Reconhecimento do falante

Trabalho realizado por:

Luís Filipe Moreira

Índice

1. Evolução dos sistemas de Reconhecimento do Falante

1.1 História

A investigação no campo do Reconhecimento do Falante tem vindo a ser feita há 50 anos. Contudo, até à data ainda não foi realizado qualquer sistema em computador eficaz. Além disso, os problemas relacionados com a adequação à verificação do falante são tão severos como os relacionados com a identificação do falante.

Naturalmente tem sido feito algum progresso, sendo o mais notável quando os Criminalistas peritos em Fonética se juntaram aos Engenheiros Electrotécnicos e de Som. Actualmente parece óbvio que a junção de ambas as equipas será necessária se se pretender o desenvolvimento de sistemas com sucesso – sistemas que consigam simultaneamente lidar com os desafios resultantes da unicidade e das variações no comportamento humano, a motricidade da fala, as situações forenses e as características dos computadores.

1.2 Introdução

O reconhecimento do falante apresenta problemas que se provaram ser extremamente difíceis de resolver. Este desafio foi exacerbado pelo facto desta área ser composta por dois sub-domínios que, embora similares no seu essencial, exibem na realidade um grande número de diferenças. Infelizmente estas diferenças raramente são focadas; desse modo, muita da investigação feita não foi adaptada às características únicas de cada uma. As duas áreas são a **Identificação do Falante** e a **Verificação do Falante**.

Na essência das duas áreas, está, naturalmente, a necessidade fundamental de reconhecer um falante a partir de sinais acústicos – e para fazê-lo partindo da análise desses sinais. Por outro lado, a principal diferença entre as duas tarefas derivam do facto de ambas serem orientadas a um objectivo (*goal-oriented*).

É sabido que a verificação está relacionada com o processo de reconhecer um falante que ser reconhecido e, desse modo, o processo é conduzido por um falante cooperativo, normalmente num meio altamente estruturado para o efeito. Deste modo, a quantidade e o tipo de material de fala, as características acústicas e do canal, bem como o sistema de processamento podem ser sofisticados e cuidadosamente controlados. Por outro lado, quando a tarefa é identificar um falante, este é, geralmente, não cooperativo, a captura/processamento do sinal é feita em condições muito desfavoráveis. Neste caso, o ruído, gravações pobres, disfarce e muitas outras distorções do canal e do falante podem estar – e regra geral estão – presentes.

Infelizmente, muita do investigação feita não tem em conta estas diferenças gritantes. Geralmente vê-se pesquisa onde são empregues técnicas refinadas de processamento de sinal aplicadas a material que foi degradado por diversos tipos de distorções; em que factores incontroláveis são inseridos no projecto. Pelo contrário, as abordagens de investigação na tarefa da identificação são por vezes rudimentares, subjectivas, ou, no mínimo, não quantitativas. Em muitos casos, pouca investigação fundamental é feita.

Finalmente foi reconhecida a necessidade de separar de modo claro as aplicações de âmbito comercial das aplicações de âmbito criminal.

Basicamente os forenses concentram-se na área da identificação enquanto que os engenheiros centram-se sobretudo na área da verificação do falante. Tal sucede porque os engenheiros (electrotécnicos, de som, de computadores) tendem ser ligados a sistemas; assim sendo, eles são atraídos a desafios de assuntos processuais e, desse

modo, para a verificação. De facto, para a verificação, o falante é controlado à medida que as características acústicas (que são o produto da motricidade da fala) são sujeitas aos mais variados algoritmos que se pensem ser efectivos.

Os fonéticos, por seu lado, trabalham com sistemas de comportamento respondendo desse modo à dinâmica do falante e do próprio acto de falar. Eles consideram que o comportamento humano é a chave e tendem a gravitar para o processo de identificação do falante com todas as suas variações. A sua atenção está mais nas características que permitem aos falantes ser reconhecidos por quaisquer meios e menos em sistemas que possam ser criados para fornecer soluções processuais.

