Reconhecimento Automático
De Fala Contínua

Trabalho realizado por:
Luís Filipe Moreira

Índice

1 Introdução

1.1 Introdução

O que é a fala? Esta pergunta é aparentemente ridícula, pois falar é algo tão natural para nós humanos que nem nos damos ao "trabalho" de a colocar, convém, para efeitos do Reconhecimento da Fala, dar-lhe um tratamento mais "científico". Há um modelo de comunicação geralmente aceite, designado *cadeia de fala*, que considera três fases fundamentais neste processo de comunicação: a produção, a transmissão e a recepção da fala.

No processo da produção distinguem-se, basicamente, dois processos: no primeiro o falante transforma a informação que pretende transmitir em símbolos de uma estrutura linguística. O segundo processo consiste em materializar esses símbolos em unidades acústicas. Para tal, são accionados os músculos necessários à geração de fluxo de ar proveniente dos pulmões e para haver uma modelação desse fluxo através da geometria das cordas vocais e do tracto vocal. Deste modo é produzida uma onda de pressão acústica. De notar que este sinal é realimentado no próprio falante através do seu aparelho auditivo, permitindo-lhe, assim, avaliar e controlar o processo de produção de fala.

Na transmissão, o sinal de fala é normalmente afectado pelos mais diversos tipos de ruído (por exemplo, a fala de outras pessoas) e por distorções do canal de transmissão (por exemplo, o corte, a frequências superiores a 4 KHz, numa comunicação via telefone).

Por fim, na recepção da fala, o(s) ouvinte(s) capta(m) a onda de pressão acústica pelo ouvido e tenta(m) extrair a informação nela contida por um processo inverso ao da produção.

O problema do Reconhecimento da Fala consiste em decidir o que foi falado, tendo por base as medições de uma onda de pressão acústica, que, usualmente, é uma versão amostrada e quantisada de um sinal de um microfone. Isto é, de facto, um problema de modelação de um sistema inverso. Para resolver este problema, a análise deve usar técnicas que apliquem conhecimentos acerca das restrições que ocorrem no mecanismo de produção da fala. Tal conhecimento pode estimar os parâmetros articulatórios que ocorreram durante a produção da frase analisada e, em última análise, permite a reconstrução do evento de fala original.

Um factor que pode complicar a resolução deste problema é o facto de a fala gravada raramente ser completa por si só – um facto que está relacionado com a própria natureza da fala como sendo uma ferramenta de comunicação humana. Mais, sendo o meio de comunicação humana por excelência, a fala é uma actividade que envolve uma aprendizagem desde o nascimento; consequentemente, o sinal de fala em si pode não conter toda a informação da mensagem que se pretende transmitir, pois uma parte considerável do processo de comunicação é constituída por outros meios explícitos (por exemplo, gestos, sorrisos, olhares, etc.) e, assim sendo, essa parte é partilhada pelo emissor e pelo receptor da mensagem. Deste modo, para perceber a fala, uma máquina necessitaria de um modelo do mundo similar ao nosso, um problema que é parte do campo da Inteligência Artificial.

O processamento da fala é um campo em rápido desenvolvimento, sendo conduzido por aplicações muito esperadas nas telecomunicações, na interacção homem-máquina, no armazenamento e recuperação de informação, tal como na vida real.

1.2 Desafios

A fala, como processo de comunicação, é o resultado de muito treino desde o nascimento de uma pessoa, treino esse que é aperfeiçoado pelo que parece ser uma capacidade natural para aprender.
O Reconhecimento Automático da Fala (ASR – *Automatic Speech Recognition*) baseia-se na identificação de padrões de fala. Assim sendo, é necessário definir um conjunto finito desses padrões – frases, palavras, fonemas, alofonemas, etc. – que sejam a "unidade" da fala. Contudo, estas unidades são quase sempre difíceis de localizar num sinal e quase impossíveis de segmentar através de meios automáticos. Além disso, apresentam uma grande variabilidade, sendo muito dependentes quer do falante, quer do contexto em que são proferidas. Ainda não foi definido um conjunto de unidades óptimo. Um fonema é uma unidade capaz de induzir uma diferença mínima significativa entre duas locuções (por exemplo, fazer vs. lazer). Os fonemas são dependentes da língua, pois cada uma tem o seu próprio conjunto de fonemas (por exemplo, o português tem o fonema /r/ de raio que o inglês, por exemplo, não tem). Por vezes, a pronunciação dos diferentes fonemas é diferente (por exemplo, o fonema /s/ de suma e de cima é bastante diferente), pelo que a sua manifestação física em contextos diferentes envolveria um conjunto de variantes alofónicas do fonema base. Usa-se, para isso, os alofonemas (por exemplo, um trifone é um alofonema que tem em atenção o que é pronunciado antes e depois do fonema básico).

