2
Fundamentos do Reconhecimento de Voz

O processo de reconhecimento de voz requer, como todas as atividade de investigação, os respectivos fundamentos teóricos e práticos para sua realização.

Porém, falar deste processo de reconhecimento em geral implica ter uma quantidade de informação muito ampla, tornando-se uma tarefa difícil devido ao fato da complexidade matemática, informática, linguística, etc.

Este capítulo ressalta as características principais do processo de reconhecimento de voz, apresentando sua estrutura básica e analisando os problemas de execução mais frequentes no processo da fala.

2.1
A comunicação oral

A comunicação oral é o meio através do qual compartilha-se informação diariamente para nos desenvolver. No caso mais simples, a comunicação oral se dá entre duas pessoas: um emissor e um receptor. A Fig. 2.1 apresenta o fluxo que origina este processo de comunicação.

Na mente do emissor, por meio de uma sequência de palavras, cria-se uma mensagem e uma vez estabelecida, o cérebro ativa os músculos vocais através de impulsos nervosos, convertendo-os num discurso de palavra transmitido através do sinal sonoro. Esse sinal sonoro é recebido pelo receptor, que faz o processo inverso: o movimento da membrana basilar no ouvido do receptor é convertido num impulso elétrico, o qual é transmitido ao cérebro mediante os nervos auditivos [3].

A dificuldade começa quando o ambiente influencia o conteúdo do sinal, fazendo com que a mensagem que o emissor quer transmitir com precisão não seja realmente a que o receptor vai escutar, provocando com isso um conflito de interesse.

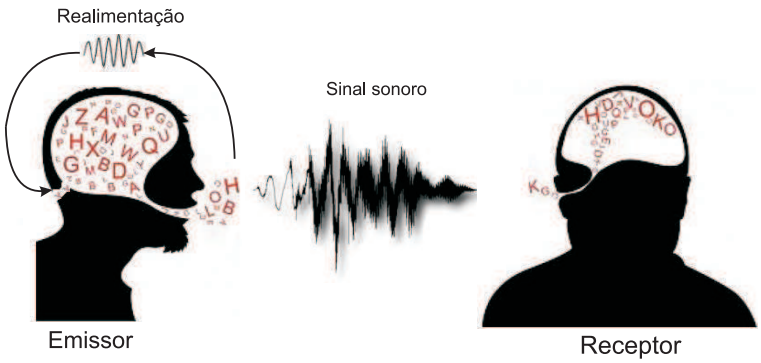

Figura 2.1: Processo de comunicação oral

2.2 O sinal de voz

A voz é uma característica que só os humanos possuem, e baseia-se na produção de sons articulados, originando assim uma linguagem que é a fonte da comunicação. A voz não só transmite informação léxica, mas também expressa emoções, como dor e alegria, através de sua entonação.

Este sinal de voz propaga-se através de uma onda de pressão acústica com limites de frequência entre 20Hz e 20KHz, chamados limites de audição [4] e se produz quando uma coluna de ar proveniente dos pulmões excita o conduto vocal, o qual se comporta como uma cavidade ressonante, convertendo-se assim, numa onda sonora que pode estimular o ouvido humano para ser percebida no cérebro como uma sensação acústica.

Nesta seção faz-se uma descrição do aparelho fonador, os tipos de som da fala, os problemas de estacionariedade do sinal e a importância do modelo de linguagem.

2.2.1 Anatomia do aparelho fonador

O aparelho fonador tem uma grande quantidade de elementos físicos, os quais intervêm na geração da voz e são divididos em três grupos de órgãos diferenciados:

- órgãos de respiração: *pulmões, brônquios e tráquea.*
- órgãos de fonação: *laringe e cordas vocais* .
- órgãos de articulação: *paladar, língua , lábios e glote.*

Estes órgãos, por sua vez, fazem parte do aparelho respiratório e também alguns do aparelho digestivo. Um diagrama esquemático da estrutura do aparelho fonador é mostrado na Fig. 2.2.

A produção da voz inicia-se com o fluxo de ar que flui desde os pulmões, impulsionado pelo diafragma, atravessando a laringe, onde se encontram dois pequenos tendões ou membranas chamados cordas vocais.

Estas cordas são consideradas um dos principais elementos para a geração de voz, que se forçam e vibram ao passo do ar. A abertura entre as cordas vocais denomina-se glote.

