Relatório Científico - 1 de 2

Título do projeto:

Estimação automática de faixa etária pelo processamento do sinal de voz

Pesquisador responsável: Ivandro Sanches

Instituição sede do projeto: Centro Universitário FEI (FEI)

Equipe de pesquisa: Ivandro Sanches, professor, FEI

Número do processo FAPESP: 2016/18700-7

Período de vigência do projeto: 01/09/2017 a 31/08/2019

Período coberto pelo relatório científico em questão: 01/09/2017 a 30/08/2018

Assinatura do pesquisador responsável:

28 agosto 2018

Resumo

Este projeto objetiva avaliar a possibilidade de se estimar a faixa etária de um locutor pelo processamento de seu sinal de voz.

Esse conhecimento poderá ser usado de forma bastante positiva ao nortear aplicações eletrônicas e/ou telefônicas a tratar os usuários de forma apropriada com a sua idade. Essa capacidade de tratar pessoas diferentemente pela sua idade é naturalmente empregada pelos humanos e o sucesso de se estimar a faixa etária automaticamente trará esse recurso a qualquer aplicação que interaja com usuários pelo o uso da voz.

Vislumbram-se inúmeros benefícios, tanto para os usuários como para os que integrarem essa tecnologia aos seus dispositivos ou aplicações. Apenas para citar alguns exemplos, uma aplicação telefônica poderá evitar que uma criança acesse conteúdo impróprio; uma empresa poderá direcionar publicidade condizente com a idade do interlocutor; um robô poderá diferenciar o seu tratamento de acordo com a faixa etária de seus colocutores. Isto é, o conhecimento da faixa etária de uma pessoa poderá ser empregado de forma conveniente em interfaces humano-máquina, humano-robô e em telefonia com vantagens para todos envolvidos.

Para isso, estão sendo exploradas técnicas de aprendizado de máquina, *machine learning*, aliadas a recursos usados em biometria e reconhecimento automático de fala.

No estágio atual de desenvolvimento, com modelagem por *Gaussian Mixture Models*, a taxa de acerto para 4 classes (criança, jovem, adulto, senior) está em 58.6 %. Espera-se melhorar essa taxa com a introdução da frequência fundamental como parâmetro adicional na modelagem e pela implementação e uso de técnica baseada em variáveis latentes, como *i-vectors*.

Realizações no período

No período houve a implenentação, na linguagem C, da técnica de modelos de misturas de Gaussianas, GMM (*Gaussian mixture models*). A escolha da linguagem deu-se pela relativa eficiência e rapidez de processamento provida por essa linguagem de programação. A implementação nessa linguagem é mais longa e trabalhosa mas espera-se que os resultados poderão ser produzidos em tempo real, viabilizando a integração em aplicações interativas com os usuários. A escolha da modelagem por GMM deve-se ao sucesso desse modelo estatístico no problema da biometria de locutor. Os resultados obtidos servirão de *baseline* para comparação com os resultados de futuras implementações com técnicas mais avançadas e variações de características que serão apresentadas em mais detalhes na seção sobre atividades do próximo período.

Implementação

Um importante fator no processo de modelagem é a escolha dos parâmetros, *features*. Candidatos naturais são os utilizados em reconhecimento de fala e autenticação de locutores, nominalmente coeficientes mel-cepstrais, *mel-frequency cepstral coefficients*. Assim, um sinal poderá ser decomposto em uma sequência de vetores de coeficientes como indicado na expressão

$$\mathbf{X} = \{X_1, X_2, \ldots, X_T\}$$

Cada vetor $X_i, i = 1, 2, \ldots, T$ corresponde a um segmento de voz de cerca de 20 a 30 ms, eventualmente com sobreposição de 10 a 15 ms entre segmentos adjacentes. Cada vetor X_i é composto por dezenas de coeficientes mel-cepstrais que extraem e condensam informações relevantes dos segmentos do sinal analisado.