Tal como na comunicação entre dois humanos, é necessário haver uma cooperação por parte do falante (por exemplo, um professor quando está a fazer um ditado para uma sala cheia, se necessário, fala mais pausadamente do que falaria numa conversa entre um grupo restrito de pessoas).
Por outro lado ambientes acústicos adversos – ruídos de fundo, canais, etc. – em que os sistemas de reconhecimento de fala operam, constituem um grande obstáculo, sendo este, talvez, o maior desafio.
Há que ter em atenção que o Reconhecimento da Fala é diferente da Percepção da Fala – por exemplo, a frase "O porco do João!" pode ser interpretada como sendo referido ao porco que o João tem, ou como referindo-se ao facto de o João ser uma pessoa suja.

1.3 Sistemas de Reconhecimento Automático de Fala

A função de um reconhecedor automático de fala, como referido anteriormente, é identificar a sucessão de símbolos linguísticos, dado um sinal de fala definido num dado intervalo de tempo. Este processo pode ser visto como um reconhecedor de padrões dinâmico.
Existem dois módulos fundamentais: um correspondente à análise do sinal e outro correspondente à sua classificação.
No primeiro módulo, chamado módulo de análise, faz-se a conversão do sinal de entrada numa representação que seja adequada ao processo de classificação.
Se *s(t)* for um sinal de fala definido num dado intervalo de tempo, o módulo de análise transforma-o numa sucessão de T vectores de características de dimensão D,

$$\boldsymbol{X}_1^T = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_T\}, \qquad \boldsymbol{x}_i \in \Re^D$$

Este módulo deve extrair somente as características do sinal que sejam relevantes para o processo de classificação. As componentes do vector de características,

$$\boldsymbol{x}_i = \{x_i^1, x_i^2, ..., x_i^D\}$$

devem ser tais de forma a que haja pouca variação das características dentro de cada classe e de modo a que essa variação seja grande para classes diferentes. Para além de fazer esta discriminação entre classes, este módulo, geralmente, gera uma representação mais compacta do sinal original, o que permite acelerar o processo de classificação, bem como o uso de algoritmos mais poderosos – que são, normalmente, mais pesados em termos computacionais. Deve ter-se em atenção o intervalo de tempo que decorre entre as observações do sinal de modo a evitar a perda de informação entre vectores de características sucessivos.
O segundo módulo, o módulo de classificação, tem como função transformar a sucessão dos vectores de características numa sucessão de símbolos linguísticos pertencentes a um vocabulário Γ e relacionados com as classes padrão:

$$\boldsymbol{W}_1^K = \{w_1, w_2, ..., w_T\}, \qquad w_i \in \Gamma_w$$

1.4 Restrições aos sistemas actuais

Os sistemas actuais e que apresentam resultados interessantes consideram diversas restrições, em simultâneo, ou não, a saber: dependência do falante, palavras isoladas, vocabulário reduzido, gramática muito restritiva.
A independência do falante é vista por muitos, como a maior dificuldade, pois a generalidade das representações paramétricas da fala são fortemente dependentes do falante. Para se ter uma ideia deste obstáculo, refira-se que, como regra empírica, o mesmo sistema apresenta para a mesma tarefa, uma taxa de erro três a cinco vezes superior se for independente do falante, do que se for dependente do falante.
O reconhecimento de fala contínua é significativamente mais difícil do que outros problemas do reconhecimento da fala, essencialmente pelos seguintes motivos:

- as margens das palavras são difíceis de definir, quando não impossíveis
- os efeitos de co-articulação são muito acentuados (quando esse efeito ocorre entre palavras, o que se acentua se se aumentar a velocidade da fala, é ainda mais difícil de tratar)
- as palavras conteúdo (nomes, verbos, adjectivos,…) são frequentemente salientadas, ao contrário das palavras função (artigos, preposições, pronomes,…) que são pobremente articuladas (os fonemas que integram essas palavras são frequentemente encurtados, distorcidos ou eliminados).