Segundo [5] a articulação glótica é o que faz possível distinguir variações muito sutis de ironia, dor, alegria, tristeza medo ou vergonha.

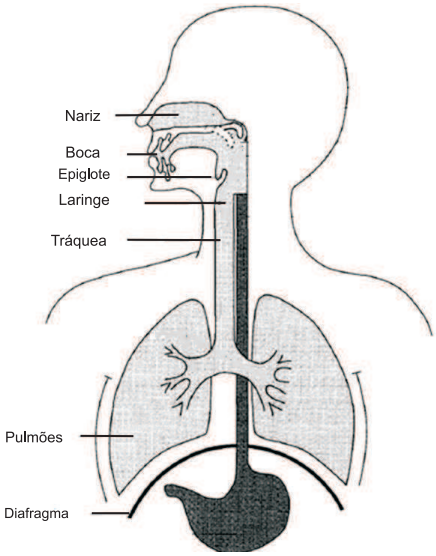

Figura 2.2: Estrutura do aparelho fonador

2.2.2 Os sons da fala

Ao falar, emitem-se sons e sua interpretação gráfica são as letras. Na atualidade existem muitos alfabetos e é quase impossível encontrar sons que se representem exclusivamente com uma letra. Por exemplo, no espanhol, as letras 'b' e 'v' têm o mesmo som. Este som é um *fonema*.

Um *fonema* é a unidade sonora mais simples da língua, e divide-se em vogais, semivogais e consonantes. Isolado, o fonema não representa significação própria, entretanto estabelece contraste de significado para diferenciar palavras.

Cada língua possui um número distinto de fonemas. Para explicar com um exemplo, considerem-se as frases abaixo:

"Ontem comi um pão no café da manhã"
"Ontem comi um cão no café da manhã"

Graficamente, observa-se que a única diferença entre elas é a oposição da letra **p** à letra **c**. Lendo as frases, percebe-se que a essa diferença gráfica corresponde uma diferença sonora que faz mudar o sentido da palavra, apenas mudando um de seus elementos básicos

Os sons da fala, dependendo da presença ou ausência de vibração das cordas vocais (Fig.2.3), podem se dividir em dois tipos: **sonoros** ou **surdos**, apresentando características atenuadamente distintas [6].

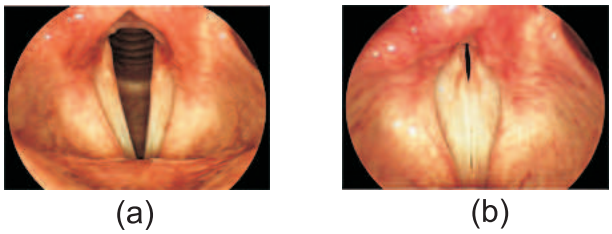

Figura 2.3: Cordas vocais **(a)** Glótis aberta e cordas vocais separadas gerando sons surdos. **(b)** Glótis fechada e cordas vocais em vibração gerando sons sonoros

Diz-se que o som é sonoro quando a corrente de ar que vem dos pulmões encontra as cordas vocais fechadas, fazendo-as vibrar. Por exemplo, na palavra 'Bato', percebe-se este som sonoro devido ao fonema /**B**/.

E o som é surdo quando a corrente de ar que vem dos pulmões encontra as cordas vocais relaxadas (abertas), não ocorrendo vibração, por exemplo na palavra 'Prato' percebe-se este som surdo devido ao fonema /**P**/.

Além de dividir os sons em sonoros e surdos, eles podem ser classificados também como: vogais, nasais, fricativos, oclusivos e líquidos.

– **Vogais**: São sons produzidos sem obstáculos para a passagem de ar livremente pela boca, desde o pulmão. Sua emissão é independente de outro fonema, por isso constitui a base da sílaba.

  Estes sons produzem-se a partir do diferentes posicionamentos dos músculos da boca, constituídos pela língua, pelos lábios e pelo véu palatino. A Fig 2.4 dá uma ideia aproximada dos sons vogais.

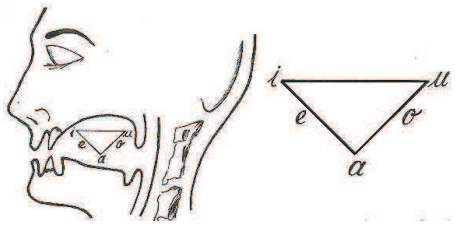

Figura 2.4: Triângulo vogais de Hellwag.