Assim, dada a observação $\mathbf{X}$ de um locutor, a tarefa de estimação da faixa etária, s, pode ser posta como um teste de hipótese entre H_0 e H_1,

$$H_0 : \mathbf{X} \text{ corresponde à faixa etária } s$$
$$H_1 : \mathbf{X} \text{ não corresponde à faixa etária } s$$

Na modelagem por GMM, as hipóteses H_0 e H_1 são representadas pelos modelos de "impressão etária" λ_s e modelo universal λ_0. Esses modelos são gerados na fase inicial de treinamento. Portanto, para um conjunto de vetores de observação $\mathbf{X} = \{X_i | i = 1, 2, \ldots, X_T\}$, o teste de hipótese é realizado ao se avaliar a seguinte razão de verossimilhança

$$\frac{p(\mathbf{X}|\lambda_s)}{p(\mathbf{X}|\lambda_0)} \begin{cases} \geq \tau \text{ aceitar } H_0 \\ < \tau \text{ rejeitar } H_0 \end{cases}$$

onde τ é um limiar de decisão. Usualmente a razão de verossimilhança é feita em escala logarítmica, resultando na *log-likelihood ratio*

$$\Lambda(\mathbf{X}) = \log p(\mathbf{X}|\lambda_s) - \log p(\mathbf{X}|\lambda_0).$$

Detalhes sobre as fases de treinamento e teste dos modelos GMM podem ser obtidos em diversas refências. Uma referência recente e que pode servir de ponto de partida para outros trabalhos é **Speaker Recognition by Machines and Humans** de John H. L. Hansen e Taufiq Hasan, publicada em *IEEE Signal Processing Magazine*, novembro de 2015.

Adicionalmente, scripts em Python foram desenvolvidos para análise dos resultados obtidos na fase de teste. A figura 1 apresenta trechos do programa desenvolvido em C, à esquerda, e trecho de script em Python à direita, no qual se obtém curva de falsa aceitação e falsa rejeição a ser apresenta na seção de resultados preliminares.

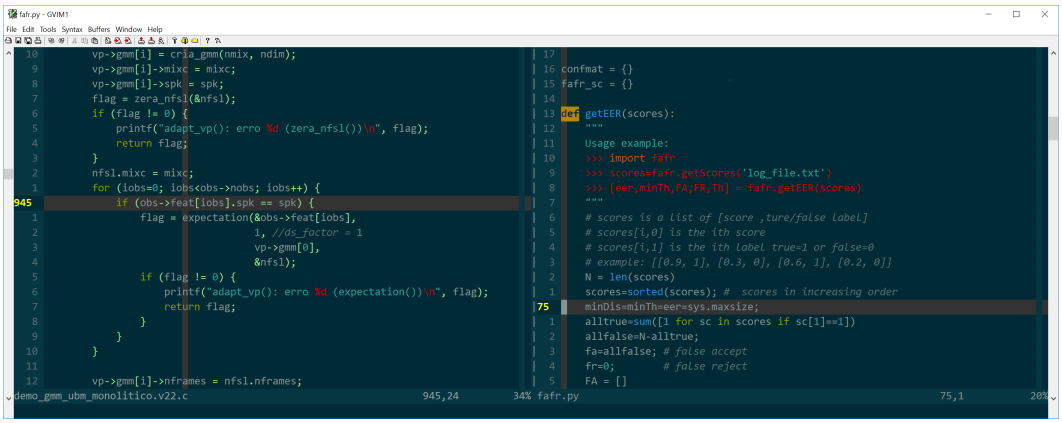

Figura 1: Trechos de códigos da implementação em C, à esquerda, e scripts em Python para análise dos resultados, à direita.

Até o momento, entre código em C e scripts em Python foram implementadas cerca de 3400 linhas de código. Ainda não foram considerados, por exemplo, os inúmeros scripts de suporte necessários para gerar o ambiente de experimentos contendo diversas listas de milhares de arquivos de áudio que participam das fases de treinamento e de testes dos algoritmos. Apenas para complementar informação, o comando a seguir apresenta a quantidade de linhas nos arquivos selecionados. O arquivo '`.c`' contém a implementação do processo de treinamento e produção de resultados. Os arquivos '`.py`' contêm os scripts em Python de suporte para análise dos resultados.

```
=> wc -l demo_gmm_ubm_monolitico.v22.c fafr.py read_save_features.py
2736 demo_gmm_ubm_monolitico.v22.c
 323 fafr.py
 362 read_save_features.py
3421 total
```

Para o teste da implementação é necessária uma base de dados estatisticamente significativa. A seção seguinte apresenta a base usada na avaliação da abordagem implementada.

