Relatório Científico - 2 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: 01/09/2018 a 31/08/2019

26 setembro 2019

Resumo

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

Tal conhecimento pode ser usado de forma bastante positiva ao guiar 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 possa interagir 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, foram exploradas técnicas de aprendizado de máquina (*machine learning*) aliadas a recursos usados em biometria e reconhecimento automático de fala.

Neste projeto foram implementadas técnicas de aprendizado de máquina como **Gaussian Mixture Models** (GMM) e **i-vectors**. As técnicas foram escolhidas devido ao bom desempenho quando aplicadas em biometria vocal e parametrização (*features*) utilizada em reconhecimento automático de fala. A linguagem de programação escolhida para implementação foi C, visando eficiência de processamento e portabilidade, vislumbrando, no futuro, implementações em dispositivos embarcados.

Foi desenvolvida técnica inovadora mais robusta e precisa de estimação de *pitch*, comparada a equivalentes e já consagradas em uso acadêmico [1]. A técnica emprega o algoritmo de Viterbi e será descrita sucintamente na próxima seção.

No estágio atual, este projeto atingiu taxas de acerto comparáveis a trabalhos correlatos que empregaram outras técnicas, como *deep learning* [3], e se utilizaram da mesma base de dados que este projeto.

Resultados foram avaliados por diversas métricas, como taxa de acerto, matriz de confusão, curvas de falsa aceitação *versus* falsa rejeição, *precision*, *recall* e *F1-score*. Constatou-se que um aumento da base de dados implicará melhora nos índices de acerto obtidos. Com a base atual, que consiste de frases curtas e qualidade de canal telefônico (limitadas em 8 kHz), obteve-se uma taxa de acerto promissora de 87.8 % na distinção entre crianças, homens e mulheres. Mais resultados serão apresentados adiante.

Realizações no período

No relatório referente ao período anterior houve a descrição da base de dados utilizada e relatou-se sobre a implementação e resultados da técnica GMM (*Gaussian mixture models*).

Neste período houve a implementação da técnica **i-vectors** e a concepção de método de estimação de **pitch** baseado no algoritmo de Viterbi. **Resultados** foram obtidos para diversas configurações das técnicas implementadas com e sem a utilização da informação de pitch. As seções seguintes apresentam com mais detalhes essas atividades.

i-vectors

O termo i-vector [2] é uma forma reduzida do termo "identity vector", com respeito ao problema de identificação de locutor. Esta técnica busca determinar os fatores latentes (não observáveis) que governam as correlações entre os componentes da parametrização (vetores de padrões, ou *feature vectors*) extraídos dos sinais de voz (observáveis). A técnica baseia-se na intuição de que o número de fatores latentes seja razoavelmente inferior ao tamanho do vetor de padrões empregado. Assim, um conjunto de vetor de padrões M de um locutor pode ser modelado pela soma do vetor do modelo universal, M_{UBM}, apresentado no relatório anterior, com a transformação do i-vector, w, pela matriz de variabilidade total T,

$$M = M_{UBM} + T\,w.$$

A dimensão do vetor w corresponde ao número de fatores latentes e, consequentemente, a matriz T será uma matriz retangular de baixo posto (*low rank*) com número de linhas igual ao tamanho do vetor M. Número típico de fatores latentes no problema pode variar de 10 a 400, enquanto que o tamanho do vetor M pode atingir 51200 (no caso em que o UBM possua 1024 misturas Gaussianas multivariadas de dimensão 50). A matriz T é estimada durante a fase de treinamento e os vetores w de cada classe do problema são estimados na fase de adaptação. Os i-vectors ainda poderão ter sua dimensão reduzida com a aplicação de LDA, *Linear Discriminant Analysis*.

Na primeira parte deste projeto, a implementação da técnica GMM resultou em cerca de 2800 linhas de código em C. Agora, com a introdução de i-vector, esse número foi para 9200 linhas de código em C. Esse número não foi maior devido à utilização da biblioteca *GNU Scientific Library*, necessária para as operações de cálculos de autovetores, autovalores, matrizes inversas generalizadas, etc.