– **Nasais**: Ao gerar um som nasal, por exemplo, o /nh/, ocorre uma maior distribuição de energia sonora no trato vocal, já que o ar sonorizado será dirigido para a cavidade nasal, comportando-se como um ressoador em paralelo com o aparelho vocal, reduzindo a tensão da laringe e da faringe.

– **Fricativos**: Produzem-se quando se faz um estreitamento entre dois órgãos articulatórios produzindo a fricção. Estes sons diferenciam-se dos outros por que possuem altas frequências, não têm estrutura periódica e apresentam turbulências que têm certa similaridade com o sinal do ruído.

– **Oclusivos**: existem sons oclusivos sonoros e surdos. Caracterizam-se por haver uma obstrução total da corrente de ar seguida de uma liberação repentina dela (explosão). Após esta explosão, os oclusivos assemelham-se aos fricativos de corta duração.

– **Líquidos**: também conhecidos como laterais, têm pouca resistência de ar e são sempre sonoros. A língua obstrui o centro da boca deixando que o ar circule pelos lados.

A Fig. 2.5 ilustra a forma de onda da palavra em inglês 'she' que é formada pelo fonema /sh/ que é fricativo e surdo e o fonema /ix/ que é vogal segundo [7].

Uma análise desta forma de onda mostra que os sons sonoros, no caso o fonema /ix/, têm uma alta energia devido à excitação das cordas vocais e têm conteúdo frequencial na faixa dos 300 Hz a 4000 Hz para voz em telefonia, sendo quasiperiódicas Fig.2.5 **(b)**.

Por outro lado, a Fig.2.5**(a)** refere-se a um som fricativo ou surdo, do fonema /sh/, o qual apresenta um comportamento aleatório em forma de ruído branco, tendo uma alta densidade de cruzamento por zero.

2.2.3
Estacionariedade do sinal de voz

O sinal de voz é especial, já que codifica mediante sons a linguagem falada. Estes sons podem ser considerados aleatórios, representados por uma série de amostras temporais e caracterizados mediante funções densidade de probabilidade. Por isso o sinal de voz é considerado como um processo estocástico.

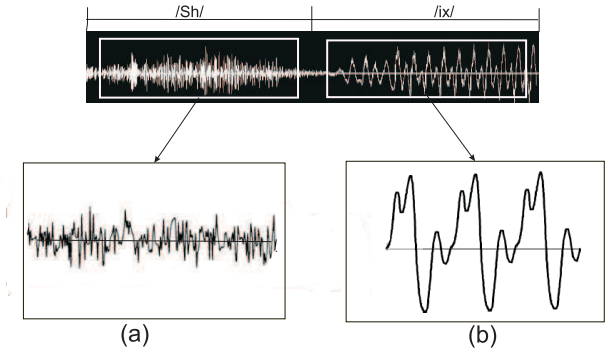

Figura 2.5: Formas de onda dos sons sonoros e surdos **(a)** fonema /sh/ **(b)** fonema /ix/.

Entretanto, este processo é não estacionário, ou seja, que o sinal varia no tempo de acordo com os sons emitidos pelo aparelho fonador, mudando suas propriedades estatísticas. Porém, pode-se assumir que realizando estimações a curto prazo, conseguem-se blocos com propriedades estatisticamente constantes, quasiestacionárias, permitindo analisar e processar o sinal de voz como um sinal estacionário.

Para atingir a quasiestacionariedade do sinal de voz, o tamanho da janela temporal de estimação deverá ser escolhido convenientemente Fig.2.6, de forma que seja o suficientemente curta para que apresente as características espectrais instantâneas da estimação e o suficientemente longa para minimizar a variância na estimação dos parâmetros, e minimizar a taxa de informação a enviar.