Apesar do tamanho do vocabulário a reconhecer não ser a melhor medida da dificuldade do problema, é um factor a considerar, pois a confusão entre as palavras está positivamente correlacionada com a dimensão do vocabulário. Por outro lado, com o aumento do vocabulário, deixa de ser possível modelar cada palavra individualmente devido à limitação da base de treino e do espaço de memória.
Por último, a naturalidade da tarefa é condicionada pelo grau de restrição que se pode estimar através da perplexidade imposta pela gramática (não é uma medida perfeita para a dificuldade da gramática, mas é o melhor standard existente). O aumento da perplexidade (entropia) resulta num substancial acréscimo da taxa de erros. Uma grande dificuldade no estabelecimento de gramáticas mais naturais, e portanto com maior perplexidade, é que exigem bases de dados enormes (por exemplo, o sistema TANGORA usa uma gramática 3-Gram calculada a partir de 25.000.000 palavras) para treinar robustamente o reconhecedor e obter, assim, resultados aceitáveis.

2 História do Reconhecimento Automático da Fala

Conceptualmente, o desenvolvimento do reconhecimento da fala está intimamente relacionado com outros desenvolvimentos da ciência da fala e da engenharia e, como tal, pode ser visto como tendo raízes em estudos da Antiga Grécia.

A primeira máquina a reconhecer fala com algum interesse foi um brinquedo comercial chamado Radio Rex, que foi fabricado na década de 1920. Este brinquedo consistia num cão celulóide, com uma base de ferro segurada pela parte interior do revestimento, contra a força de uma mola. Fluía uma corrente eléctrica pelo magneto através de uma barra de metal disposta de forma a formar uma ponte com dois membros suporte. Esta ponte era sensível a energia acústica de 500 ciclos por segundo (Hz) que a fazia vibrar, interrompendo desse modo a corrente e libertando o cão. A energia à volta dos 500 Hz, contida na vogal da palavra Rex, era suficiente para activar o dispositivo quando se pronunciasse o nome do cão. Apesar de muito simples, este brinquedo incluía um princípio fundamental dos reconhecedores de fala durante muitos anos: armazenar uma representação de uma característica distinta do som desejado e implementar um mecanismo de comparar esta característica com a fala recebida.

O primeiro reconhecedor, propriamente dito, foi um sistema feito nos laboratórios Bell, em 1952, pois permitia ser treinado para reconhecer dígitos de um dado falante. Este sistema media uma simples função do espectro energético no tempo, em duas vastas bandas, correspondentes, de uma forma grosseira, às duas primeiras ressonâncias do tracto vocal (os chamados formantes). O sistema funcionava do seguinte modo: a fala era filtrada em duas componentes, uma de altas e outra de baixas-frequências e cada componente era altamente saturada de forma a que a sua amplitude fosse independente da "força" do sinal. A frequência de corte em cada um dos casos era de 900 Hz, que corresponde a uma fronteira razoável entre o primeiro e o segundo formante para adultos machos. Para cada uma das bandas era contado o número de passagens por zero (*zero crossing*) e o sistema usava este valor para estimar a frequência central para cada banda. O número das baixas-frequências era quantisado numa das seis sub-bandas de 100 Hz cada (entre os 200 e os 800 Hz). O número das altas-frequências era quantisado numa de cinco sub-bandas de 500 Hz cada, começando em 500 Hz. Juntos, estes dois valores quantisados correspondiam a um dos 30 pares de frequências possíveis. Este sistema era analógico.

Em 1958, Dudley fez um classificador que avaliava o espectro de um modo contínuo, em vez de o fazer por aproximações aos formantes. Este novo paradigma foi uma prática comum posteriormente. De facto, genericamente falando, o paradigma actual dominante para o reconhecimento da fala utiliza a funcionalidade de um espectro local variando ao longo do tempo como sendo a representação da fala que entra no sistema.

Em 1959, Denes, acrescentou probabilidades gramaticais adicionalmente à informação acústica. Por outras palavras, ele anotou que a probabilidade de ser pronunciada uma unidade linguística em particular pode ser dependente da unidade linguística anterior, pelo que a probabilidade de uma palavra não tem de ser unicamente dependente da entrada acústica.