A decisão pela implementação de i-vector deveu-se ao seu bom desempenho no problema de biometria vocal. Contudo, como os resultados mostrarão, seu desempenho não foi superior ao dos modelos GMM implementados na primeira parte deste projeto. Além do desempenho não ter sido superior, as necessidades de memória e tempo de processamento revelaram-se mais custosas para a técnica i-vector.

Estimação de pitch

Pitch é a frequência de vibração das cordas vocais na produção de um trecho sonoro da fala. Seu valor médio é uma característica pessoal e função da idade, sexo, estrutura do trato vocal dentre outros fatores. Abaixo são apresentadas faixas de valores usuais para crianças, mulheres e homens.

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

Nota-se que embora haja sobreposição entre as faixas de frequência, esse parâmetro pode contribuir e atuar como um bom indicador na distinção entre sexos e na distinção entre crianças de adultos.

Toda palavra falada possui pelo menos um trecho em que há vibração das cordas vocais, normalmente as vogais, e eventuais 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. Intuitivamente, devido à riqueza do sinal e maior componente energético, assume-se que trechos sonoros (em que há vibração das cordas vocais) contêm mais informação da constituição do trato vocal do que trechos surdos, nos quais não há vibração das cordas vocais. Por esse motivo, um detector preciso de pitch poderá ser empregado para diferenciar trechos sonoros de surdos. Assim, apenas trechos sonoros serão permitidos prosseguir no processamento com as vantagens de redução na carga computacional, por se evitar o processamento de trechos surdos que naturalmente contêm relativamente menos informações estruturais do locutor e com o ganho adicional da informação do pitch. Será visto na seção de resultados que os experimentos visarão comprovar a influência positiva do pitch pela comparação dos resultados do arranjo de 3 situações: 1 - o sinal inteiro será processado; 2 - apenas trechos sonoros serão processados; 3 - apenas trechos sonoros serão processados e o valor do pitch correspondente será mais uma característica (*feature*) adicionada ao vetor de padrões.

A figura 1 ilustra um trecho de sinal de voz onde o detector de pitch está em ação de forma a detectar trechos sonoros e, tão importante quanto, estimar o correspondente valor de pitch para incluí-lo como mais um componente do vetor de padrões.

A seguir, apresenta-se a proposta de estimação de pitch pelo algoritmo de Viterbi.

Estimativa de pitch pelo algoritmo de Viterbi

O algoritmo de Viterbi é empregado em inúmeras aplicações. Por exemplo, no processo de decodificação de um sistema de reconhecimento de fala baseado em *hidden Markov models* e na decodificação de códigos convolucionais em CDMA e GSM, dentre muitas outras aplicações. Vislumbrou-se a possibilidade de utilizar esse algoritmo no processo de estimação de pitch.

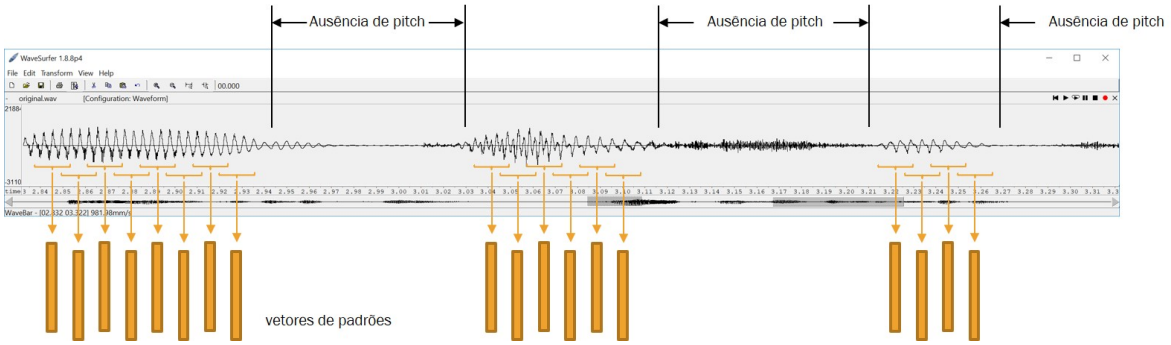

Figura 1: Vetores de padrões, representados por retângulos, correspondentes a trechos sonoros do sinal de voz.

