2
Sistemas de Reconhecimento de Voz

O desenvolvimento de interfaces homem-máquina controladas pela voz visa substituir, em certas aplicações, as interfaces tradicionais tais como teclados, painéis e dispositivos similares. Neste cenário se insere o reconhecimento de voz que permite a realização de uma interface, mais natural, entre os sistemas computacionais e o homem.

A pesquisa em reconhecimento automático da fala (Automatic Speech Recognition – ASR), apresentou avanços significativos nas últimas décadas. Já existem sistemas desenvolvidos para o reconhecimento da fala para uma ampla gama de aplicações, envolvendo desde o emprego de vocabulários pequenos para o reconhecimento através de linhas telefônicas, até a compreensão espontânea da fala [1].

Existem três tipos básicos de sistemas de reconhecimento de voz:

Reconhecedor de palavras isoladas – são os mais simples e pode-se, até, considerar que, para vocabulários pequenos e ambiente limpo, sua tecnologia está dominada. Existem diversos sistemas comerciais amplamente utilizados, tanto na versão em *software* como na versão em *hardware*, cuja taxa de reconhecimento, independente do locutor, é de aproximadamente 100% para vocabulários muito pequenos (até 10 palavras) e de aproximadamente 95%, para vocabulários pequenos e médios (de 10 até 1000 palavras).

A principal exigência dos reconhecedores de palavras isoladas é que as locuções a serem reconhecidas devam ser pronunciadas com pausas maiores que 200 ms entre elas, ou seja, é necessário locutor cooperativo. Apesar disso e do fato de que a maioria utiliza poucas palavras, são bastante empregados nas áreas de comando e controle, reserva de passagens, consulta de saldos bancários, cartões de crédito e nas tarefas em que a interface com usuário pode se dar na base de seleção de opções.

Reconhecedor de palavras conectadas – são sistemas mais complexos que os anteriores, utilizam palavras como unidade fonética padrão e reconhecem sentenças pronunciadas de forma natural (sem pausas entre as palavras). Entretanto, essas sentenças devem ser bem pronunciadas.

Reconhecedor de voz contínua – são os mais complexos e difíceis de serem implementados, pois devem ser capazes de lidar com todas as características e vícios da forma natural de falar, dentre os quais podem ser citados: durações de palavras desconhecidas, efeitos de coarticulação (por exemplo, o /z/ de *deu zebra* e de *belíssimas zebras*) e pronúncia descuidada.

Nesta dissertação será utilizado um reconhecedor de voz de palavras isoladas, que, neste quesito, é o mais simples. Porém, não é esperado que se atinja os percentuais de reconhecimento descrito por causa da presença de outros fatores do sistema aqui considerado, os quais serão tratados no sentido de buscar aproximar o desempenho do reconhecedor ao daquele descrito anteriormente.

É importante ressaltar que, embora muito se tenha aprendido a respeito da forma como implementar sistemas práticos e úteis de reconhecimento de voz, ainda resta um árduo caminho a ser percorrido. Cinco são os principais fatores que determinam a complexidade de qualquer sistema de reconhecimento da fala [2]:

O locutor. Este talvez seja o aspecto que introduza maior variabilidade na forma de onda do sinal de entrada requerendo, portanto, que o sistema de reconhecimento seja altamente robusto. Uma pessoa não pronuncia uma locução sempre da mesma forma devido a distintas situações físicas e psicológicas (chamadas variações intralocutores). Existem, ainda, enormes diferenças entre os tipos de locutores (homens, mulheres e crianças), entre diferentes idades ou regiões de origem (chamadas variações interlocutores). Assim, a forma mais simples de aplicação é aquela em que o sistema funciona para um determinado locutor, tendo sido previamente treinado por este mesmo locutor (sistema dependente do locutor). Porém, para o sistema que será estudado aqui estará sendo utilizada uma situação mais complexa, onde o mesmo terá que ser capaz de reconhecer locuções independentemente do locutor, sem levar em conta idade,

sexo ou até mesmo variações climáticas. Para isto estará sendo usada uma base de dados de vozes com cinqüenta locutores do sexo masculino e cinqüenta locutores do sexo feminino, onde cada locutor realizou três repetições dos dígitos 0,1,2,3,4,5,6,7,8,9 e a palavra *meia*, totalizando 3300 locuções.

A forma de falar. Segundo fator a determinar a complexidade de um sistema reconhecedor. O ser humano pronuncia as palavras de forma contínua e devido à inércia dos órgãos articulatórios, que não se movem instantaneamente, são produzidos os efeitos coarticulatórios. Estes, unidos às variações introduzidas pela prosódia (pronúncia regular das palavras, em harmonia com a acentuação [3]), fazem com que haja diferença entre uma mesma palavra dita no início e no meio de uma frase.

O vocabulário. Corresponde às diferentes palavras que o sistema deve ser capaz de reconhecer. Portanto, quanto maior é o vocabulário mais árdua torna-se a tarefa de reconhecimento, por dois motivos. Primeiro, porque ao aumentar o número de palavras é mais fácil que apareçam palavras parecidas entre si. Segundo, porque o tempo de tratamento é proporcionalmente maior à medida que aumenta o número de palavras com as quais devem ser feitas comparações. Uma solução possível para este problema é utilizar unidades fonéticas menores que palavras (fones, sílabas, etc).

Nesta dissertação não será necessário utilizar unidades menores que palavras, devido ao fato do dicionário ser de pequenas dimensões, 11 palavras.

A gramática. É o conjunto de regras que limita o número de combinações permitidas às palavras do vocabulário. Em geral, a existência de uma gramática em um reconhecedor ajuda a melhorar a taxa de reconhecimento, ao eliminar ambigüidades. Isso também ajuda a diminuir a carga computacional, ao limitar o número de palavras em uma determinada fase do reconhecimento.

Não será necessário utilizar gramática nesta dissertação pois será feito reconhecimento de palavras isoladas com vocabulário pequeno.

O ambiente. Fator tão ou mais importante que os anteriores para definir o reconhecedor, o ambiente é responsável pela inserção de ruído. Trabalhando-se

fora do ambiente de laboratório tal influência torna-se inevitável, pois este ruído pode aparecer de diversas formas possíveis, desde vozes de outros locutores, sons de equipamentos, ar condicionado, luz fluorescente e, até mesmo, provocados pelo próprio locutor, tais como tosses, espirros, estalo dos lábios, suspiro, respiração forte, etc. Por esse motivo, implementar sistemas reconhecedores de voz robustos a ruído, constitui condição *sine qua non* para melhorar o desempenho dos sistemas empregados hoje em dia e para assegurar a qualidade dos que estão por vir.

No caso do sistema de reconhecimento proposto não foi realizada análise em presença de ruído externo, sendo este um cenário para trabalhos futuros abordado no Capítulo 7. O único ruído que estará sendo obstáculo ao sistema, será o inserido pelo codificador e decodificador dos sistemas celulares/voz sobre IP que trabalham a baixas taxas, degradando bastante o sinal de voz.

Com a finalidade de preparar toda a base necessária ao sistema a ser apresentado no Capítulo 3 e abordado no decorrer desta dissertação, é importante que se conheça melhor as principais características do sistema de reconhecimento de voz, o que será feito nas próximas seções.