Martin desenvolveu as redes neuronais artificiais (ANN) para o reconhecimento de fonemas em 1964.

Na década de 1960 foram desenvolvidas três técnicas de estimativa espectrais que foram de grande importância no que diz respeito ao reconhecimento, isto apesar das suas primeiras aplicações terem sido para codificação de voz, nomeadamente a Transformada de Fourier Rápida (*Fourier Fast Transform – FFT*), a análise cepstral e a Codificação de Predição Linear (*Linear Predictive Coding – LPC*). Adicionalmente, foram desenvolvidos novos métodos de comparação de padrões de sequências, nomeadamente uma abordagem determinística chamada Alinhamento Temporal Dinâmico (Dynamic Time Warp – DTW) e uma abordagem estatística chamada Modelos Escondidos de Markov (*Hidden Markov Models – HMM*).

Na década de 1970, a Agência de Projectos de Investigação Avançados (A*dvanced Research Projects Agency – ARPA*) subsidiou um projecto de compreensão da fala vasta. O objectivo era efectuar o reconhecimento automático de fala com um vocabulário de 1000 palavras, usando um pequeno número de falantes, fala contínua e uma gramática restrita com menos de 10% de erro semântico.

Em 1986, a primeira grande base de dados (a TIMIT) começou ser usada por uma grande parte da comunidade científica. Foi escolhido um alfabeto de 61 fonemas para representar as distinções fonéticas. Os dados foram gravados e segmentados foneticamente, primeiro através de uma segmentação automática seguida de uma inspecção manual e reparação dos alinhamentos. Isto resultou numa base de dados em que os limites temporais de cada fonema no sinal de voz estão marcados para cada fonema proferido por cada falante. Esta constitui, ainda hoje, uma das maiores e mais utilizadas bases de dados existentes.
Mais tarde foram elaboradas outras bases de dados para o reconhecimento da fala, nomeadamente a TIDIGITS, a RM – *Resuyrce Management*, só para referir algumas.

Os HMM's tiverem a sua origem em finais da década de 1960, tendo as primeiras aplicações no reconhecimento da fala sido feitas nos anos setenta. Nos anos oitenta, esta abordagem suscitou o interesse de uma maior parte da comunidade, tendo a investigação e o desenvolvimento nesta área levado a melhorias em sistemas em muitos laboratórios. Em meados da década de 1980, os HMM eram a ferramenta predominante no reconhecimento.

Em 1984, a ARPA financiou um segundo programa, tendo sido a primeira grande tarefa deste programa a elaboração da RM. Esta tarefa envolveu a leitura de frases derivadas de um vocabulário de 1000 palavras. Estas frases consistiam em questões e comandos elaborados para manipular uma base de dados de informação naval, apesar dos sistemas não terem interferido, de facto, com qualquer base de dados. As classificações foram baseadas no reconhecimento de palavras.
Este projecto serviu como catalizador para muitos avanços de engenharia. Em 1998, havia muitos sistemas capazes de reconhecer fala lida de novos falantes sem treino específico para cada falante, com um vocabulário de 60000 palavras em tempo real, com menos de 10% de erro de palavras.

Quando, em 1969, Minsky e Papert escreveram um livro intitulado *Perceptrons* em que provaram que o perceptrão não conseguia, sequer, realizar uma tarefa tão simples como a função lógica XOR, as redes neuronais sofreram um grande revés. Contudo, com o advento da *backpropagation*, uma técnica de treino para perceptrões em múltiplas camadas (MLP), no início dos anos 1980, o campo das redes neuronais ressurgiu.
Uma das primeiras aplicações, no início dessa década, foi o uso de redes neuronais para classificar um som de fala como sendo vozeado (*voiced*) ou não (*unvoiced*).