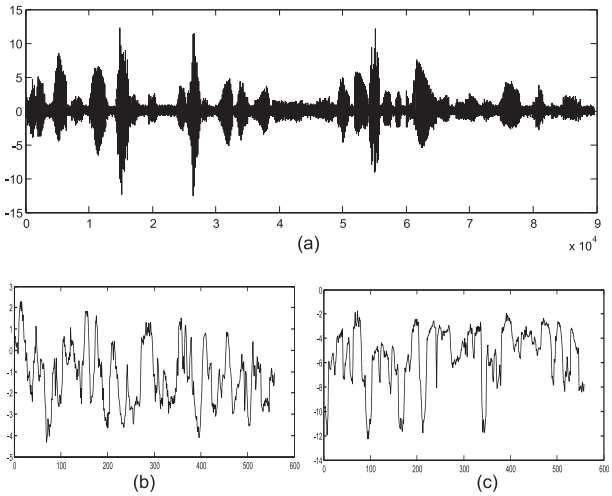

Figura 2.6: **(a)** Aspecto de sinal de voz no domínio do tempo;**(b) e (c)** análise com janelas de 25 ms, sinal quasiestacionario.

Em geral, blocos de 20 a 30 ms são adequados para a maior parte das aplicações. Em todos os casos, realiza-se uma superposição entre blocos adjacentes, garantindo de forma favorável a estacionariedade dos mesmos e limitando ainda mais as variações do aparelho fonador.

2.2.4
Modelo da linguagem

O modelo de linguagem, ou também chamado gramática, utiliza o contexto das palavras e a informação da frequência com que elas são pronunciadas, com o fim de encontrar opções prováveis que indiquem quais palavras, têm mais chances de vir antes ou depois de uma outra.

Por exemplo, existem duas ondas sonoras com sons quase iguais "norte" e "morte". Entretanto se antes da palavra encontra-se a frase "no pólo..." o modelo da linguagem determina que "norte" é a palavra certa. Desta forma pode-se dizer que as restrições impostas pelo modelo de linguagem podem melhorar consideravelmente o rendimento do reconhecedor, reduzindo significativamente o espaço de busca da frase correta.

Em geral, o modelo da linguagem terá a tarefa de estimar a probabilidade de uma palavra $P(W)$ em uma sentença, dadas todas as palavras que a procedem $W_1, W_2...., W_n$.

Usando as regras elementares da teoria da probabilidade, pode-se expressar $P(W)$ da seguinte forma:

$$P(W) = \prod_{i=1}^{n} P(w_i|w_1, w_2, ..., w_{i-1}) \tag{2-1}$$

onde $P(w_i|w_1, w_2, ..., w_{i-1})$ é a probabilidade de que w_i seja escolhida depois da sequência de palavras $(w_1, w_2, ..., w_{i-1})$.

Uma forma mais usada e simples, porém efetiva de se obter estas probabilidades, é com a utilização de *n-gramas*, na qual a probabilidade de cada palavra em uma sentença depende apenas das $n-1$ palavras anteriores a ela. Por exemplo (2-1) pode se decompor como:

$$P(w_1)P(w_2|w_1)P(w_3|w_1w_2)...P(w_n|w_1...w_{n-1})$$

Com n=2 tem-se o bigrama

$$P(W) = P(w_1)P(w_2|w_1)P(w_3|w_2)...P(w_n|w_{n-1})$$

Com n=3 tem-se o trigrama

$$P(W) = P(w_{1))}P(w_2|w_1)P(w_3|w_2w_1)...P(w_n|w_{n-2}...w_{n-1})$$

Este tipo de decomposição é feito a fim de estabelecer as possíveis combinações de palavras ou unidades acústicas a serem reconhecidas. é por isto que neste modelo têm-se que ter em conta tanto a sintaxes como a gramatica do linguagem.

2.3 Problemas do reconhecimento de voz

Antes de falar dos problemas, deve se responder, *O que é o reconhecimento de voz?*

O reconhecimento de voz é uma parte da inteligência artificial que tem como objetivo permitir a comunicação falada entre seres humanos e computadores eletrônicos, através de processos de classificação de sinais em sequências de padrões a fim de processar a mensagem contida na onda acústica.

A naturalidade com que os seres humanos se comunicam faz pensar que o reconhecimento de voz é uma tarefa simples. Porém, ele requer um processo complexo devido ao número de considerações a ter em conta para adequar o sinal e extrair suas características de uma forma fácil e eficiente.

Além disso, o reconhecimento de voz requer conhecimentos de áreas como Psicologia, Fisiologia, Acústica, Processamento de sinal, Teoria da informação, Linguística, Informática, etc. [8].

Escolher o nível de reconhecimento segundo a necessidade do sistema é uma das dificuldades principais do reconhecimento automático de voz, já que ele pode ser caracterizado por vários parâmetros, tais como palavras isoladas, palavras conectadas e de fala contínua, o que aumenta o nível de complexidade do reconhecedor, já que tem que delimitar palavras e frases.