Base de dados

A criação de uma base de dados local revelou-se inviável pela quantidade de trabalho necessário, duração do projeto, recursos disponíveis e dificuldades inerentes à tarefa. Por exemplo, como dificuldades inerentes claramente nota-se o grande e elaborado trabalho logístico que seria a gravação de um número considerável de crianças e em diferentes sessões de gravação.

Para contornar esse problema, recorreu-se a Deutsche Telekom AG Laboratories, Berlim, Germany. Após aceitar condições de uso da base de dados para pesquisa e desenvolvimento no Centro Universitário FEI, foi permitido acesso a uma base de dados de 47 h de gravação no canal telefônico de 770 participantes. Os áudios foram gravados à frequência de amostragem de 8 kHz, com amostras de 8 bits em A-law. A distribuição das faixas etárias obedece à tabela 1.

Tabela 1: Faixas etárias

		núm. locutores	
Classe	idade	treino	teste
child	7-14	69	38
young	15-24	114	69
adult	25-54	138	86
senior	55-80	150	106

Os participantes tiveram suas falas gravadas em 6 sessões, sendo cada sessão separada de outra por pelo menos 1 dia para evitar muita uniformidade nas pronúncias. Em cada sessão de gravação eram pronunciadas (lidas) 18 frases extraídas de um conjunto maior previamente estabelecido. As ligações foram feitas de aparelhos celulares, com 3 sessões feitas em ambiente fechado e 3 sessões feitas em ambiente externo para aumentar a variabilidade das condições de gravação. Os locutores foram remunerados pela participação e as ligações eram gratuitas.

A título de informação, o comando a seguir apresenta a quantidade total de arquivos de áudio, `'.raw'`, presentes no diretório `wav_traindevel/` e usados nos experimentos.

```
=> find wav_traindevel/ | grep .raw | wc -l
53076
```

Apesar da língua falada na base adotada ser o Alemão, isso não é um impedimento para a análise do problema pois não são empregadas quaisquer informações linguísticas na metodologia adotada neste projeto. Com isso, quer-se enfatizar que a informação etária do locutor possa ser extraída do sinal de voz independentemente do que se esteja falando (semântica), nível intelectual do locutor, sotaques, regionalismos, etc.

Resultados preliminares

A figura 2 ilustra o processo de avaliação, que consiste na fase inicial de **treinamento**, ou aprendizado, onde os modelos de cada faixa etária predefinida (modelos de impressão etária ou '*age prints*') são estimados; e na fase de **teste** na qual arquivos de áudio que não participaram da fase de treinamento serão usados para avaliar a eficiência do método.

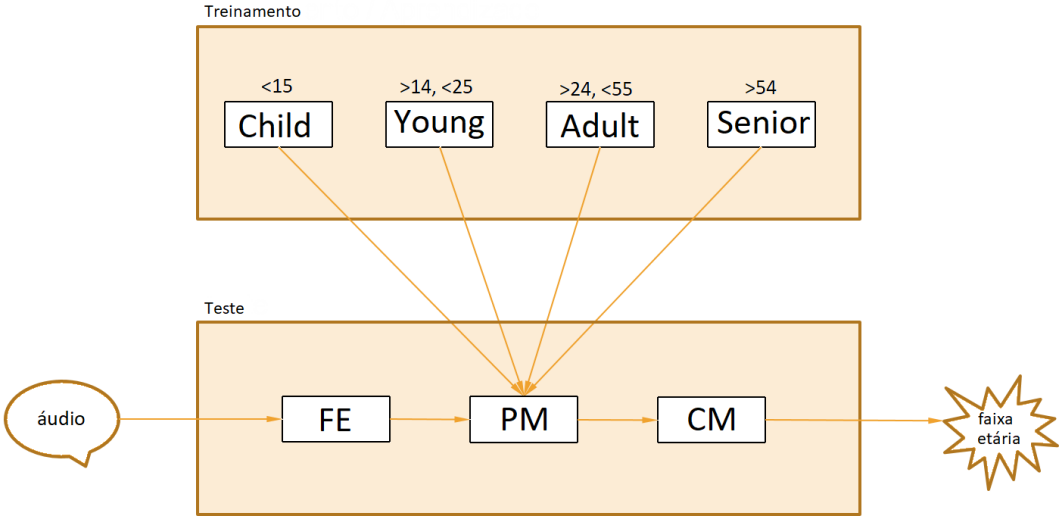

Figura 2: Fases de treinamento e teste.