Isto confunde ainda mais um problema que já é de si complexo. Os que trabalham na área da verificação irão usar, por vezes, material da área forense e ao fazê-lo introduzem variáveis incontroláveis no processo de investigação. Claro que por vezes a verificação do falante vai para além do sector comercial (aviões, campos de guerra, astronautas, aquanautas, etc.). Contudo mesmo aqui a tarefa mantém-se como sendo cooperativa. Por outro lado, os que trabalham na área da identificação são conhecidos por estudar a mímica ou a simples discriminação e esses são assuntos claramente relacionados com a verificação do falante. Resumindo, ambos os processos e os profissionais envolvidos tendem a se sobreporem.

Uma das poucas relações entre as duas áreas é que os desafios presentes na tarefa de identificação são muito mais severos que os associados à verificação do falante. De facto, se se desenvolvesse um sistema totalmente válido e eficiente para a identificação do falante, então o problema da verificação estaria essencialmente resolvido. Considere-se o caso em que se tem de avaliar duas amostras de voz. No primeiro caso a pessoa fala num ambiente barulhento, ao telefone e disfarça a sua voz, enquanto que a sua amostra exemplar é feita quando essa pessoa fala para um gravador de baixa fidelidade numa sala cheia de reverberações e está assustada e deprimida. Se o sistema de identificação utilizado for suficientemente robusto para fornecer informação válida acerca do "acasalamento" (*match*) ou não destas duas vozes, porque é que ele não seria igualmente eficaz em "acasalar" a fala de um indivíduo com um pouco de tosse que apanhou devido a uma constipação com a amostra dessa mesma pessoa previamente guardada? Não seria também eficaz em expor um impostor que tenta identificar-se como o piloto de um avião mesmo que o avião esteja em pleno voo? De facto, é razoável assumir que se o problema da identificação do falante pode ser resolvido, disso resultará uma boa verificação do falante.

1.3 Soluções possíveis

Quando se tem em atenção os problemas e as confusões citados, é fácil perceber porque é que as soluções para o desafio que é o reconhecimento do falante têm sido lentas no seu desenvolvimento. O que é que pode ser feito para acelerar o processo especialmente agora que as perspectivas de "dificuldade" foram estabelecidas?

Pelo que foi dito, talvez fosse melhor deixar os Fonéticos tomar conta da área em exclusivo. Ao fim de contas são eles quem percebe a dinâmica do comportamento humano e são versados na natureza e nas variações da motricidade da fala humana, percepção e por aí fora. De facto, alguns destes especialistas são cientistas rigorosos; alguns até percebem a complexidade do processo forense. A maioria usa procedimentos de instrumentação e processamento de sinal; alguns fazem-no exaustivamente. Mas, serão as suas aptidões individuais e colectivas suficientemente robustas para encarar todos os desafios encontrados e poderão suceder em tempo razoável? Provavelmente, não.

Bom, então, talvez deva ser dada aos Engenheiros a responsabilidade de desenvolver métodos efectivos de reconhecimento do falante. Ao fim de contas, como um grupo eles têm quer o treino quer a experiência para conceber novos sistemas; estão familiarizados com o processamento de sinal (incluindo operações envolvendo distorções e ruído) e têm um bom conhecimento de computadores e de aplicações para computadores. Alguns lidaram directamente com dinâmicas humanas e/ou tiverem experiências no campo forense. Novamente, contudo, não irão os profissionais desta área exibir tendências e deficiências que irão bloquear soluções com sucesso nesta área?

O teste destes dois postulados reside na resposta à seguinte questão: será que meio século de esforço de ambos os grupos produziu uma abordagem efectiva ao reconhecimento do falante? Infelizmente, a resposta tem de ser negativa.