2.1. Características e Modelo de Produção do Sinal de Voz

Um caminho para se ter um bom desempenho no reconhecimento de voz passa pelo conhecimento do sinal a ser estudado, ou seja, sons da fala e suas classificações. É também importante que se conheça um modelo de produção da fala, o qual é a base dos codificadores paramétricos de voz. Estes conhecimentos são de grande importância neste trabalho, pois são a base dos parâmetros de reconhecimento de voz que são apresentados no Capítulo 4.

2.1.1. Sons da Fala

A linguagem humana compreende um número de elementos básicos de sons, denominados *fonemas*. Esses elementos são caracterizados pelo fato de que duas palavras diferem, se pelo menos um de seus elementos básicos diferirem.

A Tab. 2.1 apresenta exemplos de fonemas do português brasileiro, língua sobre a qual estará se fazendo o reconhecimento:

Tipos de Sons	Exemplos
Vogais	/i /- v*i*, /ɛ/- v*e*la, /e/- v*ê*, /a/- v*a*la, /u/- *u*va, /ô/-bobo, /o/- b*o*la
Fricativas Sonoras	/v/- chu*v*a, /z/- *z*elo, / ž /- ge*l*o
Fricativas Surdas	/f/ - *f*aca, /s/- *s*ala, /š/- *ch*uva
Oclusivas Sonoras	/b/- *b*ato, /d/- *d*edo, /g/- *g*ola
Oclusivas Surdas	/p/- *p*ato, /t/- *t*atu, /k/- *c*apa
Nasais	/m/- *m*ala, /n/- *n*ada, /ñ/- ma*nh*a
Laterais	/l/- ca*l*a, / ľ / - ca*lh*a
Vibrantes	/r/- ca*r*a, / R̃/ - ca*rr*o

Tabela 2.1 – Fonemas do Português brasileiro

Os sons da fala podem ser classificados em *continuados* e *não-continuados*. Os continuados são as vogais, as fricativas e as nasais e são classificados desta forma pois durante a produção do som a forma do aparelho vocal é invariante no tempo. No caso dos sons não-continuados – como os sons oclusivos – o aparelho vocal muda a sua configuração, não sendo invariante no tempo.

Dependendo da presença ou ausência de vibração das cordas vocais, os sons são divididos, respectivamente, em *sonoros* ou *surdos*. Essas duas categorias apresentam características acentuadamente distintas e são de fundamental importância na construção do modelo de produção que será apresentado posteriormente. A Fig 2.1 ilustra amplitudes de formas de onda de voz de sons sonoros e surdos com 64 ms de duração:

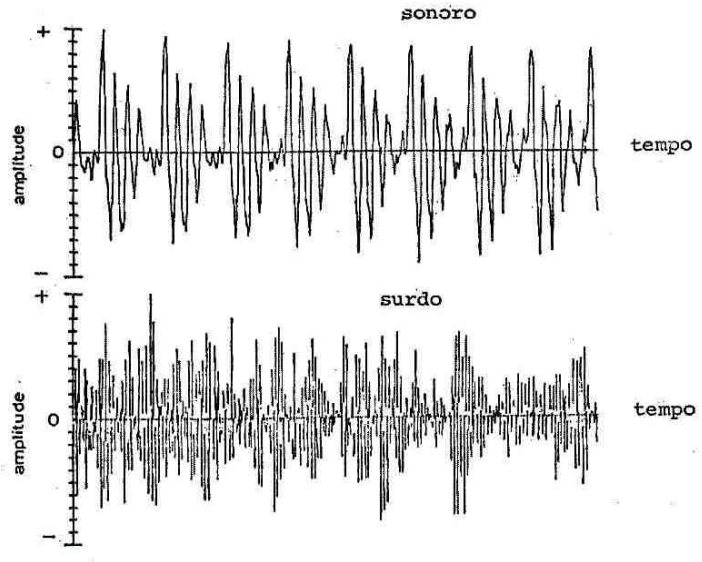

Figura 2.1 – Formas de onda de sons sonoro e surdo

Observa-se que, para os sons sonoros, a forma de onda é aproximadamente periódica. O intervalo T_0 entre os picos principais fornece uma medida do *período fundamental* para aquele locutor em particular. O inverso de T_0 corresponde à freqüência fundamental F_0, que pode apresentar variações de uma oitava no decorrer de uma sentença falada por uma mesma pessoa. Freqüências fundamentais médias típicas para homens e mulheres são, respectivamente, 120 Hz e 220 Hz. A duração do período fundamental pode variar mesmo de um período para o seguinte. Estas variações determinam a característica de entonação, que é extremamente importante para a naturalidade da fala humana. Apenas máquinas são capazes de gerar uma voz perfeitamente monótona. A Fig. 2.2 ilustra as fases sucessivas de um período de vibração das cordas vocais.

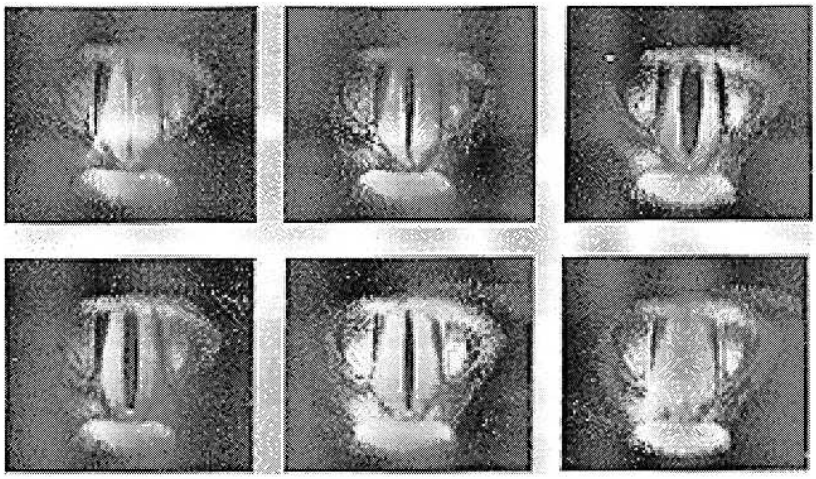

Figura 2.2 – Fases sucessivas em um período de vibração das cordas vocais

Os sons da fala são caracterizados por sua forma – a envoltória do sinal – no domínio da freqüência, que está univocamente relacionada à configuração do aparelho vocal. Uma análise das formas de onda ilustradas na Fig 2.1 mostra os sons sonoros com características que se aproximam de oscilações amortecidas, conseqüência das ressonâncias espectrais – também denominadas *formantes* – da cavidade vocal e refletem características particulares do aparelho vocal do locutor. Já sons surdos exibem características de ruídos, apresentando concentração de energia em altas freqüências.

Dependendo da caracterização acústica, os sons da fala podem ser classificados em vogais, nasais, fricativas e oclusivas, cujas características são descritas abaixo:

Vogais – são sons sonoros – resultado de vibrações das cordas vocais – e relativamente longos, apresentando duração tipicamente maior que 150 ms. Este tempo depende entretanto da vizinhança fonética.

Fricativas – podem ser sonoras – como /v/ em chuva – ou surdas – como /f/ em faca – dependendo da presença ou ausência de vibração das cordas vocais. Ambas, entretanto, são caracterizadas por uma fonte de ruído contínuo produzido através de turbulência criada em constrições estreitas no aparelho vocal. O local das constrições distingue os vários sons.