O bloco **FE** corresponde ao *front-end*, responsável pela extração dos parâmetros (*features*) do áudio. Como já mencionado, adotou-se o uso de mel-frequency cepstral coefficients, MFCC. Abaixo encontra-se a configuração usada na obtenção desses parâmetros obedecendo à nomenclatura usada no pacote HTK da Universidade de Cambridge, UK (`http://htk.eng.cam.ac.uk/`).

```
# Exemplo: parametriza wavefile raw 8kHz para mfc no formato htk
# HCopy -C thisfile 1001/2/a11001s1.raw 1001/2/a11001s1.mfc
HEADERSIZE     = 0
BYTEORDER      = VAX
SOURCEKIND     = WAVEFORM
SOURCEFORMAT   = RAW
SOURCERATE     = 1250
ZMEANSOURCE    = TRUE
TARGETKIND     = MFCC_E_D_A
TARGETFORMAT   = HTK
TARGETRATE     = 100000
WINDOWSIZE     = 250000.0
NUMCHANS       = 24
ENORMALISE     = FALSE
SAVEWITHCRC    = FALSE
```

Resumidamente, um sinal de áudio `WAVEFORM RAW` amostrado em 8 kHz (`SOURCERATE`

= 1250) será convertido em MFCC (TARGETKIND = MFCC_E_D_A) usando banco de filtros com 24 canais (NUMCHANS = 24), tomando-se segmentos de áudio de 25 ms (WINDOWSIZE = 250000.0) a uma taxa de 10 ms (WINDOWSIZE = 100000). Esta extração de parâmetros deve ser idêntica para os sinais utilizados nas fases de treinamento e teste. Com isso, para cada arquivo de áudio será produzido um arquivo binário contendo sequência de vetores MFCC de dimensão 39, correspondentes a 12 coeficientes cepstrais estáticos mais 1 de energia, e os correspondentes delta e aceleração ($3 \times (12 + 1) = 39$).

O bloco **PM** corresponde ao algoritmo de *Pattern Matching* empregado na abordagem atual que é a modelagem por misturas de Gaussianas, GMM.

O bloco **CM** corresponde à medida de confiança no resultado *Confidence Measure*. Esse bloco é importante pois indica a confiança que se tem no resultado. Esse bloco é que permitirá realizar a importante análise de falsa aceitação e falsa rejeição.

Apresenta-se a seguir exemplo de arquivo de configuração para a modelagem por GMM na realização de um experimento.

```
train_file  = train_file1.txt
adapt_file  = adapt_file1.txt
tests_file  = tests_file1.txt
ubm_file    = ubm1.2.bin
vp_file     = voice_printsv1.2.bin
log_file    = log_configv.txt
feature_dir = ../features
feature_ext = .mfc
trace       = 2
nmix        = 256
final_niter = 10
rfactor     = 10
dfactor     = 1
vfactor     = 0.1
```

Resumidamente, a configuração acima instrui o processo a buscar a lista de arquivos de treinamento em train_file1.txt, a qual contém 32527 arquivos de áudio listados para a construção de um modelo universal, UBM; a lista de arquivos de adaptação em adapt_file1.txt, a qual contém 15411 arquivos de áudio listados, cada um com a indicação de qual faixa etária pertence para a estimação de um modelo para cada faixa etária, *age print*; e a lista de arquivos de teste em tests_file1.txt, a qual contém 5138 arquivos de áudio a ser usados no teste. As listas são mutuamente exclusivas. Um parâmetro importante é o número de misturas por modelo nmix = 256. Dessa forma, cada modelo será composto por 256 Gaussianas de dimensão 39. Como saída, o processo salva os modelos estimados de cada faixa etária no arquivo binário vp_file = voice_printsv1.2.bin e produz arquivo texto log_file = log_configv.txt que contém informações importantes do processamento para a análise dos resultados referentes aos arquivos de teste. A figura 3 apresenta uma saída típica de processamento para nível 2 de informação de saída, (trace = 2). Quando maior o nível, mais

informações de processamento são colocadas no arquivo gerado. Essas informações permitirão a obtenção de métricas para avaliar a eficácia da metodologia adotada:

- matriz de confusão
- curvas de falsa aceitação e falsa rejeição
- índices de *precision*, *recall* e *F1*

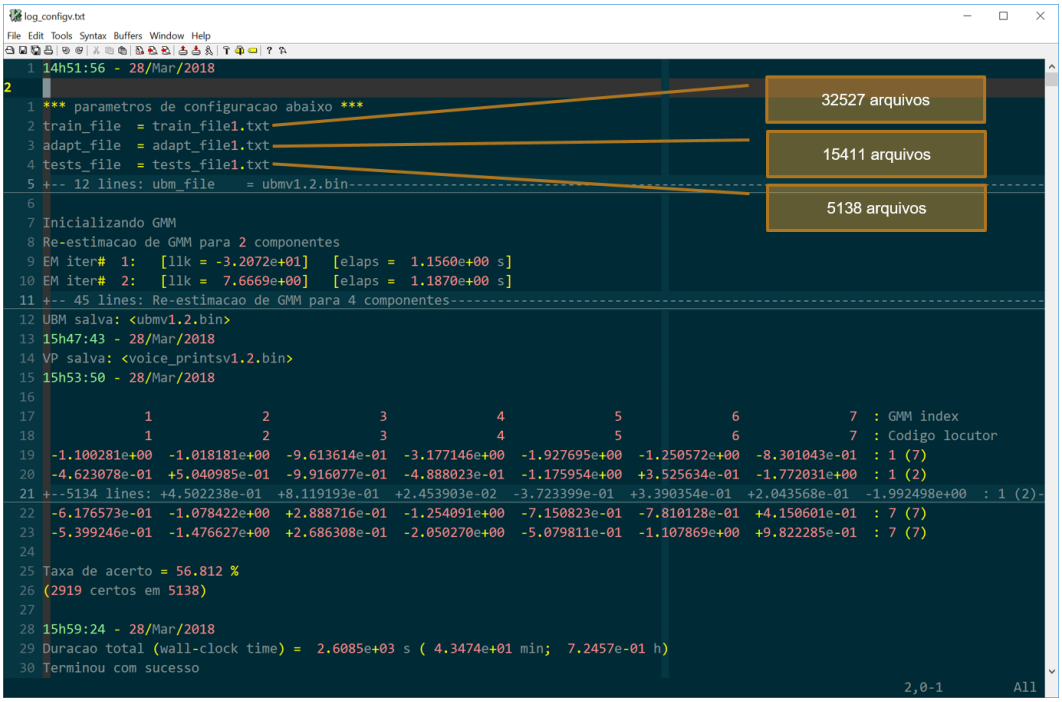

Figura 3: Arquivo texto com saída do processamento.

Na figura 3 nota-se um bloco com 7 colunas com números em notação científica. Nota-se que há uma compressão de 5134 linhas por motivo de espaço e portanto somadas com as 4 linhas visíveis totaliza o número de 5138 linhas que correspondem às saídas para cada um dos 5138 arquivos de teste. Cada coluna contém a log-verossimilhança de cada um dos modelos estimados para cada um dos arquivos de entrada, representados logo após ':' na extrema direita de cada linha. Entre parênteses aparece o modelo que produziu o melhor resultado. Assim, a título de exemplo, nota-se que as duas primeiras linhas resultaram em erro pois os sinais de entrada correspondem ao modelo 1 e o sistema acusou como sendo do modelo 7 e 2, respectivamente. Por outro lado, as duas últimas linhas acusaram corretamente a faixa etária do sinal de teste (modelo 7).

A tabela 2 apresenta a correspondência entre o número do modelo e a faixa etária. A divisão em 7 classes foi conveniente para que o número de arquivos entre as classes ficasse bem distribuído, uma vez que há menos arquivos para a classe das crianças (a qual não foi subdividida em masculino e feminino).