1.4 Recomendações

Uma solução razoável do problema parece possível. Em primeiro lugar, as abordagens dinâmicas actuais devem ser continuadas. Especialmente importante são aquelas que servem para clarificar os assuntos e as fronteiras envolvidos e que permitem uma melhor organização e estruturação dos domínios constituintes do reconhecimento. Em segundo lugar, é necessário aceitar que a chave para o reconhecimento do falante reside na identificação com sucesso e que o ênfase na investigação deveria ir nessa direcção. Claro que se deve continuar a investigação em programas de verificação. Contudo, tendo a orientação global virado da verificação à identificação, tal deverá conduzir a uma solução global mais eficaz.

2. Processos de Reconhecimento do Falante com senha

2.1 Definição do problema do reconhecimento utilizando senha

Admita-se que se dispõe de *A* amostras de som pronunciadas por *F* falantes, e nas quais se conhece o início e o fim. As amostras referem-se a *C* classes distintas de som, por exemplo palavras, com durações não necessariamente iguais. De entre as *A* amostras, A_i refere-se à classe *i*. Naturalmente

$$\sum_{i=1}^{C} A_i = A$$

Coloca-se a questão de, recolhido um novo som pronunciado por um dos *F* falantes iniciais ou não, mas com a garantia de pertença a uma das *C* classes, identificar a qual das classes pertence.

Suponha-se que os sons são amostrados a uma frequência *f* e divididos em sectores n_i (i=1,2,…) chamados *frames*, cada um dos quais contendo *N* pontos, de tal forma que o sinal possa ser considerado praticamente estacionário em cada sector. Vai ocorrer uma sobreposição parcial dos sectores. Seja *s* o intervalo (*frame shift* ou *overlap*) entre os inícios de cada par de sectores consecutivos.

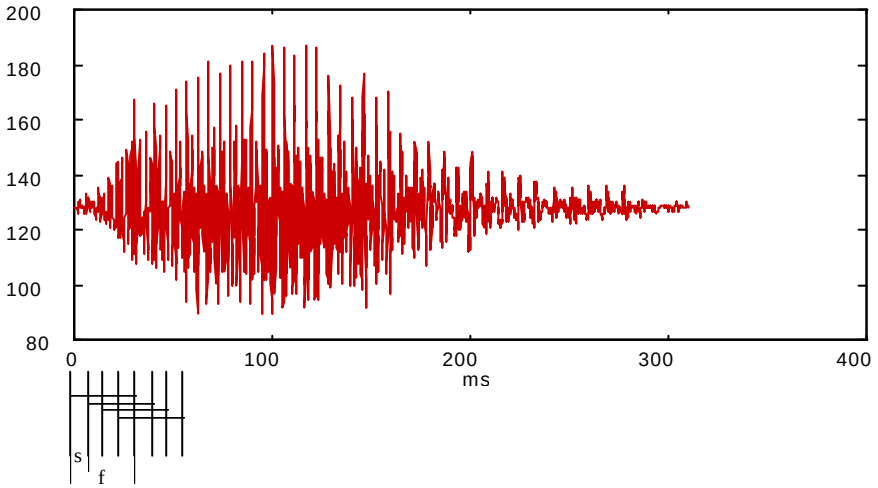


Fig. 1

Considere-se uma das *A* amostras disponíveis. Do som de cada sector extrai-se um vector de características (*features*), tais como LPC, energia, coeficientes cepstrais, delta-cepstrais, ou outras quaisquer. Assim ao sector n_i corresponde o vector de características **r**(i).

Existem, essencialmente, três processos (com alguns híbridos e variantes), razoavelmente eficazes, de resolver o problema, processos esses baseados em:

- Alinhamento temporal não linear (*Dynamic Time Warping* - DTW)
- Redes neuronais (*Artificial Neural Networks* - ANN)
- Modelos (ou cadeias) escondidos de Markov (*Hidden Markov Models* - HMM).

2.2 DTW

Imagine-se que é apresentado um som para ser identificado e que há a garantia de que esse som pertence a uma das *C* classes consideradas. Esse som será, naturalmente, também ele dividido em sectores e será extraído o vector de características de cada sector: por exemplo do sector *i*, extrair-se-á o vector **t(i)**. A identificação será feita comparando os vectores de características dos vários sectores das amostras.