Oclusivas – também podem ser sonoras – como /b/ em bola – ou surdas – como /p/ em pato. Estes sons são caracterizados por um excesso de pressão criado em um ponto do aparelho vocal, seguido de um desprendimento repentino de ar. O acréscimo de pressão é observado como um intervalo de silêncio e, no caso de oclusivas sonoras, ocorre vibração das cordas vocais. Após o desprendimento de pressão, a oclusiva geralmente se assemelha a uma fricativa de curta duração.

Para os sistemas de reconhecimento é necessário que se conheça as unidades fonéticas, pois é a partir delas que se define a estrutura do reconhecedor e com que unidade básica o mesmo irá trabalhar.

2.1.1.1. Unidades Fonéticas

Um sistema reconhecedor de voz pode ser caracterizado, entre outros fatores, pela unidade fonética por ele empregada [4,5].

Existem dois critérios importantes que devem ser analisados durante a escolha dessas unidades:

Consistência – a unidade deve ter características similares em sentenças diferentes. É importante porque permite uma discriminação efetiva entre unidades distintas.

Treinabilidade – devem existir amostras suficientes para o treinamento e a criação de um modelo com bom desempenho nos testes. Sua importância reside no fato de os modelos atualmente usados no reconhecimento exigirem grandes quantidades de dados de treinamento.

A seguir, serão descritas algumas das unidades mais utilizadas segundo esses dois critérios.

Palavras – são unidades mais naturais da voz porque são exatamente elas que se quer reconhecer. Além disso, os sistemas baseados em palavras são capazes de captar os efeitos das coarticulações que existem dentro delas. Desse modo, quando existem dados suficientes para o treinamento, um sistema utilizando palavras como unidades básicas apresenta melhor desempenho [6]. Entretanto, quando o vocabulário é grande, existe um grande problema que contra-indica o seu emprego: o tempo de busca. Como todas as palavras do vocabulário têm que ser comparadas com a palavra em teste, o tempo de comparação torna-se excessivo. À medida que o vocabulário do sistema vai aumentando, esta demora torna-se cada vez mais problemática.

Outro problema, não tão grave quanto o anterior, mas que também possui importância, é que a memória utilizada por tais sistemas aumenta linearmente com o aumento do vocabulário (número de palavras). Além disso, para muitas tarefas, é conveniente prover o usuário com a opção de acrescentar novas unidades ao vocabulário; pois, se este sistema for utilizado, o usuário terá que produzir muitas repetições de cada nova palavra para o treinamento, o que será muito inconveniente. Conseqüentemente, embora estes sistemas sejam uma opção natural, além de modelarem bem as transições fonéticas que acontecem dentro da palavra, devido ao fato de que cada palavra do vocabulário irá corresponder a um modelo, eles não são práticos para reconhecedores de voz contínua que utilizem grandes vocabulários. Por exemplo, para serem obtidos modelos confiáveis de cada palavra do vocabulário, o número de sentenças do conjunto de treinamento

precisa ser suficientemente grande, de modo que cada palavra do vocabulário precisa aparecer em cada contexto fonético possível, várias vezes. Somente desta maneira a variabilidade acústica do início e do fim pode ser modelada apropriadamente.

Usando como exemplo o caso de reconhecimento de dígitos, sabe-se que cada um pode ser precedido e seguido por qualquer outro, assim, um vocabulário de 11 dígitos (zero a nove mais meia) possui exatamente 121 contextos fonéticos (alguns dos quais essencialmente idênticos, como ocorre com a seqüência seis quatro sete e a seqüência três quatro seis). Com um conjunto de treinamento de vários milhares de cadeias de dígitos é realístico e prático ver cada dígito em cada contexto fonético várias vezes. Agora, considerando-se um vocabulário de 1000 palavras com uma média de 100 contextos fonéticos para o início da palavra e 100 para o fim, para que cada palavra seja considerada em cada contexto possível, serão necessárias: 100 x 1000 x 100 = 10 milhões de sentenças.

Além disso, ao se considerar um vocabulário grande, os conteúdos fonéticos da palavra inevitavelmente irão se sobrepor, ou seja, o número de informações redundantes irá aumentar. Assim, ao armazenar e comparar palavras inteiras serão utilizadas mais informações que o necessário. A solução mais promissora, então, é trabalhar com unidades fonéticas menores que, quando combinadas, possam formar todas as palavras do vocabulário.

Fones independentes do contexto – Uma das subpalavras mais utilizadas é o fone. Existem diversas maneiras de utilizá-lo. A mais simples utiliza os modelos independentes do contexto. Nessa abordagem, é assumido que um determinado fone em qualquer contexto é igual ao mesmo fone em um outro contexto. Por exemplo, o fone / t/ em *teto* seria idêntico ao fone /t/ em *metal*. Obviamente, isto não é verdade pois uma palavra não é uma simples concatenação de fones devido ao efeito coarticulatório. Como os articuladores não conseguem se mover instantaneamente de uma posição a outra, um fone é fortemente influenciado por seus vizinhos.

Como existem aproximadamente em torno de 50 fones [7] na língua portuguesa, eles podem ser suficientemente treinados a partir de poucas centenas de sentenças. Entretanto, Bahl et al [8] mostraram que modelos baseados em palavras possuem desempenho melhor que baseados em fones. Esses resultados

demonstram que, enquanto os sistemas baseados em palavras carecem de generalização, necessitando de um modelo para cada palavra, os sistemas baseados em fones generalizam demais.

Fones dependentes da palavra – Os sistemas baseados em fones dependentes da palavra são um meio termo entre os sistemas baseados em palavras e os sistemas baseados em fones independentes do contexto. Os parâmetros desses modelos dependem da palavra onde os fones ocorrem.

Se uma determinada palavra não aparece freqüentemente nas sentenças de treinamento, os parâmetros podem ser interpolados com os parâmetros dos sistemas baseados em fones independentes do contexto. Isto certamente diminui a necessidade de observar cada palavra e facilita novas adições ao vocabulário.

Estes sistemas possuem um desempenho superior aos sistemas baseados em palavras [9]. Uma das razões é que, enquanto os sistemas baseados em palavras possuem modelos que podem ser insuficientemente treinados, os sistemas baseados em fones dependentes da palavra podem ser interpolados.

Difones e Trifones (Fones dependentes do contexto) – Modelos de fones dependentes do contexto são semelhantes aos modelos de fones dependentes da palavra, exceto que no lugar de se considerar a palavra como contexto, consideram-se os vizinhos. Existem três tipos de modelos que levam em consideração diferentes contextos:

- modelos de difones à esquerda – que levam em consideração apenas o fone à esquerda do fone em questão;
- Modelos de difones à direita – que levam em consideração apenas o fone à direita do fone em questão; e
- Modelos de trifones – que levam em consideração os fones à esquerda e a direita do fone em questão.

Bahl et al [8] mostraram que sistemas baseados em trifones reduzem a taxa de erro em 50 % em relação aos sistemas baseados em palavras. Esta melhoria no desempenho é obtida porque os trifones captam melhor os efeitos da coarticulação existentes entre as palavras. Entretanto, existem três problemas associados ao uso de trifones: o primeiro diz respeito à quantidade de memória necessária para

armazenar o grande número existente; o segundo é que muitos deles podem ser insuficientemente treinados; e o terceiro é que embora existam muitos trifones similares, não se tira proveito deste fato.