Em finais dos anos 1970, princípios dos anos 1980, várias abordagens baseadas na codificação do conhecimento humano, tipicamente sob a forma de regras, foram amplamente usadas num grande número de disciplinas. Alguns investigadores da área da fala usaram sistemas de reconhecimento usando conhecimento acústico-fonético para desenvolver regras de classificação para sinais de fala. Por exemplo, as consoantes /k/ e /g/ seguidas de uma vogal eram discriminadas com base na proximidade da segunda e terceira ressonâncias no fim da vogal. Uma das potenciais vantagens desta abordagem é que as características da fala usadas para fazer a discriminação não eram limitadas à acústica de uma simples *frame.*

Ainda é relativamente cedo para se fazer um julgamento do trabalho efectuado na última década, mas pode-se desde já destacar que houve vários programas do tipo ARPA – que prosseguiu como DARPA. Além disso, têm-se realizado conferências com maior regularidade e projecção.
Na última década tem sido feito um grande esforço para melhorar a robustez no reconhecimento através de diferentes canais, bem como atender às diferenças dos microfones usados para a captação do sinal de fala e o ruído acústico.
Tem havido um aumento na ênfase de assuntos como a pronúncia, o modelo de diálogo, dependências de longas distâncias dentro de sequências de palavras e modelação acústica, só para referir alguns grandes tópicos.
Tem havido uma rápida expansão da investigação noutras tarefas relacionadas com o reconhecimento automático da fala, tais como a identificação e verificação do falante e identificação da língua.

3 Processamento de sinal no Reconhecimento de Fala

Uma das medidas chave usadas no processamento da fala é o espectro de curta-duração. Em todas as suas formas, tal consiste no cálculo de uma estimativa qualquer do espectro local, usualmente feita sobre uma curta porção da fala (tipicamente secções de 20 a 30 ms).
Há três abordagens básicas: banco de filtros, processamento cepstral e Coeficientes de Predição Linear (LPC).

3.1 Banco de filtros

Esta é a mais antiga das abordagens e consiste em fazer estimativas de energia atenuadas temporalmente através de um banco de filtros. A grande inspiração para esta abordagem foi o sistema auditivo humano, por isso convém interpretá-lo como um banco de filtros.
Algumas experiências sugerem a existência de um filtro auditivo na vizinhança de um dado tom que impede a informação extrínseca a esse tom de interferir com a sua detecção. Esta vizinhança e chamada *banda crítica* e pode ser vista como a banda de passagem de um filtro (centrada no tom a medir). Os resultados experimentais mostram que a banda crítica aumenta com a frequência central do tom. Algumas medições mostram que para baixas frequências esta banda varia entre os 100 e os 800 Hz.
Patterson demonstrou que estes filtros auditivos podiam ser representados por um filtro do tipo:

$$|H(f)|^2 = \frac{1}{\left[(\Delta f / a)^2 + 1\right]^2}$$

em que o parâmetro α é uma medida da selectividade do filtro. Esta função é simétrica numa escala de frequências linear. Este filtro é um filtro APGF (*all-pole gamma-tone filter*), cuja transformada de Laplace é

$$H(s) = \frac{K}{\left[(s-p)(s-p^*)\right]^N}$$

com N=2.
Quantos filtros deste tipo serão necessários para emular o banco de filtros auditivo? A reposta a esta questão tem sido dada com base na experiência e erro. Assim, um sistema com 30 canais demonstrou ter bons resultados, mas este sistema não aproveitou a resolução em frequência variável do ouvido. Em sistemas com maiores bandas de passagem para frequências centrais mais elevadas, conseguiu-se reduzir esse número para 16.

3.2 Análise cepstral

A estrutura base de um sinal de fala pode ser identificada como uma excitação – trem de impulsos à saída das cordas vocais – que é a entrada de um sistema de produtores de ressonância – cavidades do tracto vocal. A convolução da excitação com a resposta impulsional dos ressoadores produz o sinal da fala. É pois perfeitamente natural, contemplar uma análise do sinal como ele é inicialmente – a excitação. A análise cepstral

faz a deconvolução do sinal de fala, permitindo obter representações da excitação inicial e dos filtros.

Considere-se o sinal de fala, em que X se refere ao espectro desse sinal, E à componente de excitação e V à modelação que a excitação sofre no tracto vocal. Assim sendo, a magnitude do sinal fala vem