Tem de se considerar as seguintes questões:

- É preciso definir um critério de comparação, isto é, uma medida da distância entre vectores de características
- Os sons, mesmo os da mesma classe, não têm todos a mesma duração, pelo que é indispensável saber que sectores da amostra padrão e que sectores da amostra a classificar, devem ser comparados. Esta tarefa equivale a um alinhamento dos dois sons
- Tem-se várias amostras de cada classe e torna-se necessário aproveitar da melhor forma essa informação disponível, gerando a partir dela, de alguma maneira, uma ou mais amostras padrão. Este processo corresponde a um treino do sistema.

A distância entre os vectores **t**(i) e **r**(i) mais utilizada em DTW é a **euclidiana**, definida para o caso do processamento da voz do seguinte modo:

$$\mathrm{d}\left(\mathbf{t}(i),\mathbf{r}(i)\right)=\sqrt{\left(\mathbf{t}(i),\mathbf{r}(i)\right)^{T}\mathbf{C}_{i}^{-1}\left(\mathbf{t}(i),\mathbf{r}(i)\right)}$$

em que **C** é a matriz de covariância do vector de características.

A cada uma das classes está associada uma amostra padrão do som que será reconhecido. A definição de amostras padrão é essencial para conseguir a economia computacional suficiente que permita um funcionamento real on-line. A cada amostra padrão está associada uma classe, por exemplo uma palavra. Pretende-se comparar o som a classificar com todas as amostras e decidir a que classe pertence.

Se o novo som fosse pronunciado da mesma forma que a amostra correspondente à classe a que na verdade pertence, em particular se tivesse a mesma duração temporal, bastaria calcular as distâncias entre vectores de sucessivos sectores do som das amostras e, por exemplo, somá-las. A soma com menor valor corresponderia à amostra da classe.

Infelizmente, o problema não é assim tão simples. Contudo, deve reter-se a ideia de calcular distâncias entre vectores de sectores e somar.

Assim sendo, torna-se necessário fazer um alinhamento de sectores do som com sectores da amostra.

Uma primeira solução poderia ser fazer um alinhamento linear. Porém, não só a palavra pode ser dita mais depressa ou mais devagar, como podem variar de forma desigual os tempos gastos, por exemplo, na pronúncia das suas diferentes sílabas. É preciso, portanto, um alinhamento não linear.

Suponha-se, então, que a amostra se divide em *I* sectores e que o som a classificar se divide em *J* sectores. Tem-se, assim, *I* vectores de características extraídos do som e *J* da amostra.

$$\mathbf{t}(1), \mathbf{t}(2), \ldots, \mathbf{t}(I)$$
$$\mathbf{r}(1), \mathbf{r}(2), \ldots, \mathbf{r}(J)$$

A classe a que se concluirá pertencer o som será a correspondente à amostra para a qual for menor uma quantidade *D*, que será a soma de *K* distâncias entre vectores de características extraídos de sectores do som e da amostra.

$$D = \sum_{k=1}^{K} d_k$$

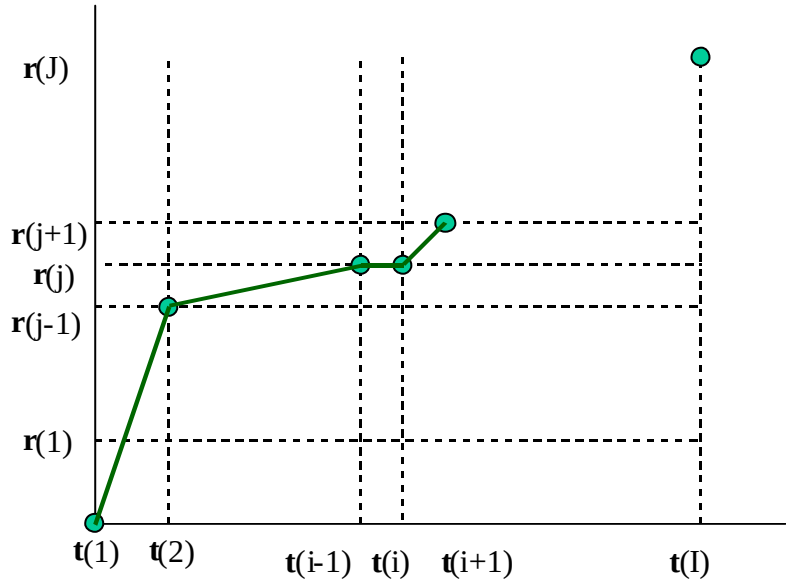
A questão é determinar que sectores, e portanto que vectores do som e da amostra, vão ser comparados. Note-se que *K*, *I* e *J* não são necessariamente iguais.