Por falta de espaço, mais detalhes e a descrição matemática formal e precisa do método serão deixados para publicação futura com o devido reconhecimento à IBM e FAPESP. Na figura 2 temos uma disposição de pontos, os quais correspondem a possíveis valores de frequência de pitch f_i, $i = 1, 2, \ldots, N$, em função de instantes de tempo correspondente aos índices dos quadros do sinal de voz, que correspondem a trechos de cerca de 20 ms a cada 10 ms, enumerados de 1 a T. Para toda transição de quadros são computadas todas as transições possíveis de valores de frequência de pitch. Por exemplo, seja f_i^t o valor de frequência no quadro t e f_j^{t+1} o valor de frequência para pitch no instante $t + 1$. A confiabilidade do valor corrente de pitch estimado também depende do índice V, o qual é uma combinação de fatores como a energia do trecho corrente e sua razão com a amplitude do segundo pico da função de autocorrelação do trecho de voz sendo processado. Apenas os valores estimados cujos valores de V estiverem acima de um limiar serão considerados confiáveis. O algoritmo de Viterbi provê o mecanismo de se traçar e obter ao final do processo o melhor (no sentido probabilístico) de todos os caminhos possíveis de transições de valores de pitch. Dessa forma, tendo-se o melhor caminho obtém-se a melhor sequência de valores de pitch para cada quadro do sinal de voz sendo analisado. No exemplo ilustrativo apresentado, detectou-se os seguintes pares de índice de quadro e valor de pitch: $\{(3,\ f_4),\ (4,\ f_4),\ (5,\ f_3),\ (6,\ f_3),\ (11,\ f_3),\ (12,\ f_3),\ (13,\ f_4),\ (14,\ f_4),\ (15,\ f_3)\}$.

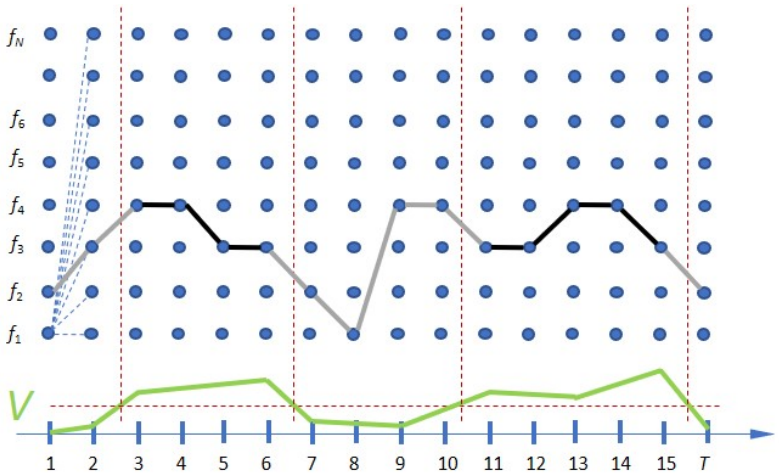

Figura 2: Ilustração do processo de estimação de pitch via algoritmo de Viterbi.

Para ajudar no traçado do melhor caminho na malha apresentada, será inserido

um fator fisiológico de enorme importância na restrição das possibilidades de transição entre valores de pitch. Apesar do pitch poder variar consideravelmente durante uma elocução, principalmente quando há algum sentimento envolvido (interjeições), perguntas, etc., as transições entre valores de pitch em instantes de tempo próximos (quadros vizinhos) estão limitadas pela própria fisiologia do trato vocal. Isto é, concretamente, é fácil intuir que dado que no quadro t a frequência de pitch seja f_i^t, então a probabilidade de transição para a frequência f_j^{t+1} no quadro $t+1$ será maior para valores próximos de f_i^t, isto é, não há transições muito abruptas de valores de pitch. Assim, pode-se modelar uma probabilidade condicional de transição entre os estados do algoritmo. Bons resultados foram obtidos ao se assumir que essa probabilidade condicional seja uma densidade de probabilidade Gaussiana de média f_i^t e desvio padrão σ_p,

$$\Pr(f_j^{t+1}|f_i^t) = \frac{1}{\sqrt{2\ \pi}\ \sigma_p} e^{-\frac{1}{2}(\frac{f_j^{t+1}-f_i^t}{\sigma_p})^2}.$$