$$|X(w)| = |E(w)| \cdot |V(w)|$$

3.3 Coeficientes de Predição Linear (LPC)

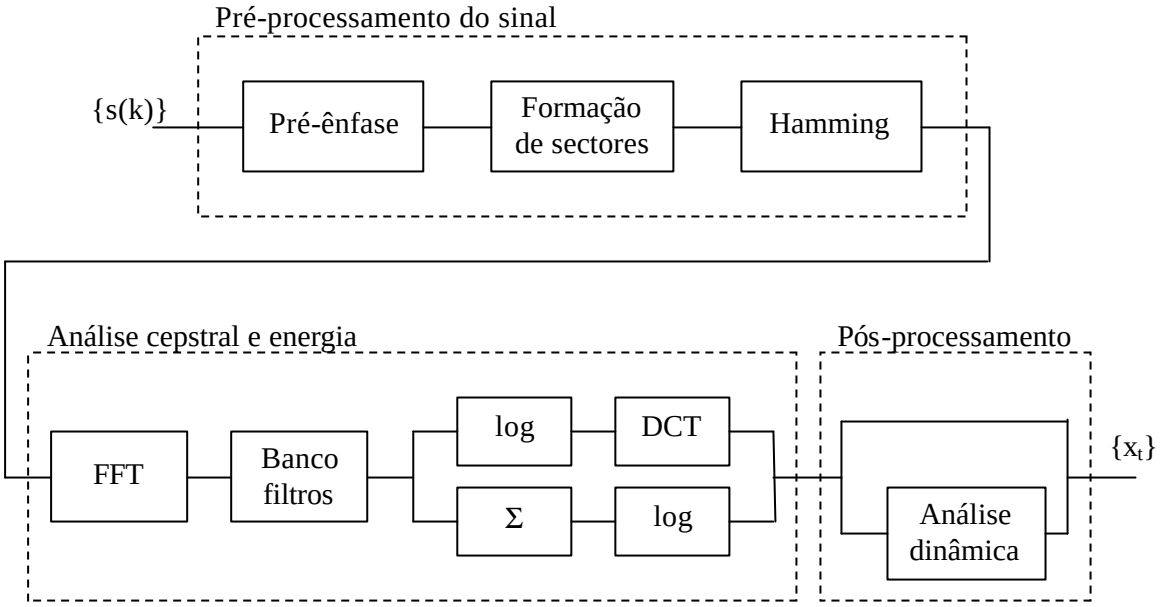
O trato vocal pode ser visto como um tubo não uniforme constituído por múltiplos pequenos tubos concatenados, com áreas de corte transversal diferentes, mas com o mesmo comprimento. Este tubo resultante teria um conjunto de ressonâncias semelhantes às do tracto vocal real (na forma requerida para produzir um dado som).
Os tractos vocais reais geram ressonâncias cujo número pode ser razoavelmente predito por modelos de tubos.
Supondo que cada frequência de ressonância – formante – pode ser representada por uma função de transferência só com um pólo do tipo:

$$H_i(z) = \frac{1}{1 - b_i z^{-1} - c_i z^{-2}}$$

Este modelo funciona bem para sons vozeados (vogais), pois filtros de oitava ou décima ordem representam bem quatro formantes em boas condições. Contudo, quando o som é nasalado, há um acréscimo significativo de zeros pelas cadeias laterais, fazendo com que o resultado seja pior. Os sons não-vozeados – *unvoiced* – têm um espectro mais simples, ocorrendo uma sobre parametrização. Mesmo assim, os LPC funcionam bem nestes casos.
Nos LPC o efeito de *aliasing* é "encorajado" ao definir a frequência dos cantos a 50% da frequência de Nyquist, em vez dos habituais 40%. Esta escolha colocará uma característica de filtro passa-baixo íngreme dentro da banda modelada pela análise LPC. Essencialmente, a análise só tem um número fixo de pólos a colocar de modo a modelar o espectro e, tendo um filtro assim tão íngreme, vai reduzir o grau de liberdade para modelar a fala.
Os formantes estão muito associados com a identificação fonética, em particular para as vogais. Contudo, há uma grande variação nestas frequências entre as pessoas (homens vs mulheres vs crianças), pelo que há a tendência a haver uma dependência do falante nas características baseadas nos LPC.