Como se garante que, quer a amostra, quer o novo som, estão perfeitamente limitados, pode-se afirmar que os primeiro e último sectores identificam realmente, nos dois casos, o principio e o fim do que se quer classificar, por exemplo palavras. Logo sabe-se que vectores comparar no início e no fim. Isto é

$$d_1 = d\left(\mathbf{t}(1), \mathbf{r}(1)\right)$$

$$d_K = d\left(\mathbf{t}(I), \mathbf{r}(J)\right)$$

Mas como obter d_2? Deve-se calcular a distância entre que vector da amostra e entre que vector de som?



A cada ponto da grelha está associado um custo: a distância entre os vectores da abcissa e da ordenada. O objectivo é o de encontrar o caminho de menor custo entre o ponto (1,1) e o ponto (I,J). Dada a dependência temporal nem todos os caminhos são possíveis à partida: por exemplo, o caminho nunca pode voltar para trás. Por outras palavras, na passagem de um nó para o seguinte, i e j não podem diminuir. Há assim, para cada nó, um conjunto de nós anteriores legais e um conjunto de nós posteriores legais.

Um resultado essencial para a eficiência do processo de busca do caminho de custo mínimo foi obtido por Bellman e pode resumir-se facilmente. Se se pretender determinar o menor custo entre (1,1) e (I,J) basta determinar o mínimo dos custos dos caminhos de menor custo para os nodos antecessores legais e somar-lhe a distância entre **t**(i) e **r**(j). O custo total mínimo é então D = g(I,J).

2.3 Redes Neuronais

As redes neuronais (NN) são um método relativamente generalizado utilizado aquando do reconhecimento de padrões dificilmente modeláveis, isto é, quando não há um modelo matemático preciso capaz de resolver o problema. Tal faz com que sejam um instrumento utilizado no Reconhecimento do Falante.

Torne-se a olhar para o nosso problema de reconhecimento como o enunciado no princípio.

Suponha-se, para já, que todos os sons e amostras têm o mesmo comprimento ou o mesmo número de sectores.

Associe-se, a cada amostra ou cada som, um vector formado pelos vectores de características colocados em série. Esses vectores serão os vectores de entrada na rede a cada um dos quais corresponde um vector de saída identificador da classe a que pertence a amostra. A rede é treinada com este conjunto e espera-se que, apresentado um novo som à entrada, ela generalize, produzindo à saída um vector próximo do vector de código da respectiva classe.

Subsiste um problema: é que os sons não têm todos a mesma duração. Existem arquitecturas de redes que ajudam a resolver o problema, nomeadamente redes recorrentes como a rede de atraso temporal ('time delay neural network' - TDNN).

Irá procurar-se, em vez de utilizar essas redes, alinhar os sons. Isso poderia ser feito por DTW. Note-se que se devem reduzir todos os sons a um mesmo número de sectores, porque todos os vectores de entrada terão de ter o mesmo comprimento.

Usar-se-á um método diferente do DTW, conhecido como segmentação do traço.