O desvio padrão σ_p dá uma margem para quanto o valor de pitch pode variar entre quadros consecutivos. Uma faixa de valores para σ_p que produziram bons resultados vai de 20 Hz a 70 Hz.

As figuras seguintes fazem uma comparação entre o método proposto e um médoto referência usado por aplicativo conceituado na academia e por desenvolvedores em reconhecimento de fala [1]. Foram selecionados aleatoriamente 3 arquivos representantes de uma criança, mulher e homem.

A figura 3 apresenta um exemplo de voz de criança. A parte superior divide-se em três janelas: a primeira é o sinal de voz no tempo em função do índice de amostra, a segunda apresenta o melhor caminho resultante do algoritmo de Viterbi aplicado ao problema como elucidado anteriormente e a terceira janela são índices que indicam confiabilidade no valor corrente estimado para o pitch. Nestas duas janelas o eixo das abscissas representa o índice do quadro de voz, lembrando que cada quadro está espaçado do seguinte em 10 ms. A parte inferior da figura é o painel do aplicativo referência, que se divide em duas partes: a parte superior são os valores de pitch estimados e a parte inferior é o sinal de voz em função do tempo, em segundos. No método proposto, os valores de pitch válidos são os que se encontram sobre a região hachurada e concomitantemente quando a distância entre as linhas pontilhadas diminui (maior confiabilidade no resultado). Vê-se que ambos métodos produzem praticamente os mesmos valores de pitch para os mesmos instantes de tempo. Contudo, no método proposto constata-se uma melhor precisão em termos de falsa aceitação, que pode ser constatada em torno do instante de tempo 0.43 s, onde claramente não há vibração das cordas vocais mas houve detecção de pitch erradamente pelo método de referência.

A figura 4 apresenta um exemplo de voz feminina. Neste caso também se observa uma melhor precisão no método proposto. Devido às restrições fisiológicas mencionadas e inseridas no processo de decodificação do algoritmo de Viterbi, no método proposto não ocorreu a abrupta transição de valor de pitch em torno de 1.5 s. Vê-se que no método referência o valor de pitch saltou de um quadro de voz para o seguinte de cerca de 100 Hz a 180 Hz. Essa grande variação instantânea, provavelmente errônea, devido à fisiologia e inércia inerente dos mecanismos de produção de voz pelo trato vocal, não aconteceu nos resultados do método proposto.