A utilização de cada uma das unidades fonéticas apresentadas possui os seus prós e contras. Em particular, considerando as características da base de dados de vozes a ser utilizada nesta dissertação (poucas palavras com uma quantidade suficientemente grande de repetições), a utilização de palavra como unidade fonética no reconhecimento é bastante adequada.

2.1.2. Modelo de Produção da Fala

Com base nas características do mecanismo vocal humano, foi desenvolvido um modelo linear de produção da fala, em que a fonte de excitação e o aparelho vocal são considerados como dois sistemas separados. De acordo com este modelo, o sinal de voz $s(t)$ é a resposta dos sistemas de filtragem do aparelho vocal a uma ou mais fontes de excitação, e suas propriedades são especificadas ao longo do tempo em termos de características individuais da fonte e do filtro. O sinal de voz $s(t)$ é, portanto, a convolução – no tempo – da forma de onda que caracteriza a excitação $e(t)$ com uma resposta impulsional do filtro $h(t)$. A Fig 2.3 ilustra um diagrama em blocos deste modelo, já considerando sinais e sistemas discretos no tempo, ou seja, $s(n)$, $e(n)$ e $h(n)$.

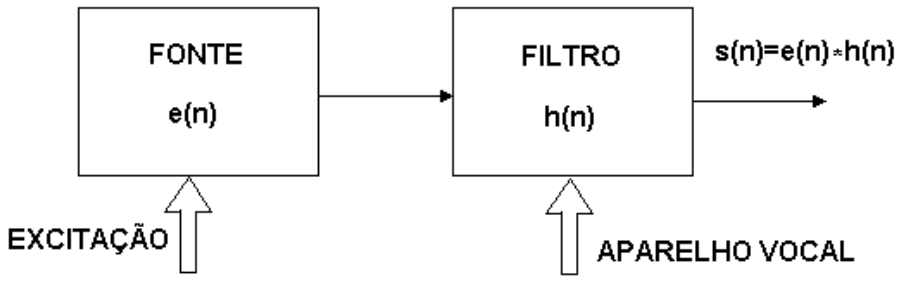

Figura 2.3 – Diagrama de blocos do modelo de produção da fala

Um modelo simples para a excitação $e(n)$ consiste em utilizar, em trechos de voz sonoros, uma seqüência de pulsos unitários. Para sons surdos, é empregada uma seqüência que represente um ruído com espectro plano.

O modelo do aparelho vocal e suas variações ao longo do tempo podem ser descritas por uma seqüência de formas estacionárias em intervalos de curta duração. Cada uma dessas formas pode ser caracterizada pela função de transferência de um filtro linear discreto, definido por seus pólos e zeros. Porém, é importante notar que a contribuição de cada zero pode ser aproximada por múltiplos pólos. Sendo assim, o modelo do aparelho vocal é usualmente caracterizado por um filtro só de pólos (filtro de síntese), contendo p pólos, cuja função de transferência é dada por

$$H(z) = \frac{\sigma}{A(z)} = \frac{\sigma}{1 + \sum_{n=1}^{p} a_n z^{-n}} \tag{2.1}$$

onde $A(z)$ é o filtro de análise ou filtro inverso de $H(z)$ multiplicado pelo fator de ganho σ e a_n são os coeficientes do filtro ou parâmetros LPC (*Linear Predictive Coefficients*). Cabe ressaltar que o filtro H(z) deve ser estável, pois representa um sinal que é limitado em energia e em potência. Isso significa que todos os seus pólos devem estar dentro da circunferência unitária.

2.2. Reconhecimento Automático de Voz através do Modelo de Markov Escondido

O reconhecimento de voz é essencialmente um problema de reconhecimento de padrões, realizado a partir de uma seqüência de parâmetros que caracterizam o sinal de voz. Como na maioria dos métodos de reconhecimento de padrões, envolve dois passos, como pode ser visto na Fig. 2.4: treinamento e reconhecimento a partir da comparação com um padrão. O "conhecimento" da fala é inserido no sistema através do procedimento de treinamento. O conceito é o seguinte: se versões suficientes do padrão a ser reconhecido (seja ele um som, uma palavra, uma frase, ...) forem incluídas no conjunto de treinamento fornecido

ao algoritmo, o procedimento de treinamento deverá ser hábil em caracterizar adequadamente as propriedades acústicas do padrão (a despeito de qualquer outro padrão presente no procedimento de treinamento). Este tipo de caracterização da fala através do treinamento é chamado de classificação de padrões, porque a máquina aprende quais propriedades acústicas são confiáveis e repetem-se ao longo de todos os sinais de treinamento do padrão. A fase de comparação nada mais é do que uma comparação direta da fala desconhecida (a fala a ser reconhecida), com cada padrão possível aprendido na fase de treinamento e a classificação desta fala desconhecida de acordo com a qualidade e o poder de acertos dos padrões [1].

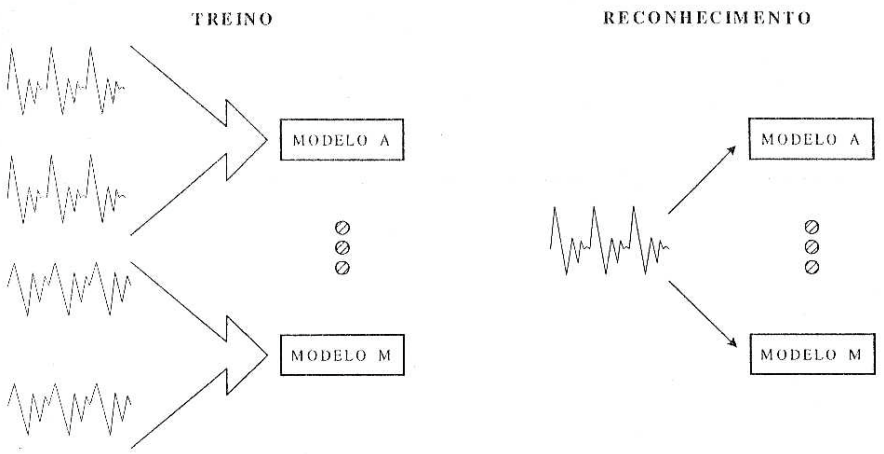

Figura 2.4 – As duas fases de um sistema de reconhecimento de fala

O Modelo de Markov escondido (HMM – *Hidden Markov Model*) é um método estatístico de reconhecimento de voz, sendo um dos mais utilizados [4], devido ao bom compromisso que apresenta entre a carga computacional e sua eficiência e flexibilidade.

Define-se como HMM, um processo de Markov que possui um número contável de estados [10], e no qual uma dada observação não é estado e sim uma função probabilística deste. Em outras palavras, o HMM é um processo estocástico que só pode ser observado através de um outro conjunto de processos estocásticos que produzem a seqüência de observações; daí o nome *hidden* (oculto, escondido). Um exemplo bastante elucidativo é o seguinte [11]: suponha que uma pessoa esteja em uma sala, tenha três moedas e atire-as numa dada seqüência – coroas (T) e caras (H). Esta sala está fechada e somente tem-se

acesso (através do telão colocado do lado de fora da sala) aos resultados TTHTHHTT ... o que será chamado de *seqüência de observações*. Os observadores no exterior da sala não sabem a seqüência em que a pessoa no interior do recinto está atirando as moedas, nem sabem se estas são "viciadas". Para avaliar como o resultado depende das características individuais e da ordem em que se atiram as moedas, suponha que se diga que a terceira moeda tem maior propensão a produzir caras (H) e todas as moedas são atiradas com igual probabilidade. Então, espera-se naturalmente, que haverá um número maior de caras do que de coroas (T) na seqüência de saída. Agora, se for dito também que a probabilidade de passar para a terceira moeda (estado), a partir da primeira ou da segunda moedas (estados), é zero, e assumindo-se que parte-se do primeiro ou do segundo estados, as caras e coroas aparecerão com igual probabilidade. Logo, fica claro que a seqüência de saída depende de três fatores: as características individuais, as probabilidades de transição entre os vários estados e o estado inicial escolhido. Estes três fatores servem para caracterizar o chamado HMM do experimento em questão.