A cada sector corresponde, como se sabe, um vector N-dimensional que identifica um ponto num espaço dessa dimensão. Um som pode ser definido como uma trajectória nesse espaço, com tantos pontos quanto os sectores. Pode-se utilizar a distância euclidiana para medir o comprimento da trajectória entre cada dois pontos. O comprimento total da trajectória será a soma dessas distâncias; quanto maior for o número de pontos, mais fina será a avaliação. Note-se que uma distância pequena entre dois pontos significa que eles são, para o que interessa, muito próximos.

Suponha-se então que se quer reduzir todos os sons a g sectores. Então vão-se reter sectores igualmente espaçados ao longo da trajectória. Divide-se o comprimento total da trajectória por g, o que dá a distância a que devem estar cada dois pontos d. Começa-se pelo primeiro sector e vai-se avançando e somando distâncias ao longo do som. Quando a distância acumulada ultrapassar d, o sector correspondente, o anterior ou o logo a seguir, será o segundo sector a ser retido, eliminando-se todos os outros entre o primeiro e este. O processo decorre até ser alcançado o último sector que será sempre retido.

2.4 Modelos Escondidos de Markov

Suponha-se que as variáveis aleatórias, os índices de vectores, podem tomar valores de entre o mesmo conjunto de D elementos. A gama de vectores é a mesma em todos os instantes, isto é, a variável aleatória correspondente ao instante (sector) t=1, designada por $x(1)$, pode tomar valores x_1, x_2, …, x_D. Qualquer outra variável, no

instante genérico t, *x(t),* pode tomar os mesmos valores. Isso é o que acontece de facto com o som. Um vector de características pode ocorrer em qualquer instante e não há instantes em que um determinado vector não possa ocorrer.

No entanto, admite-se que a probabilidade de *x(t)* assumir um qualquer dos seus valores só depende do valor assumido pela variável *x(t-1)*. Um processo estocástico deste tipo, diz-se um processo de Markov de 1ª ordem. Isto já não ocorre necessariamente no som, em que há sequências de um ou mais sectores que são muito frequentes, o que significa que essa probabilidade não depende apenas do instante imediatamente anterior. Simplesmente, o custo computacional de levar isso em consideração é demasiado pesado.

Como o conjunto de valores que as variáveis aleatórias podem tomar é discreto e finito, o processo diz-se cadeia de Markov.

Suponha-se, também, que as probabilidades referidas não dependem do tempo, isto é, a probabilidade de a variável *x(t)* assumir por exemplo o valor x_j depois de a variável *x(t-1)* ter assumido o valor x_i é a mesma de a variável *x(t-k)* assumir o valor x_j depois de a variável *x(t-k-1)* ter assumido o valor x_i. O processo de Markov diz-se então homogéneo.

Um processo estocástico como este pode ser representado por uma máquina de estados, mais concretamente um autómato finito.

Estes modelos são muito simples, mas em algumas situações, particularmente no estudo do som, adaptam-se mal aos fenómenos em estudo, essencialmente porque não admitem nenhuma variação das probabilidades com o tempo.

Determinadas sequências de sectores são mais frequentes quando se pronuncia por exemplo a vogal a do que quando se pronuncia a vogal e. Mas a distribuição das vogais num discurso pode ser considerada aleatória.

Em resumo, se se suspeita que num determinado fenómeno há uma estrutura estatística com esta complexidade de dois processos estocásticos correlacionados, é preciso desenvolver modelos mais complexos. De alguma forma é o equivalente a, dados pares de pontos, precisar-se de os ajustar por uma parábola, por exemplo, porque a recta não os representa bem, isto é, é o equivalente a aumentar o número de parâmetros de que se dispõe para afinar o ajuste a uma colecção de dados complexa.

No modelo inicial tinham-se como parâmetros as probabilidades iniciais e as probabilidades de transição de estado. Agora tem-se as iniciais, as de transição de estado e as de emissão de saídas. Tem-se muitos mais parâmetros para ajustar, logo é mais fácil modelizar comportamentos estocásticos mais complexos.

Os estados não estão assim observáveis. Por isso se chama a estes modelos, modelos escondidos de Markov (*Hidden Markov Models* - HMM).