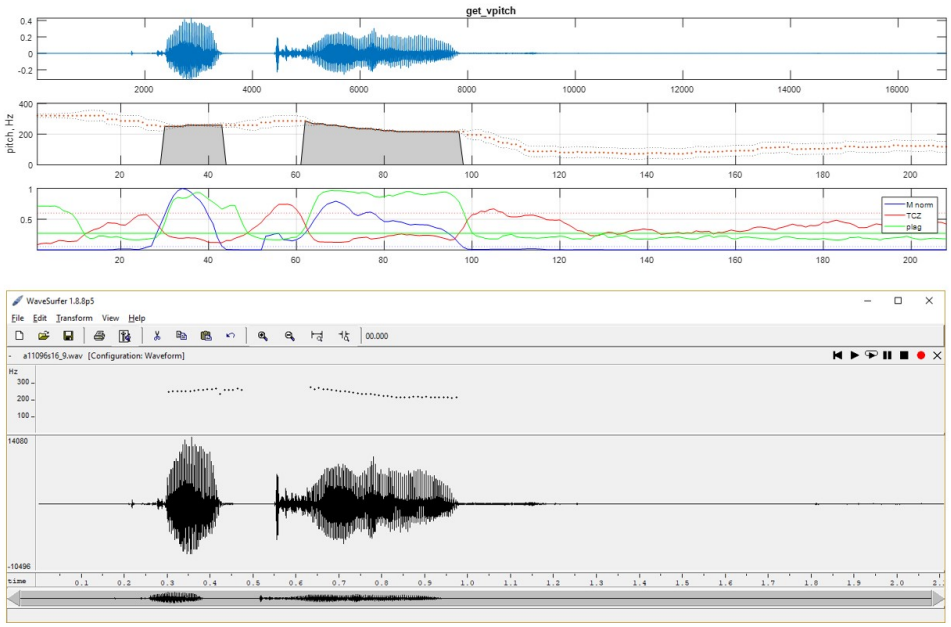

Figura 3: Voz de criança. Pitch acima dos 200 Hz.

Por fim, a figura 5 ilustra resultados para voz masculina. Neste caso, o método proposto revelou-se mais robusto ao não acusar existência de pitch logo no início do sinal, onde claramente é ruído de fundo, e também ao final, em torno de 1.8 s e 1.9 s.

Resultados

Os resultados da técnicas GMM e i-vector serão apresentados em 3 experimentos distintos:

exp1: todo o sinal de voz de cada arquivo de áudio será processado e usado nos processos de treino, adaptação e teste

exp2: apenas os quadros em que ocorre vibração das cordas vocais de cada arquivo de áudio serão processados.

exp3: apenas os quadros em que ocorre vibração das cordas vocais de cada arquivo de áudio serão processados e, adicionalmente, o valor estimado de pitch será adicionado ao vetor de padrões do quadro correspondente.

Ressalta-se que em **exp2** e **exp3** ocorrerá significativa redução no número de quadros (dados) para processamento, ocasionando maior rapidez de resposta e, possivelmente, ganho de precisão nos resultados pois serão tratados apenas quadros de voz com maior informação vocal dos locutores, e desprezados quadros com trechos surdos e ruído de fundo (ausência de voz).

A base de dados, a parametrização aplicada nos arquivos e detalhes da metodologia experimental foram descritos no relatório científico anterior. A tabela 1 recorda as classes (faixa etária mais gênero) modeladas no problema.

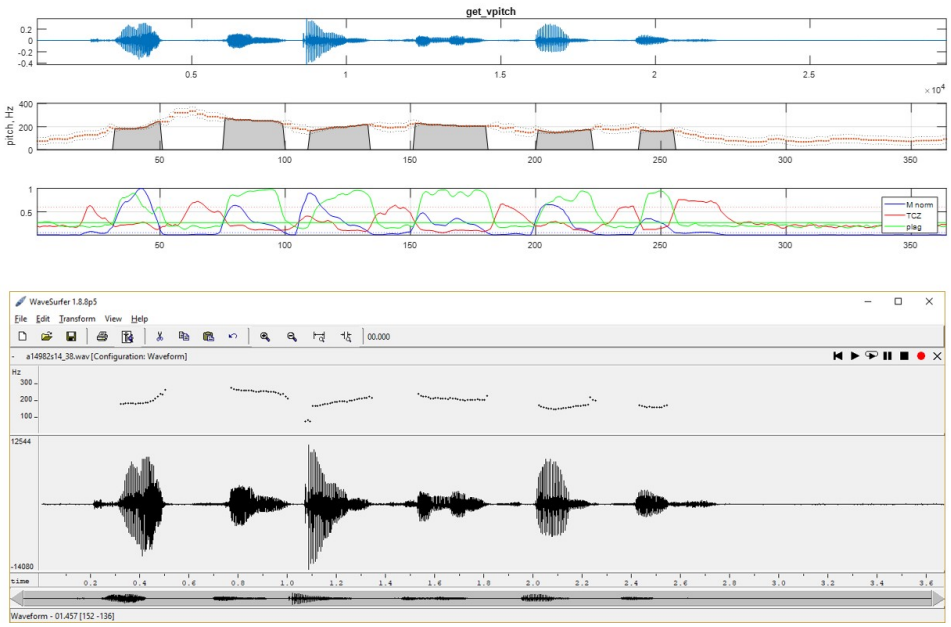

Figura 4: Voz de mulher. Pitch médio um pouco inferior aos 200 Hz.