Portanto, dos dados de saída registrados no arquivo `log_configv.txt` constroi-se a matriz de confusão entre as classes, apresentada na tabela 3. Uma primeira constatação animadora é que os valores na diagonal principal da matriz 7×7 são dominantes,

Tabela 2: Faixas etárias em termos do número do modelo

Modelo	faixa etária
1	child
2	young female
3	young male
4	adult female
5	adult male
6	senior female
7	senior male

indicando consistência na estimação e que a faixa etária encontra-se caracterizada nos modelos, embora ainda dando margem a melhoria no processo de estimação, que será tratada na seção sobre as atividades do próximo período do projeto.

Tabela 3: Matriz de confusão. Na horizontal são os dados reais e na vertical os estimados.

		estimado						
		1	2	3	4	5	6	7
	1	325	116	38	40	18	41	22
	2	42	439	19	110	3	74	15
	3	7	5	301	10	113	11	80
real	4	27	152	27	428	20	173	10
	5	5	7	116	19	358	16	123
	6	44	143	39	157	10	459	19
	7	6	3	111	34	182	12	609

Pode-se estimar uma precisão global para as 7 classes dividindo-se o número total de arquivos de áudio corretamente estimados (soma dos valores na diagonal principal = 2919) pelo número total dos arquivos de teste (5138)

$$\text{Precisão global para as 7 classes} = \frac{2919}{5138} \times 100\% = 56.8\ \%$$

Note que o valor de 56.8 % é bem superior a uma eventual estimação aleatória entre as 7 classes: $\frac{1}{7} \times 100\% = 14.3\ \%$.

Aglutinando-se as 7 classes da matriz de confusão da tabela 3 nas 4 classes originalmente propostas produz-se a matriz de confusão apresentada na tabela 4.

Analogamente, a precisão global para as 4 classes originais pode então ser obtida

$$\text{Precisão global para as 4 classes} = \frac{3013}{5138} \times 100\% = 58.6\ \%$$

que é apenas marginalmente superior à precisão global com 7 classes. A vantagem do uso de 7 classes é que a informação sobre gênero (masculino, feminino) é também revelada.

Tabela 4: Matriz de confusão. Na horizontal são os dados reais e na vertical os estimados.

		estimado			
		child	young	adult	senior
real	child	325	154	58	63
	young	49	764	236	180
	adult	32	302	825	322
	senior	50	296	383	1099

A figura 4 apresenta as curvas de falsa aceitação e falsa rejeição em função de limiar. O limiar $\tau = 0.177$ resulta no *equal error rate*, EER, de 20.59 %.

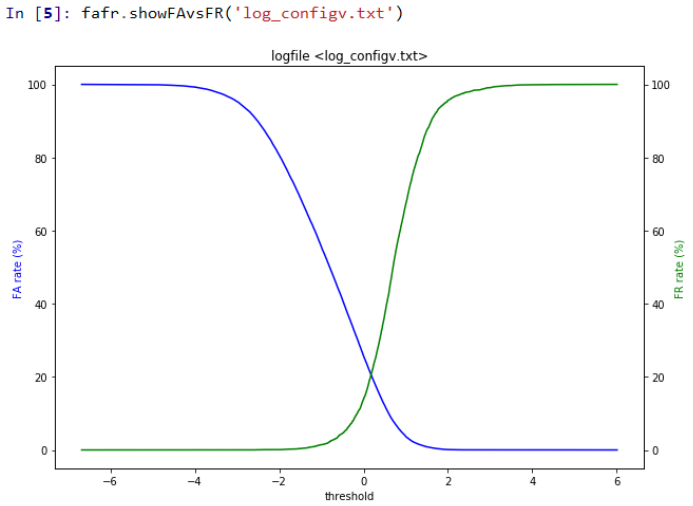

Figura 4: Curvas de falsa aceitação, FA, e falsa rejeição, FR.

Finalmente, as métricas *precision*, *recall* e *F1* são computadas na tabela 5.