Um HMM deve, assim, ser caracterizado por:

- Um número de estados *N*. Esses estados podem ter à partida algum significado físico ou podem ser apenas abstracções, as necessárias para ajustar estatisticamente os dados disponíveis.
- Um número *M* de valores que pode assumir a variável aleatória observável. É o número de saídas ou símbolos observáveis em cada estado. Designaremos esse conjunto por $V=\{v_1, v_2, \ldots, v_M\}$.
- Uma matriz de probabilidade de transição de estados **A**.
- Um vector de probabilidades iniciais de estados p .
- Um vector de probabilidades de saídas ou símbolos ou observações por estado $\mathbf{b}_j$, formando uma matriz **B**.
-

Um modelo pode ser abreviadamente representado por

$$M\ (N, M, \mathbf{A}, \mathbf{B}, p)$$

e o conjunto de parâmetros do modelo por

$$\lambda(\mathbf{A}, \mathbf{B}, p)$$

Agora, os problemas de estimar os parâmetros do modelo (treino) e determinar a probabilidade de uma sequência de observações dado o modelo, isto é, determinar a sua verosimilhança, são muito mais complicados. Considere-se o problema do treino: dispõe-se de várias sequências de observações. Como se viu, se se estivesse a trabalhar com estados observáveis, estimavam-se directamente a partir dessas sequências todos os parâmetros, probabilidades iniciais e probabilidades de transição de estado. Mas aqui tem que se estimar, com as mesmas sequências, probabilidades iniciais, de transição de estado e de produção de saídas em cada estado. Uma sequência de observações pode ser gerada de diferentes maneiras, dependendo dos estados que forem sendo percorridos. A determinação da sua verosimilhança tem de levar em conta todas essas maneiras.

Os grandes problemas do reconhecimento da fala usando HMMs são, assim:

- Escolha do número de estados N
- Estimação das probabilidades do modelo $\lambda(\mathbf{A}, \mathbf{B}, p)$ – Treino
- Determinação da máxima verosimilhança de uma sequência de observações.

Uma vez resolvidos esses três problemas o reconhecedor pode ser construído.

3. Processos de Reconhecimento do Falante sem senha

Uma amostra de som também pode ser entendida como um processo estocástico. Os sectores estão indexados de *t=1* até *T*. O índice é, no fundo, uma medida de tempo porque os sectores têm duração igual. A cada sector corresponde um vector de características que pode ser entendido como uma realização ou uma observação de uma variável vectorial aleatória. Dessa forma a amostra é um processo estocástico, uma sucessão de variáveis aleatórias.

Irão ser estudadas situações em que as variáveis aleatórias são não vectoriais, são escalares.

A gama de vectores de características possíveis é contínua. De facto, por exemplo, o valor de um coeficiente cepstral pode ser qualquer, dentro de certos limites. A gama é contínua, não é limitada.

É necessário reduzir essa gama a um número finito. O processo não é tão estranho como isso. É feito constantemente com escalares, porque a precisão de medida e de cálculo que se pode atingir é finita. Ao truncarem-se todos os resultados de uma medida até, digamos à 2ª casa decimal, porque os algarismos seguintes já não são significativos, estão-se a substituir todos os possíveis resultados por uma gama finita. Diz-se, então, que é feita uma quantificação escalar.

O que se tem de fazer com vectores é em tudo paralelo. Substitui-se a gama contínua por um conjunto finito numerado de vectores, os centróides que resultem da divisão da gama em *k* classes. A variável aleatória será o índice identificativo do vector típico, que pode ser codificado. Constitui-se, assim, um dicionário de vectores (*codebook*).