Os HMMs, para o reconhecimento da fala, têm demonstrado muita eficiência na caracterização das propriedades temporais e espectrais do sinal de voz, e seu uso é baseado nas seguintes assertivas [12]:

1. A fala pode ser segmentada, dividida em estados, nos quais a forma de onda do sinal de voz pode ser considerada estacionária. Assume-se que a transição entre tais estados seja instantânea.
2. A probabilidade de uma certa "observação" ser gerada depende apenas do estado atual e de nenhum símbolo gerado anteriormente.

É possível usar HMMs para representar qualquer unidade da fala. Para os sistemas de reconhecimento de vocabulários pequenos, normalmente utiliza-se HMM para modelar diretamente as palavras, enquanto que para grandes vocabulários ele é utilizado para modelar sub-palavras.

A representação mais usual de um HMM é utilizada para máquinas de estados finitos, a saber: conjuntos de nós (que representam os estados) e arcos (transições permitidas entre os estados). Um tipo de modelo especialmente apropriado para reconhecimento de voz é o esquerda-direita (modelo Bakis – o

uso deste modelo se faz necessário pela restrição temporal); modelos em que, uma vez abandonado um estado, não mais se pode voltar a ele [11].

A Fig. 2.5 apresenta um exemplo de um HMM simples que poderia ser o modelo de uma palavra curta, onde se assume ser composta por três partes estacionárias. Este HMM tem uma topologia esquerda-direita simples, em que as transições só podem ocorrer para o mesmo estado ou para o estado seguinte, ou seja, não são permitidos "saltos". Neste exemplo, procurou-se explicitar alguns dos parâmetros responsáveis pela caracterização do HMM e que serão mais detalhados na seção 2.2.1: o conjunto de estados q_i, as densidades de probabilidade de emissão, $p(x_n/q_i)$, associadas a cada estado e as probabilidades de transição, $p(q_j/q_i)$, para cada transição permitida, do estado q_i para o estado q_j:

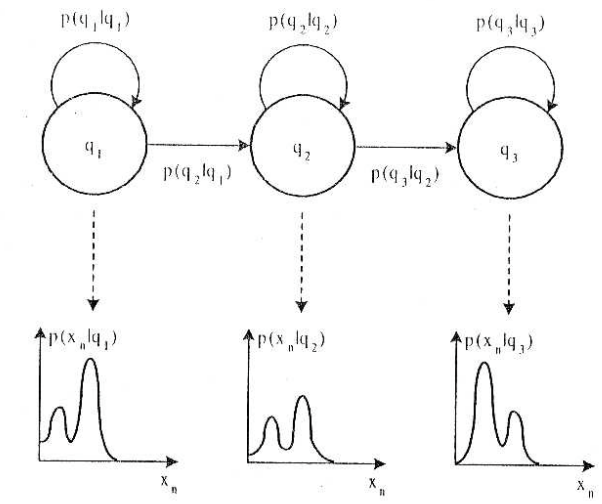

Figura 2.5 – Exemplo de HMM usado para modelar o sinal de voz

2.2.1. Elementos de um HMM

No experimento das moedas apresentado anteriormente, T(Coroa) ou H (Cara) são chamados *símbolos de observação*. Como foram consideradas moedas, só é possível a existência de dois símbolos. Para generalizar, considere-se um conjunto de N urnas, cada qual com um determinado número de bolinhas de gude, as quais possuem M diferentes cores. Dentro de cada urna as bolinhas apresentam

cores distintas. O experimento consiste em sortear bolinhas dessas urnas em uma dada seqüência; somente a seqüência de bolas sorteadas é apresentada aos observadores. As bolinhas, aqui correspondem a T ou H, e as urnas equivalem às moedas, quando comparado com o exemplo anterior.

Agora pode-se, então, definir formalmente os elementos de um HMM, usando dados do exemplo anterior, de forma a tornar mais elucidativa às exposições.

Os HMMs são caracterizados por [11, 12]:

1. $N \Rightarrow$ número total de estados (urnas) no modelo. Os estados individuais são denotados por $s = \{s_1, s_2, \ldots, s_N\}$, e o estado no tempo t é representado por q_t.

2. $M \Rightarrow$ Se o HMM for discreto, M representa o número total de símbolos de observação distintos por estado (bolinhas de gude de M diferentes cores), indicando o tamanho do alfabeto discreto. Estas observações correspondem à saída física do sistema a ser modelado e são denotados por $V = \{v_1, v_2, \ldots, v_M\}$. Se o HMM for contínuo, M representa usualmente o número de misturas da função densidade de probabilidade (fdp) associada a cada estado.

3. $A \Rightarrow$ Matriz de transição de probabilidades inter-estados. $A = \{a_{ij}\}$, onde:

$$a_{ij} = P[q_{t+1} = s_j \,|\, q_t = s_i], \qquad 1 \le i, j \le N \tag{2.1}$$

Esta matriz define a estrutura do modelo: cada um de seus elementos, a_{ij}, define a probabilidade de se passar do estado i para o estado j. Ela é uma matriz quadrada $N \times N$, mas não é obrigatoriamente cheia, a não ser quando todas as transições são permitidas.

A matriz de transição contém a informação temporal associada ao HMM e, em virtude da sua definição, seus elementos devem obedecer à seguinte restrição:

$$\sum_{j=1}^{N} a_{ij} = 1 \tag{2.2}$$

4. $B \Rightarrow$ Distribuição de probabilidade dos símbolos de observação no estado j, ou distribuição de saída associada ao estado j. $B = \{b_j(k)\}$, onde:

$$b_j(k) = P[v_k \ em \ t \mid q_t = S_j], \quad 1 \le j \le N, \ 1 \le k \le M \tag{2.3}$$

Se as observações forem contínuas será necessária a estimação das fdp que geram o conjunto de observações dado o estado j, ou seja, $b_j(\mathrm{O})$, onde O é o vetor de observações no caso contínuo. Normalmente, utiliza-se uma mistura de gaussianas e, nesses casos, a distribuição de saída pode ser descrita usando-se os seguintes parâmetros:

c_{jk} - peso da componente k da mistura

μ_{jk} - média da componente k da mistura (vetor de ordem n)

C_{jk} - matriz de covariâncias, $n \times n$, da componente k da mistura

onde $1 \le k \le M$ e n é a dimensão do vetor de observações.

Dado um vetor de observações O, sua função densidade de probabilidade gerada pela componente k da mistura de gaussianas, $b_{jk}(\mathrm{O})$, é:

$$b_{jk}(\mathrm{O}) = \frac{1}{(2\pi)^{\frac{n}{2}} \mid C_{jk} \mid^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathrm{o}-\mu_{jk})' \ c_{jk}^{-1}(\mathrm{o}-\mu_{jk})} \tag{2.4}$$

onde $|C_{jk}|$ é o determinante da matriz de covariâncias e $(\mathrm{o}-\mu_{jk})'$ é o transposto do vetor $(\mathrm{o}-\mu_{jk})$.