3.4 Extracção de características

Nos métodos tradicionais destacam-se dois grupos: os baseados no espectro de predição linear (LPC) associado ao modelo de produção acústica da fala (por exemplo, os LPC) e os baseados no espectro de Fourier do sinal de fala (entre eles os coeficientes mel-cepstrais – MFCC). Estes últimos são menos exigentes computacionalmente.
Um esquema para a extracção de características pode ser o seguinte:



Pré-processamento

Pré-ênfase: logo após a amostragem e quantificação, o sinal é submetido a um filtro linear de primeira ordem com o objectivo de nivelar a representação espectral do sinal. A função de transferência desse filtro é:

$$H(z) = 1 - az^{-1}, \quad a \in [0{,}91;0{,}99]$$

Formação de sectores e janela: a sucessão de vectores de características é determinada analisando uma sucessão de sectores do sinal, a qual é necessário definir. Seja *s(k)* o sinal à saída do filtro de pré-ênfase e seja *h(k)* uma função também de variável discreta, correspondente a uma janela:

$$h(k) = a - (1-a)\cos\left(2p\,\frac{k-1}{L-1}\right), \qquad 1 \le k \le L$$

Assim a sucessão de sectores $\{s_n(k); \quad n = 1,2,...; \ k = 1,2,...,L\}$ é definida por:

$$s_n(k) = s(k + nD + \Delta)h(k), \quad n = 1,2,...; \ k = 1,2,...,L$$

A sucessão de sectores é o resultado da observação do sinal em intervalos regulares (tipicamente 10 ms), sendo cada observação realizada através de uma janela com uma duração que varia entre os 20 e os 30 ms, tipicamente.
A duração de cada sector é definida de modo a obter um compromisso razoável entre a resolução nas representações temporal e espectral do sinal, à semelhança do que sucede com o sistema auditivo humano.

4 Outras disciplinas

Para que a generalidade dos sistemas de reconhecimento automático de fala – em especial os utilizados em aplicações mais exigentes – possam atingir um desempenho aceitável, é necessário que sejam capazes de processar eficientemente a informação que se encontra associada ao sinal da fala em diferentes níveis

Há dois grandes níveis de processamento: um acústico e outro linguístico.

Ao nível acústico, o reconhecedor tenta fazer a relação directa entre vários segmentos do sinal de fala e os módulos acústicos que representam internamente esses segmentos. Este nível é normalmente considerado básico, pois as correspondências feitas têm por base unicamente regras numéricas de similitude entre as características físicas do sinal de fala e os módulos acústicos.

Ao nível linguístico há quatro grandes sub-níveis: o lexical, que transforma os fonemas em palavras; o sintáctico, que contém regras que definem a sucessão de palavras (frases); o semântico, que tem em atenção o significado das palavras e a evolução do seu sentido no conteúdo da mensagem; o pragmático, que tenta extrair o significado da mensagem no contexto em que é produzida. De notar que os níveis lexical e sintáctico são altamente dependentes da língua (cada língua tem o seu próprio conjunto de palavras, bem como a sua própria estrutura das palavras nas frases).
Existem várias estruturas para estes sub-níveis: algumas são hierárquicas com bidireccionalidade de informação; outras são não hierárquicas; outras são ainda híbridas.

5 Tipos de reconhecimento

O Reconhecimento Automático da Fala pode ser visto como um caso particular de reconhecimento sequencial. Assim, para uma sequência de observação de sectores X, deseja-se associá-la a uma segunda sequência Q de unidades linguísticas e em que Q é escolhida de modo a minimizar um dado critério de erro. Para simplificação, considere-se a segunda sequência restringida a outra sequência de observações usada como referência.

5.1 Reconhecimento de Palavras Isoladas

Dadas duas sequências indexadas de vectores de características e a distância métrica entre dois vectores quaisquer, e caso as sequências tenham o mesmo número de vectores, pode determinar-se a distância (ou distorção) global entre as duas, como sendo o somatório das distâncias locais entre cada sector correspondente das sequências a ser comparadas:

$$D\left(X_k^{ref}, X^{in}\right) = \sum_{i=1}^{N} d\left(x_{ik}^{ref}, x_i^{in}\right)$$

em que *d* é a distância local, *D* a distância total e *N* o tamanho das sequências.
Contudo, na fala, raramente o número de sectores em duas sequências é igual, além de que as sílabas, por exemplo, têm durações diferentes em observações diferentes. Logo, o alinhamento temporal linear não é o mais aconselhado para esta tarefa. É preferível um alinhamento temporal dinâmico (DTW) de modo a alinhar as unidades fonéticas das palavras proferidas. Assim sendo, o DTW consiste em determinar o melhor caminho entre o ponto de partida e o ponto de chegada (que são iguais para as sequências que se pretende comparar).

