz não vozeada ser insuficiente para continuar a usar o tempo de treino anterior. Passaram a ser usados 301 vectores de MFCCs em vez de 578.
- Análise comparativa dos resultados anteriores, selecção do número de gaussianas mais indicado em cada um dos cenários testados.
- Início dos testes com NRDs. Adaptação do algoritmo de extracção de NRDs e MFCCs fornecido (construído por Ricardo Sousa e Professor Aníbal Ferreira) ao sistema de reconhecimento em uso.
- Os primeiros testes foram feitos no software Weka, que utiliza o método Nearest Neighbour e 10-fold cross-validation para calcular a taxa de identificações correctas. Os resultados para as vogais retiradas da TIMIT não foram satisfatórios – nem em termos de níveis de performance geral (que se registou muito abaixo dos valores obtidos com GMMs, mesmo utilizando apenas MFCCs), nem em termos de contribuição dos NRDs ao desempenho do sistema – de facto a inclusão de NRDs prejudicou o desempenho.

- Face aos resultados anteriores foi utilizada a base de dados em que os NRDs foram já testados pelo colega Ricardo Sousa e cujos resultados foram publicados no artigo [1] – base de dados constituída por 8 oradores e 5 vogais cantadas por orador. Os resultados no Weka foram satisfatórios e até o desempenho geral foi algo superior, devido à substituição do algoritmo de extração de MFCCs pelo disponível na Voicebox. Não foi possível, no entanto, obter resultados semelhantes usando o classificador baseado em GMMs e no valor de log-likelihood. Isto deve-se provavelmente à escassez dos dados, que não permitem um treino adequado das componentes gaussianas.

- Teste em ambiente Weka das características NRDs e MFCCs, utilizando agora uma base de dados de voz falada. Esta é constituída por 6 oradores masculinos, em que cada um diz as cinco vogais. Cada segmento apresenta a parte estável da vogal, e tem precisamente 100 ms. Os resultados foram satisfatórios na medida em que os NRDs contribuíram para o desempenho do sistema (desempenho utilizando NRDs e MFCCs foi maior do que o desempenho utilizando apenas MFCCs). No entanto, a percentagem de identificações correctas foi mais baixa em cada um dos cenários comparativamente aos resultados com as vogais cantadas. Utilizando o classificador GMM os resultados foram idênticos aos descritos para as vogais cantadas.

- Estudo da influência da frequência de amostragem, do número de NRDs utilizado e da selecção prévia dos NRDs mais significativos (através do software Weka) para as várias bases de dados mencionadas.

Dificuldades Encontradas:

- As anotações da base de dados TIMIT não delimitam a parte estável de vozeamento nas vogais, e incluem por vezes zonas de silêncio, de arranque e outras partes não desejáveis para o teste do desempenho com NRDs. A selecção manual é demasiado morosa, de forma que não é uma opção viável nesta fase do trabalho. As bases de dados com essa delimitação feita adequadamente não contêm dados suficientes para o estudo do desempenho do sistema com modelação e classificação GMM. Desta forma a continuação do estudo terá de ser feita com base em ambiente Weka e com classificador Nearest Neighbour, e com outras bases de dados que não a TIMIT.

Próximas Tarefas:

- Organização dos dados obtidos nos testes descritos.
- Teste em Weka das vozes contidas na base de dados com vogais faladas, considerando também os sujeitos femininos e as crianças contidas na base de dados.
- Estudo do tempo de treino "limite" dos GMMs, através da variação entre 150 e 550 vectores de MFCCs para tempo de treino. O limite inferior é de 150 pelo facto do tempo de teste a partir do qual há uma descida mais acentuada de desempenho do sistema ser 150 vectores MFCCs (conclusão retirada a partir de testes feitos anteriormente), e pelo facto de não fazer sentido utilizar tempo de treino inferior ao tempo de teste.

Bibliografia:

[1] Sousa, R. e Ferreira, A., *Singing Voice Analysis Using Relative Harmonic Delay*, 2010