As densidades de cada componente da mistura são combinadas de tal forma a fornecer a função densidade de probabilidade do estado:

$$b_j(\mathrm{O}) = \sum_{k=1}^{M} c_{jk} b_{jk}(\mathrm{O}) \tag{2.5}$$

5. $\pi \Rightarrow$ Distribuição de probabilidades do estado inicial $\pi = \{\pi_i\}$, onde:

$$\pi_i = P[q_1 = S_i], \qquad 1 \le i \le N \tag{2.6}$$

Usando valores apropriados de N, M, A, B e π, o HMM pode ser utilizado como gerador de uma seqüência de observações $\mathrm{O} = \{\mathrm{O}_1 \mathrm{O}_2 \ldots \mathrm{O}_T\}$ (onde, no caso discreto, cada observação O_t é um dos símbolos de V, e T é o número de símbolos observados). Ainda tomando o experimento das urnas como exemplo, a seqüência de observações seria gerada da seguinte forma: inicialmente escolhe-se uma das urnas (de acordo com a distribuição de probabilidade inicial π); em seguida, sorteia-se uma bolinha de gude (símbolo de observação) desta urna – este instante de tempo é tomado como $t = 1$, bem como o estado e o símbolo de observação escolhidos nesse instante são denominados de s_1 e O_1, respectivamente. Depois disso, escolhe-se uma urna (pode ser a mesma ou outra diferente da escolhida em $t = 1$), de acordo com a matriz de transição de probabilidades A e novamente sorteia-se uma bolinha de gude (denominada de O_2) desta urna, com base na distribuição de probabilidade do símbolo de observação $b_j(k)$ para a tal urna (estado). O processo é repetido até o tempo $t = T$, gerando a seqüência de observações desejada.

2.2.2.
Os Três Problemas Básicos dos HMMs

A maioria das aplicações envolvendo HMMs é baseada na solução de três problemas principais [11]:

Problema 1: Dada a seqüência de observações $\mathrm{O} = \mathrm{O}_1, \mathrm{O}_2, \ldots, \mathrm{O}_T$ e o modelo $\lambda = (A, B, \pi)$, como calcular eficientemente $P(\mathrm{O} \mid \lambda)$, a probabilidade de ocorrência da seqüência de observações, dado o modelo. Em outras palavras, dado um HMM treinado, como encontrar a verossimilhança de o modelo ter produzido uma determinada seqüência de observações. Este problema diz respeito à fase de *reconhecimento*.

Problema 2: Dada a seqüência de observações $\mathrm{O} = \mathrm{O}_1, \mathrm{O}_2, \ldots, \mathrm{O}_T$ e o modelo λ, como escolher uma correspondente seqüência de estados $Q = q_1, q_2, \ldots, q_T$ que seja ótima de acordo com algum critério que melhor "justifique" as observações. Este problema diz respeito à fase de *treinamento*.

Problema 3: Como ajustar os parâmetros do modelo $\lambda = (A, B, \pi)$ de tal forma a maximizar $P(\mathrm{O} \mid \lambda)$. Em outras palavras, dada uma série de observações de treinamento para uma dada palavra, como se treina um HMM de forma que ele a represente. Este problema obviamente diz respeito à fase de *treinamento*.

Para ilustrar o emprego das soluções dos problemas 1, 2 e 3, considere o seguinte esquema para reconhecimento de palavras isoladas: para cada palavra de um vocabulário de W palavras, deseja-se projetar um HMM de N estados. Inicialmente, representa-se o sinal de voz de uma determinada palavra como uma seqüência temporal de vetores de atributos. A primeira tarefa será construir os modelos individuais das palavras. Esta tarefa é feita utilizando-se a solução do Problema 3 para otimização dos parâmetros do modelo. Para segmentar as seqüências de treinamento em estados, utiliza-se a solução do Problema 2. Finalmente, uma vez que o conjunto de HMM tenha sido projetado e otimizado, pode-se realizar o reconhecimento de uma palavra desconhecida utilizando-se a solução do Problema 1. Esta consiste no cálculo da verossimilhança de cada

modelo em relação à seqüência de observações em teste, selecionando, então, a palavra referente ao modelo que produziu a maior verossimilhança.

Na Fig 2.6 é apresentado o diagrama em blocos de um sistema de reconhecimento de palavras isoladas que emprega HMM. A Fig 2.7 apresenta, de forma detalhada, a fase de reconhecimento:

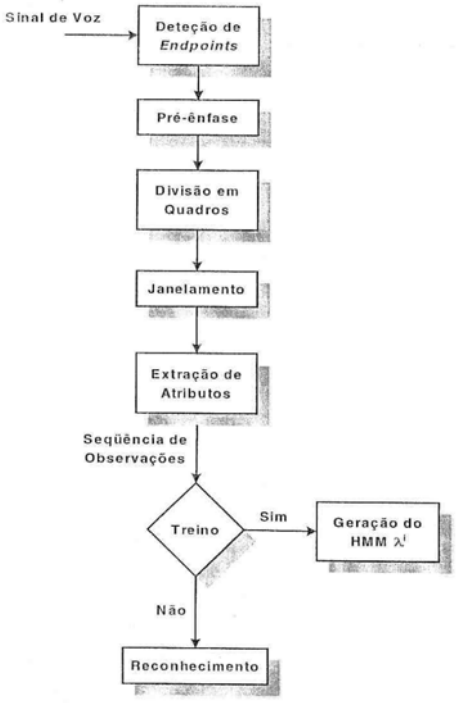

Figura 2.6 – Esquema de um sistema de reconhecimento de palavras isoladas que emprega HMM.

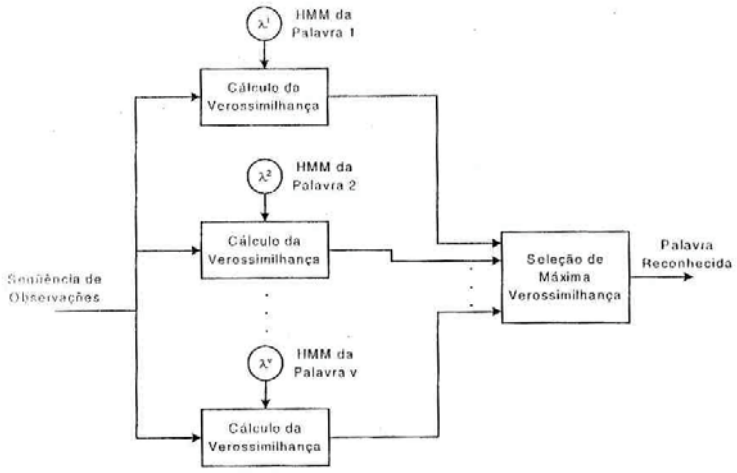

Figura 2.7 – Detalhamento da fase de reconhecimento

1. Solução para o Problema 1

A forma mais direta de se determinar $P(\mathrm{O}\,|\,\lambda)$ é obter $P(\mathrm{O}\,|\,Q,\lambda)$ para uma dada seqüência de estados $Q = q_1, q_2, \ldots, q_T$, multiplicar por $P(Q\,|\,\lambda)$, e, então, fazer a soma para todas as possíveis Qs, ou seja,

$$P(\mathrm{O}\,|\,\lambda) = \sum_{todo\ Q} P(\mathrm{O}\,|\,Q,\lambda) P(Q\,|\,\lambda) \tag{2.7}$$