Tabela 1: 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

Os resultados globais obtidos são apresentados na tabela 2. Parâmetros de configuração fundamentais usados nestes experimentos são

- número de misturas: 512
- variáveis latentes: 200
- dimensão LDA: 6

A tabela possui duas partes,

- **teste == treino** em que locutores usados para teste também foram usados no treinamento, com uso de 24500 arquivos de voz
- **teste != treino** em que locutores usados para teste não foram usados no treinamento, com uso de 20492 arquivos de voz

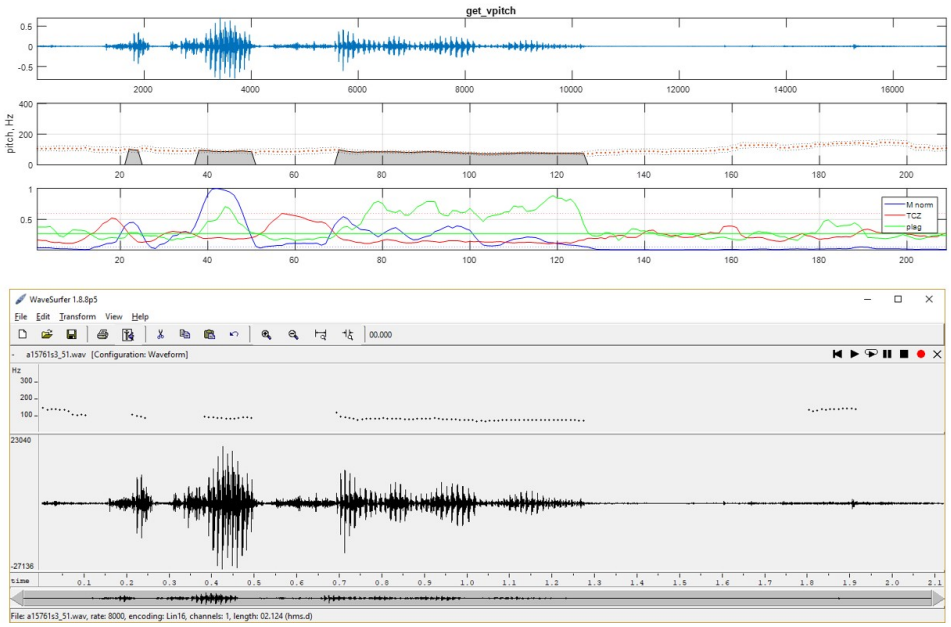

Figura 5: Voz de homem. Pitch em torno dos 100 Hz.

Tabela 2: Resultados globais em porcentagem de acerto.

#arq	24500		20492	
	teste == treino		teste != treino	
	GMM	i-vector	GMM	i-vector
exp1	73.0	57.9	41.4	46.4
exp2	73.2	51.6	44.5	46.4
exp3	74.1	53.4	48.0	47.8

Mesmo com o uso de aproximadamente duas dezenas de milhares de arquivos de cerca de centenas de locutores, a grande diferença de performance entre os casos **teste==treino** *versus* **teste!=treino** indica que um aumento na base de dados poderá trazer benefícios na modelagem com melhora na taxa de acerto.

Na porção dos resultados em que **teste!=treino**, que é a que possui maior importância para a análise, vê-se que como esperado, há uma melhora nos resultados de **exp3** sobre os demais. Isto é, a informação de pitch trouxe melhora nos resultados, embora não tão impactante quanto se esperava. Adicionalmente, a técnica i-vector não demonstrou ser superior à GMM.

Tomando-se o melhor dos resultados para **teste!=treino**, que é **exp3** GMM, a matriz de confusão resultante é dada na tabela 3, que apresenta os resultados em função do índice do modelo, como definido na tabela 1.