Tabela 5: Valores de precision, recall e F1 para 4 classes.

	precision	recall	F1
child	0.713	0.542	0.616
young	0.504	0.622	0.557
adult	0.549	0.557	0.553
senior	0.660	0.601	0.629

Descrição e avaliação do apoio institucional recebido no período

O Centro Universitário FEI propicia as condições suficientes para a realização da pesquisa. Existe o apoio e interesse institucional com o projeto, justificado principalmente pelas horas de dedicação permitidas para a sua realização. Além disso conta-se com infraestrutura e logística excelentes, como por exemplo

- bolsistas de iniciação científica mantidos pela instituição; houve uma breve colaboração de 2 bolsistas no início do projeto
- alunos de pós-graduação; atualmente há um aluno de mestrado, com bolsa de estudos "mérito FEI" desenvolvendo tema relativo ao projeto
- apoio de pessoal técnico e secretaria
- biblioteca bem estruturada e equipada
- acesso irrestrito online a material do IEEE de qualquer computador na rede da FEI
- coordenadoria de tecnologia da informação para suporte em software

Conclui-se que o apoio institucional é altamente positivo.

Plano de atividades para o próximo período

Viu-se que a taxa de acerto de faixa etária no estágio atual para 7 classes (faixas etárias vinculadas ao gênero) está em 56.8 %. Acredita-se que esse valor poderá ser elevado com a introdução de duas principais proposições: *pitch* e *i-vectors*.

pitch

Uma palavra falada possui trechos em que há vibração das cordas vocais, normalmente as vogais, e trechos em que não há vibração das cordas vocais, normalmente as consoantes. Por exemplo, na palavra **/fala/** há vibração das cordas vocais nos trechos em que a vogal **/a/** é pronunciada e não há vibração nos instantes em que a consoante **/f/** é pronunciada. A frequência de vibração das cordas vocais, *pitch*, varia consideravelmente entre uma criança e um adulto. Adicionalmente, acredita-se que a maior informação de faixa etária encontra-se em trechos da fala em que há vibração das cordas vocais. Dessa forma, *pitch* é um bom candidato a ser considerado na modelagem deste estudo. Para isso, será necessária a implementação de um método preciso, robusto e de carga computacional não proibitiva de estimação de *pitch*. Assim, cada segmento de sinal de voz poderá ser sinalizado como tendo, ou não, vibração das cordas vocais. A ideia será considerar no processo de modelagem apenas segmentos de sinal de voz em que há vibração das cordas vocais, isto é, considerar majoritariamente trechos de vogais e desprezar trechos em que não há vibração das cordas vocais: consoantes surdas e silêncio. Adicionalmente, o vetor de *features* desses segmentos terão o valor da frequência fundamental concatenado aos outros elementos desse vetor. Apenas para ilustrar, abaixo são apresentados valores médios típicos

	pitch, Hz	
crianças	de 250	a 650
mulheres adultas	de 165	a 255
homens adultos	de 85	a 180

Vê-se também que o *pitch* pode ser um bom indicador de gênero apesar de haver sobreposição entre os limites das faixas: isto é, por exemplo, pode haver homens com *pitch* de 180 Hz e mulheres com *pitch* de 170 Hz. Apesar dessa sobreposição, mantém-se a tese de que o uso desse parâmetro no processo de modelagem trará potenciais benefícios.

O estudo da influência do *pitch* na estimação automática da faixa etária pela voz conta com o trabalho sendo desenvolvido por aluno de mestrado. O aluno recebeu bolsa institucional do Centro Universitário FEI, como mencionado na seção anterior, e vem sendo orientado pelo autor deste relatório.

i-vectors

Dos métodos clássicos de *machine learning*, i-vector é o estado da arte em sistemas de biometria para autenticação de locutor. O trabalho já realizado com GMM será útil como base para extensão à nova técnica, a qual exige mais carga computacional e memória para o treinamento dos modelos (*age prints*). A técnica busca determinar os fatores latentes que governam as correlações entre os componentes dos vetores de padrões (*feature vectors*). O método baseia-se na intuição de que o número de fatores latentes seja razoavelmnete inferior ao tamanho do vetor de padrões atualmente sendo empregado. Assim, com a adoção de uma configuração experimental adequada, a expectativa é de que se obtenham melhores resultados com o uso de menos parâmetros (contudo a um custo maior de processamento e uso de memória na fase de treinamento dos modelos). Como os resultados com essa técnica são promissores na autenticação de locutores, espera-se que o mesmo se reflita neste estudo.

Participação em evento científico

Nada a relatar nesta seção no momento.

Lista de publicações

Nada a relatar nesta seção no momento.