Geralmente é construído um *codebook* para o Falante (tornando-se, assim, numa "representação" do Falante) e é relativamente a este *codebook* que irão ser determinadas as distâncias a que se encontram as amostras do som a reconhecer. Estas poderão ser as distâncias dos vectores de características do som, a distância de um *codebook* do som a reconhecer, bem como a combinação dos dois casos. De notar que para evitar discrepâncias resultantes do facto de umas frases terem maior duração que outras, podendo deste modo levar a que as distâncias dos vectores de características do som a reconhecer ao "som típico" do falante não tenham valor numérico relevante, deve normalizar-se essas distâncias, isto é, dividir essas distâncias pela duração da amostra de som a analisar (mais concretamente pelo número de sectores – *frames* – do som).

4. Conclusões

O problema do reconhecimento do falante pode ser equacionado tendo em conta a aplicação em causa:

- no caso de aplicações ditas comerciais, há que esperar com falantes cooperativos e "tolerantes", isto é, um pequeno erro não será determinante na apreciação global do sistema e a introdução de senhas com o intuito de melhorar o desempenho desse sistema pode ser relativamente tolerada. Contudo a existência destas senhas é sempre mais impopular que a sua ausência, o que leva a um maior esforço no intuito do desenvolvimento de sistemas que as não requeiram
- no caso das aplicações ditas do âmbito forense (ou criminal), o erro permitido é praticamente, se não mesmo, nulo; por exemplo, não é possível chegar-se perto de um juiz e dizer que o sistema acerta 99,9% das vezes, pois esse argumento pode ser rebatido pelo acusado com o facto de, por exemplo, numa cidade com 10 000 habitantes, haver a incerteza quanto a 10 indivíduos, ou seja, há pelo menos, 10 sujeitos passíveis de terem cometido o que quer esteja em causa por parte da acusação. Igualmente, neste caso, existirão os falantes a não colaborar com o sistema de reconhecimento, pois, obviamente, não lhe interessa serem identificados. Também neste caso, os sistemas sem senha são os mais adequados.

Acresce a tudo isto as condições em termos de ruído em que a amostra a ser reconhecida foi retirada, pois vários locais com diferentes níveis de ruído vão "introduzir" sinais diferentes, além de, normalmente aquando da recolha de amostras para treino, se verificarem situações ambiente diferentes quando comparadas aquando das do reconhecimento.

Aliás a quantidade de som a ser analisada desempenha um papel importante, pois durante muito tempo os sistemas, quer de treino quer de teste, requeriam quantidades de som, de cada falante, enormes, levando a cálculos muito morosos, não sendo desprezável este factor aquando da concepção de sistemas de reconhecimento de voz em geral. Actualmente têm-se vindo a desenvolver algoritmos para solucionar esta questão.

Há ainda outra questão fulcral no reconhecimento do falante: a variabilidade intra-falante, isto é, as variações ocorridas ao longo do tempo relativamente à mesma pessoa (por exemplo o *pitch* - frequência fundamental - de uma pessoa pode variar de alguns Hz em poucos dias de um modo completamente aleatório), além de nem sempre termos a mesma voz, nomeadamente através de constipações, tosse, envelhecimento, etc., e se nós, humanos, lidamos com valores qualitativos e, de algum modo ainda não completamente conhecido, compensamos esses factores, já os computadores lidam com valores numéricos e estes fenómenos ainda não foram "modelizados".

Concluindo, apesar de toda a evolução verificada nos últimos tempos, há ainda muito caminho a ser percorrido no campo do Reconhecimento do Falante: é como se só agora se estivessem a dar os primeiros passos nesta área.

5. Bibliografia

The challenge of effective speaker recognition, Hollien, H., Jiang, M., RLA2C - La Reconnaissance du Locuteur et ses Applications Commerciales et Criminalistiques, Avignon, 1998

Apontamentos da disciplina *Processamento Computacional da Fala*, fornecidos pelo Prof. Carlos Espain, Professor Associado da FEUP

Text independent speaker verification using string codebooks, Moreira, F., Espain, C., COST 250 MCM, Porto, 1999

VQ Speaker Verification with Sentence Codebook, Moreira, F., Espain, C., Seminário REC - Reconhecimento de Fala e suas Aplicações em Telecomunicações, Lisboa, 2000