As curvas de falsa aceitação e falsa rejeição foram obtidas e são apresentadas na figura 6.

A matriz de confusão, tabela 3, pode ainda ser manipulada para apresentar os

Tabela 3: Matriz de confusão para GMM, **exp3**, **teste != treino**. Na horizontal são os dados reais e na vertical os estimados.

		estimado						
		1	2	3	4	5	6	7
real	1	1354	467	155	192	29	122	69
	2	510	1398	24	544	10	303	14
	3	4	12	949	33	583	65	454
	4	155	944	48	1263	19	907	13
	5	2	4	681	21	901	62	895
	6	221	559	32	947	17	1663	37
	7	7	4	400	42	920	120	2317

resultados em termos de 4 grandes classes, conforme a tabela 4. A taxa de acerto correspondente pode ser computada pela soma dos valores na diagonal principal sobre o número total de arquivos usados nesse experimento:

$$\text{Taxa de acerto 4 classes} = \frac{10078}{20492} \times 100\ \% = 49.2\ \%$$

Tabela 4: Matriz de confusão para 4 classes.

		estimado			
		child	young	adult	senior
real	child	1354	622	221	191
	young	514	2383	1170	836
	adult	157	1677	2204	1877
	senior	228	995	1926	4137

As métricas *precision*, *recall* e *F1-score* são apresentadas na tabela 5 para as 4 classes.

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

	precision	recall	F1-score
child	0.601	0.567	0.584
young	0.420	0.486	0.451
adult	0.399	0.373	0.385
senior	0.588	0.568	0.578

A matriz de confusão, tabela 3, pode ser manipulada mais uma vez para apresentar os resultados em termos de 3 grandes classes, conforme a tabela 6. A taxa de acerto correspondente pode ser computada pela soma dos valores na diagonal principal sobre