5.2 Reconhecimento de Fala Contínua

Para este caso poderia pensar-se em fazer o alinhamento temporal da entrada com as referências para cada sequência de palavras possível. Isto não é, obviamente, exequível dado o número de sequências de palavras possíveis ser muito grande (teoricamente infinito, no caso geral), além de ser computacionalmente muito moroso.
Só recorrendo à estatística se pode reduzir as dimensões citadas anteriormente; um caso em que a estatística está presente é o dos HMM. Aqui o que se faz é arranjar modelos das unidades básicas e calcular a sua sequência com base em HMM's. Além disso, podem fazer-se assumpções que nos indicam um conjunto de recorrências que retêm as semelhanças das sequências acústicas. Estas semelhanças, em combinação com probabilidades à priori de sequências de palavras, indicam a hipótese da sequência de palavras que retém a probabilidade de erro mínimo.

5.2.1 Estatísticas N-Gramaticais

Para cada palavra existe uma lista possíveis palavras que se lhe seguem; uma generalização deste raciocínio é associar a probabilidade em cada palavra com o conjunto de palavras seguinte, obtendo, assim, um modelo *bi-grama*. De um modo semelhante, uma lista de palavras com probabilidades associadas que podem suceder a cada par de palavras é chamado um modelo *tri-grama*. Estes modelos são geralmente denominados de *N-gramas*, em que N é um inteiro pequeno.
Como exemplo para uma probabilidade bi-grama, tem-se:

$$P(\text{Eu comi pão}) = P(\text{eu} \mid \text{início}) \cdot P(\text{comi} \mid \text{eu}) \cdot P(\text{pão} \mid \text{comi})$$

Para N-gramas, as probabilidades são calculadas a partir da contagem de co-ocorrências das palavras:

$$P(\text{comi} \mid \text{eu}) = \frac{P(\text{eu comi})}{P(\text{eu})} = \frac{n.^{\text{o}}\ ocorrências(\text{eu comi})}{n.^{\text{o}}\ ocorrências(\text{eu})}$$

Isto faz-se com o pressuposto de que há uma transcrição de palavras no conjunto de treino. De notar que este processo só é fácil para muitas ocorrências das palavras.

6 Conclusões

Actualmente existem alguns sistemas de Reconhecimento Automático de Fala (por exemplo, o *FreeSpeech* da Philips); contudo, sistemas como este sofrem dos problemas já citados – só para dar um exemplo, ninguém tem "paciência", depois de comprar um sistema, para estar a treiná-lo durante mais de 15 minutos para obter uma taxa de acerto de palavras satisfatória e, ainda por cima, que só apresenta essa taxa para essa pessoa – adaptação ao falante.

No reconhecimento é muito importante distinguir o ruído ambiente do sinal de fala; embora muitos sistemas funcionem bem em laboratório, em cenários "reais" já não se pode dizer o mesmo: ainda falta robustez aos sistemas.

Desde os HMM's, na década de 1970, não foi criado qualquer grande novo mecanismo de reconhecimento, mas houve, sobretudo, um grande desenvolvimento das técnicas já existentes, nomeadamente na extracção de parâmetros e na estimativa probabilística. Coloca-se agora a questão de saber se o amadurecimento das técnicas actualmente em uso será suficiente para permitir grandes desenvolvimentos no futuro, ou se será necessário haver mudanças radicais neste campo para futuros desenvolvimentos.

Concluindo, só resta dizer que ainda há muito trabalho a ser feito neste campo.

7 Bibliografia

- Apontamentos da disciplina *Processamento Computacional da Fala*, fornecidos pelo Prof. Carlos Espain, Professor Associado da FEUP
- *Automatic speech and speaker recognition – advanced topics*, Chin-Hui Lee, Frank K. Soong, Kuldip K. Paliwal, Kluwer Academis Publishers, 1996
- *Automatic speech recognition: the development of the SPHINX system*, K.-F. Lee, Kluwer Academis Publishers, 1989
- *Discrete-time processing of speech signals*, John R. Deller, John G. Proakis, John H. Hansen, Macmillan Publishing, 1993
- *Fundamentals of speech synthesis and recognition: basic concepts, state of the art and future challenges*, Eric Keller, John Wiley & Sons, 1994
- *Reconhecimento de Fala Contínua com processamento simultâneo de diferentes características do sinal*, Vítor Manual Pêra, FEUP, 2001
- *Speech and audio signal processing – processing and perception of speech and music*, Ben Gold, Nelson Morgan, John Wiley & Sons, 2000