Mas, supondo que as observações são estatisticamente independentes tem-se que

$$P(\mathrm{O}\,|\,Q,\lambda) = b_{q_1}(\mathrm{O}_1) b_{q_2}(\mathrm{O}_2) \ldots b_{q_T}(\mathrm{O}_T) \tag{2.8}$$

Sabendo, também, que a probabilidade de uma dada seqüência Q pode ser escrita como

$$P(Q\,|\,\lambda) = \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} \ldots a_{q_{T-1} q_T} \tag{2.9}$$

Tem-se que,

$$P(\mathrm{O}\,|\,\lambda) = \sum \pi_{q_1} b_{q_1}(\mathrm{O}_1) a_{q_1 q_2} b_{q_2}(\mathrm{O}_2) \ldots a_{q_{T-1} q_T} b_{q_T}(\mathrm{O}_T) \tag{2.10}$$

Entretanto, a carga computacional para a resolução da equação (2.10) é muito grande, até mesmo para valores pequenos de N e T. Considerando-se um caso típico, em que $N = 10$ (estados) e $T = 50$ (observações), o número de cálculos atinge o total de 10^{52}. Este valor pode ser drasticamente reduzido se forem adotadas formas convenientes de resolução. O algoritmo *Forward-backward* resolve (2.10) de forma eficiente, reduzindo o número de operações para 10^4. O algoritmo de *Viterbi* [22] fornece um resultado aproximado para a expressão, calculando os valores de probabilidade ao

longo da seqüência de estados mais provável, em vez de fazê-lo ao longo de todas as combinações de seqüências possíveis. Ele aproveita-se do fato de que, no campo do reconhecimento da fala, não interessa o valor exato de (2.10), mas uma classificação ordenada das probabilidades de cada modelo, pois isto já permite determinar qual palavra foi reconhecida.

(a) Algoritmo *Forward-backward*

Considere a variável *forward* $\alpha_t(i)$ definida como:

$$\alpha_t(i) = P(\mathrm{O}_1\mathrm{O}_2 \ldots \mathrm{O}_t, q_t = S_i \mid \lambda) \tag{2.11}$$

isto é, a probabilidade de ocorrer, até o instante t, a seqüência parcial de observações $\mathrm{O}_1\mathrm{O}_2 \ldots \mathrm{O}_t$ e o estado em t ser S_i, dado o modelo λ. A variável $\alpha_t(i)$ pode ser calculada iterativamente da seguinte maneira:

- Inicialização:

$$\alpha_1(i) = \pi_i b_i(\mathrm{O}_1), \qquad 1 \le i \le N \tag{2.12}$$

- Indução:

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij} \right] b_j(\mathrm{O}_{t+1}), \qquad 1 \le t \le T-1$$

$$1 \le j \le N \tag{2.13}$$

- Término:

$$P(\mathrm{O} \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i) \tag{2.14}$$

De forma análoga, pode-se definir a variável *backward* $\beta_t(i)$ como:

$$\beta_t(i) = P(\mathrm{O}_{t+1}\mathrm{O}_{t+2}\ldots\mathrm{O}_T \mid q_t = S_i, \lambda) \tag{2.15}$$

isto é, a probabilidade de ocorrer a seqüência parcial de observações do instante $t+1$ até o final, dado o estado S_i no tempo t e o modelo λ. Novamente, por indução, pode-se calcular $\beta_t(i)$, como se segue:

- Inicialização:

$$\beta_T(i) = 1, \qquad 1 \le i \le N \tag{2.16}$$

- Indução:

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(\mathrm{O}_{t+1}) \beta_{t+1}(j), \qquad t = T-1,\ T-2,\ldots,1$$
$$1 \le i \le N \tag{2.17}$$

- Término:

$$P(\mathrm{O} \mid \lambda) = \sum_{i=1}^{N} \pi_i b_i(\mathrm{O}_1) \beta_1(i) \tag{2.18}$$

Cabe ressaltar que para a solução do Problema 1 necessita-se apenas da parte *forward* do algoritmo *Forward-backward*. Porém, a parte *backward* também é apresentada, pois a mesma será necessária para a solução do Problema 3.

(b) Algoritmo de Reconhecimento de *Viterbi*

O algoritmo de reconhecimento de *Viterbi* é o mais rápido, pois em vez de considerar todas as combinações de transições de estados possíveis, como é feito no algoritmo *Forward-backward*, considera somente a seqüência de estados com maior probabilidade de produzir O. De forma compacta, o algoritmo de *Viterbi* pode ser descrito pelas seguintes equações:

$$v_1(j) = \pi_j b_j(\mathrm{O}_1) \tag{2.19}$$

e

$$v_{t+1}(j) = \max_{i=1,2,..,N}[v_t(i)a_{ij}]b_j(\mathrm{O}_{t+1}), \qquad t = 1,2,\ldots,T-1 \tag{2.20}$$

O valor

$$P^v(\mathrm{O}\,|\,\lambda) = \max_{i=1,2,..,N} v_T(i), \tag{2.21}$$

é adequado para representar $P(\mathrm{O}\,|\,\lambda)$, uma vez que a melhor seqüência de estados contribui com o maior peso no processo completo de cálculo de $P(\mathrm{O}\,|\,\lambda)$. Verifica-se somente a necessidade de uma classificação ordenada dessas probabilidades para cada modelo, em vez de valores exatos, pois isto já permite determinar qual palavra foi reconhecida. A equação (2.20) é semelhante à (2.13), exceto pelo fato de a soma ter sido substituída pelo operador de maximização. Esse fato se deve à observação feita anteriormente de que não se precisa do valor da probabilidade de cada modelo e sim de uma ordenação dos mesmos que permita concluir qual palavra foi reconhecida.

2. Solução para o Problema 2

Diferentemente do Problema 1, para o qual uma solução exata podia ser dada, existem diversas maneiras de se resolver o segundo problema, que se constitui em achar a seqüência de estados "ótima" associada a uma dada seqüência de observações. Aqui, novamente, o mais utilizado é o algoritmo de *Viterbi*.

Para recuperar a seqüência de estados mais provável, deve-se armazenar $\psi_{t+1}(j)$, onde $\psi_{t+1}(j)$ é o estado mais provável no tempo *t* dado o estado j no tempo $t+1$. Daí, $\psi_{t+1}(j) = i^v$, onde i^v é o número do estado que maximiza o lado direito de (2.20). Desta forma,

$$q_T^v = \arg\max_{i=1,2,\ldots,N}[v_T(i)] \tag{2.22}$$

é o último estado da melhor seqüência de Viterbi,

$$Q^v = \{q_1^v, q_2^v, \ldots, q_T^v\} \tag{2.23}$$

a qual é recuperada após sucessivas aplicações de ψ. Já o estado anterior a q_T^v, designado por q_{T-1}^v, é obtido a partir de $\psi_T(q_T^v)$. Este processo é repetido de forma sucessiva, obtendo-se os outros estados e, assim, a seqüência de estados mais provável, que é a solução do Problema 2.