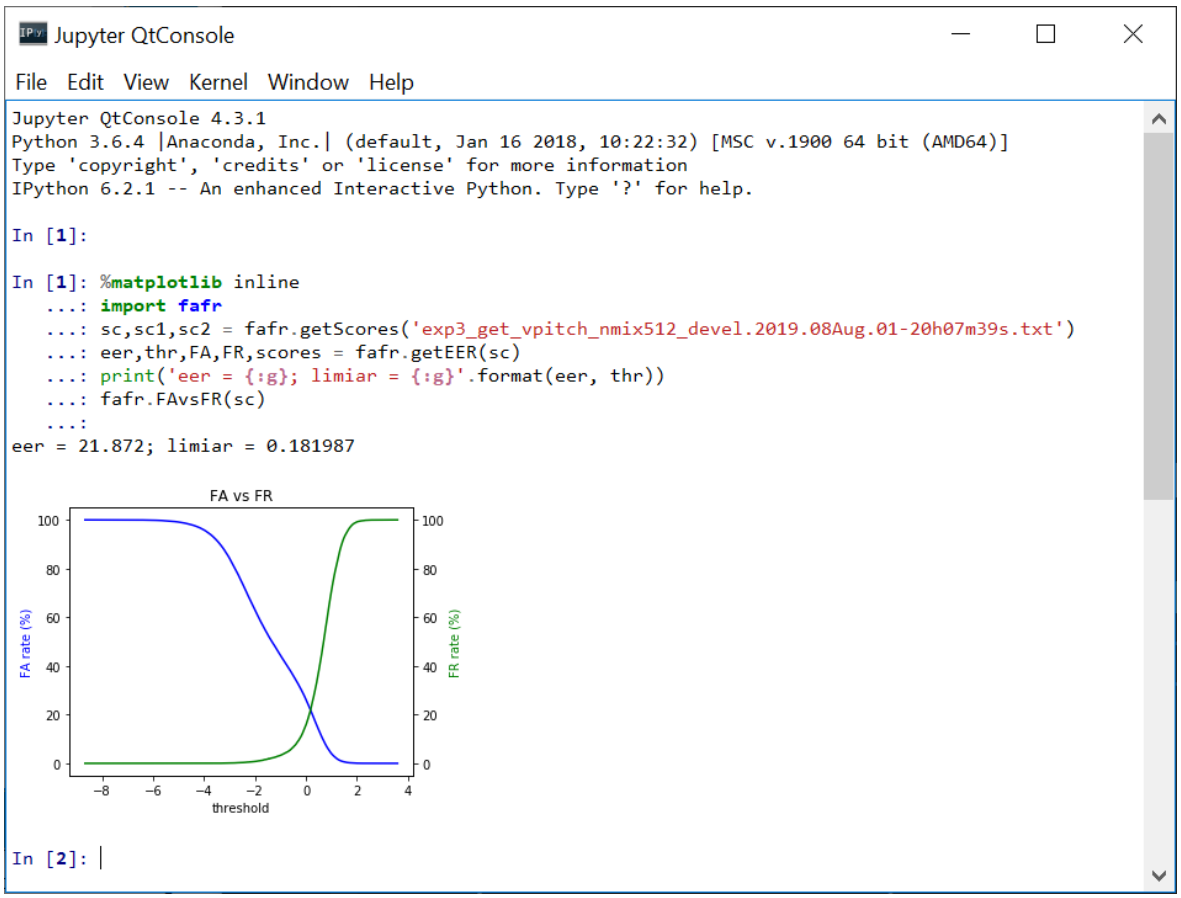

Figura 6: Curvas de falsa aceitação e falsa rejeição. No limiar de 0.182 ocorre o *equal error rate* de eer = 21.9 %.

o número total de arquivos usados nesse experimento:

$$\text{Taxa de acerto 3 classes} = \frac{17982}{20492} \times 100\ \% = 87.8\ \%$$

Tabela 6: Matriz de confusão para 3 classes.

		estimado		
		child	female	male
real	child	1354	78	253
	female	886	8528	214
	male	13	363	8100

Acredita-se que a taxa de acerto de 87.8 % na distinção das classes crianças, mulheres e homens pode ser considerada aceitável e com potencial de trazer os benefícios esperados em aplicações que se utilizarem dessa classificação. Quanto à estimação automática precisa de faixas etárias em menor resolução, ainda serão necessários mais estudo e desenvolvimento.

Referências

[1] Speech, Music and Hearing Division, KTH Royal Institute of Technology, Stockholm, Sweden. `https://www.speech.kth.se/wavesurfer/`. Acesso: 24/09/2019.

[2] Najim Dehak et al. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798, 5 2011.

[3] Zakariya Qawaqneh et al. Age and gender classification from speech and face images by jointly fine-tuned deep neural networks. *Elsevier Expert Systems with Applications*, 85:76–86, 11 2017.

Descrição e avaliação do apoio institucional recebido no período

O Centro Universitário FEI propiciou as condições suficientes para a realização da pesquisa. Houve o apoio e interesse institucional com o projeto, justificado principalmente pelas horas de dedicação permitidas para a sua realização. Além disso, contou-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" na fase da escrita de sua dissertação
- 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 e hardware

Participação em evento científico

- Participação no Workshop organizado pela IBM Research Brasil, em 11 de junho de 2019, no qual foram apresentados o desenvolvimento e os resultados deste projeto, junto com os dos outros projetos participantes do Acordo de Cooperação em Computação Cognitiva - FAPESP e IBM
- Participação de aluno de iniciação científica, Guilherme Grandesi, no IX SICFEI, a ser realizado em 17 de outubro de 2019. Aproveito a chance para mencionar que o trabalho relaciona-se com este projeto no sentido de que o aluno desenvolve uma estrutura de coleta de voz via internet que permitirá de forma relativamente barata e de fácil acesso pelos usuários na formação e ampliação de base de dados de voz.

Lista de publicações

- Aluno de mestrado, Caio Ribeiro Nakaue, encontra-se na fase de escrita de sua dissertação sobre o tema da influência do pitch na estimação automática de faixa etária pela voz.
- Planeja-se a submissão de artigo descrevendo o método de estimação de pitch baseado no algoritmo de Viterbi e os resultados deste projeto