3. Solução para o Problema 3

O terceiro e mais difícil problema inerente aos HMMs é determinar um método para ajustar os parâmetros do modelo $\lambda = (A, B, \pi)$, de forma a maximizar a probabilidade da seqüência de observações dado o modelo $P(\mathrm{O}|\lambda)$. Para qualquer seqüência finita de treinamento, não existe uma solução ótima para estimar os parâmetros do modelo. Pode-se, entretanto, escolher λ de modo que $P(\mathrm{O}|\lambda)$ seja localmente maximizado. A técnica mais utilizada para isso é o *método de reestimação de Baum-Welch*.

Para a descrição deste método define-se inicialmente $\xi_t(i,j)$, a probabilidade de estar no estado S_i no instante t, e no estado S_j no instante $t+1$, dados o modelo e a seqüência de observações:

$$\xi_t(i,j) = P(q_t = S_i, q_{t+1} = S_j \mid \mathrm{O}, \lambda) \tag{2.24}$$

A partir das variáveis *forward,* $\alpha_t(i)$, e *backward,* $\beta_t(i)$, definidas, respectivamente, em (2.11) e (2.15), pode-se reescrever $\xi_t(i,j)$ na seguinte forma:

$$\xi_t(i,j) = \frac{\alpha_t(i)a_{ij}b_j(\mathrm{O}_{t+1})\beta_{t+1}(j)}{P(\mathrm{O}\mid\lambda)} = \frac{\alpha_t(i)a_{ij}b_j(\mathrm{O}_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b_j(\mathrm{O}_{t+1})\beta_{t+1}(j)} \tag{2.25}$$

Definindo-se $\gamma_t(i)$ como a probabilidade de estar no estado S_i no instante t, dada a seqüência de observações e o modelo,

$$\gamma_t(i) = P(q_t = S_i \mid \mathrm{O}, \lambda) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N}\alpha_t(i)\beta_t(i)} \tag{2.26}$$

Pode-se relacionar $\xi_t(i,j)$ com $\gamma_t(i)$ da seguinte maneira:

$$\gamma_t(i) = \sum_{j=1}^{N}\xi_t(i,j) \tag{2.27}$$

Considerando-se o somatório $\gamma_t(i)$ ao longo do tempo, obtêm-se uma quantidade que pode-se interpretar como o número de vezes que S_i é visitado, ou equivalentemente, o número esperado de transições a partir de S_i. Similarmente, pode-se considerar que o somatório de $\xi_t(i,j)$ ao longo do tempo é o número esperado de transições de S_i para S_j.

Explicitamente:

$$\sum_{t=1}^{T-1}\gamma_t(i) \Rightarrow \text{Número esperado de transições a partir de } S_i \qquad (2.28)$$

$$\sum_{t=1}^{T-1}\xi_t(i,j) \Rightarrow \text{ Número esperado de transições de } S_i \text{ para } S_j \qquad (2.29)$$

Utilizando (2.27), (2.28), (2.29) e o conceito de freqüência relativa de ocorrência, pode-se deduzir as equações para reestimação dos parâmetros do modelo $\lambda = (A,B,\pi)$, onde A é a matriz de transição de probabilidades inter-estados, $A = \{a_{ij}\}$, em que cada um de seus elementos, a_{ij}, define a probabilidade de se passar do estado i para o estado j, B é a distribuição de probabilidade dos símbolos de observação no estado j, ou distribuição de saída associada ao estado j, $B = \{b_j(k)\}$, em que $b_j(k)$ é a probabilidade do símbolo de observação discreto k no estado j e π é a distribuição de probabilidades do estado inicial $\pi = \{\pi_i\}$, em que π_i é a probabilidade do estado i ser o inicial :

$$\overline{\pi}_i = \gamma_1(i) \qquad (2.30)$$

$$\overline{a}_{ij} = \frac{\sum_{t=1}^{T-1}\xi_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)} \qquad (2.31)$$

onde os elementos barrados $\overline{\pi}_i$, $\overline{a}_{ij}$ e $\overline{b}_j(k)$ são os mesmos definidos anteriormente, porém obtidos pelo processo de reestimação.

- Para o caso discreto, tem-se:

$$\overline{b}_j(k) = \frac{\sum_{t=1;O_t=v_k}^{T}\gamma_t(j)}{\sum_{t=1}^{T}\gamma_t(j)} \qquad (2.32)$$

- Para o caso contínuo, com mistura de gaussianas, tem-se:

$$\overline{b}_j(\mathrm{O}_t) = \sum_{k=1}^{M} \overline{c}_{jk} \overline{b}_{jk}(\mathrm{O}_t) \tag{2.33}$$

onde

$$\overline{b}_{jk}(\mathrm{O}_t) = \frac{1}{(2\pi)^{\frac{n}{2}} \mid \overline{C}_{jk} \mid^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathrm{O}_t - \overline{\mu}_{jk})' \, c_{jk}^{-1}(\mathrm{O}_t - \overline{\mu}_{jk})} \tag{2.34}$$

e

$$\overline{c}_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j,k)}{\sum_{t=1}^{T} \sum_{k=1}^{M} \gamma_t(j,k)} \tag{2.35}$$

onde em (2.34),

$$\overline{\mu}_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j,k) \cdot \mathrm{O}_t}{\sum_{t=1}^{T} \gamma_t(j,k)} \tag{2.36}$$

e

$$\overline{C}_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j,k) \cdot (\mathrm{O}_t - \overline{\mu}_{jk})(\mathrm{O}_t - \overline{\mu}_{jk})'}{\sum_{t=1}^{T} \gamma_t(j,k)} \tag{2.37}$$

Sendo que $\gamma_t(j,k)$, em (2.35) e (2.36), é dado por

$$\gamma_t(j,k) = \left[\frac{\alpha_t(j)\beta_t(j)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)} \right] \left[\frac{\overline{b}_{jk}(\mathrm{O}_t)}{\overline{b}_j(\mathrm{O}_t)} \right] \tag{2.38}$$

que denota a probabilidade de estando no estado j, no instante t, a componente k da mistura estar contribuindo para O_t.

Considerando que $\lambda = (A,B,\pi)$ o modelo atual e $\overline{\lambda} = (\overline{A},\overline{B},\overline{\pi})$, o modelo obtido através dessa reestimação, foi provado por Baum et al [13] que $P(\mathrm{O}\,|\,\overline{\lambda}) \geq P(\mathrm{O}\,|\,\lambda)$, encontrando-se um novo modelo $\overline{\lambda}$, a partir do qual é mais provável que a seqüência de observações tenha sido produzida.

Usando iterativamente $\overline{\lambda}$ no lugar de λ e repetindo-se os cálculos de reestimação, pode-se melhorar a probabilidade de O ser observado, dado o modelo, até que seja atingido algum limite. O resultado final deste processo de reestimação é chamado de *estimativa de máxima verossimilhança do HMM*.

2.3. Conclusão

Neste capítulo foi feita a apresentação dos sons da fala e suas classificações, das unidades fonéticas e dos conceitos de reconhecimento automático de voz.

No capítulo seguinte será dada continuidade a este estudo, tendo em vista o cenário de reconhecimento de voz distribuído. Serão apresentadas as técnicas de reconhecimento de voz distribuído, os parâmetros mais utilizados, bem como a estrutura do reconhecedor de voz distribuído que é o objeto de estudo desta dissertação.