Na Tabela 2.1 [9] apresenta-se uma visão global das variáveis que definem um sistema de reconhecimento automático de voz e o range de valores que podem ter.

Parâmetro	**Variedade**
Forma de falar	Palavra isolada $\longleftrightarrow$ Fala contínua
Estilo de fala	Texto lido $\longleftrightarrow$ Fala espontânea
Adaptação	Dependente de locutor $\longleftrightarrow$ Independente de locutor
Tamanho do vocabulário	Pequeno ($<$ 20 palavras) $\longleftrightarrow$ Grande ($>$20.000 palavras)
Modelo da linguagem	Estados finitos $\longleftrightarrow$ Dependentes de contexto
Perplexidade	Pequena ($<$10) $\longleftrightarrow$ Grande ($>$100)
SNR	Alta ($>$30) $\longleftrightarrow$ Baixa($<$10)
Transductor	Microfone de eliminação de eco $\longleftrightarrow$ Telefone

Tabela 2.1: Parâmetros típicos que caracterizam o sistema de reconhecimento de voz.

Além da interdisciplinaridade e às limitações dos sistemas de reconhecimento, há outros aspectos da fala que tornam o reconhecimento de voz uma

tarefa difícil [10], como a variabilidade, o ruído, a continuidade, a redundância, e a quantidade de dados a processar.

2.3.1 Variabilidade

A variabilidade do sinal de voz depende tanto de aspectos relacionados a fatores internos ao fenômeno de produção de voz como de fatores externos ao mesmo.

Entre os fatores internos à produção da voz, destacam-se:

- *Variabilidade intralocutor*, que vai depender do estado emocional, do contexto da conversação, da inclusão de ruídos (respiração, sons de admiração, dúvida, etc.).
- *Variabilidade interlocutor*, que vai depender dos distintos sotaques e forma de falar, já que cada locutor apresenta características diferentes.

Para os fatores externos à produção da voz, tem-se:

- Variabilidade na cadeia de conversão e transmissão do sinal acústico, devido às diferenças entre caraterísticas de microfones, linhas de transmissão, etc.

2.3.2 Continuidade

No processo natural de fala não existe pausa nem separação de forma automática entre fones, sílabas, até mesmo entre palavras que compõem uma frase, pois, devido ao efeito de coarticulação, os elementos são influenciados mutuamente. A separação destes elementos é feita pelo ser humano devido ao seu conhecimento prévio da lingua, o que constitui uma das principais diferenças entre o reconhecimento automático de fala e o tratamento da escrita.

2.3.3 O Ruído

Os reconhecedores em ambientes limpos fornecem resultados excepcionais. Porém, em ambientes reais, as taxas de reconhecimento diminuem por causa do ruído que degrada a qualidade da voz e altera a estatística dos vetores que a representam.

2.3.4
Redundância

A redundância contém os dados adicionais que permitem identificar o locutor, seu ambiente, seu estado emocional, seu sotaque, etc.

O maior problema do reconhecimento de voz é procurar a informação relativa à mensagem. É por isso que um sistema de reconhecimento tem que focar na extração dos parâmetros que caracterizem o tipo de informação útil para este processo. Porém, não estão definidas regras que descrevam os diferentes níveis nos quais se apresenta a informação, dificultando a análise da voz.

2.3.5
Quantidade de dados a processar

Devido aos problemas citados, o sistema de reconhecimento precisa guardar e processar uma quantidade de dados considerável para estabelecer as características que fazem a diferença entre as distintas unidades de reconhecimento.

Assim, precisa-se de sistemas com maior capacidade de armazenamento e velocidade para desenvolver aplicações que funcionem bem em condições reais.

2.4
Estrutura dos sistemas de reconhecimento de voz

Os sistemas atuais de reconhecimento voz contínua baseiam-se fundamentalmente em princípios de reconhecimento estatístico de padrões, onde os sinais acústicos são transformados em uma sequência de símbolos e saí analisados e estruturados em unidades de sub-palavras (por exemplo, fones), que os representem com a menor perda de informação possível.

A ideia principal desta estrutura é dividir o processo de reconhecimento em etapas menores, visando reduzir a carga computacional e o tempo de processamento.

A Fig. 2.7 representa a estrutura geral de um sistema de reconhecimento de voz, que consta de três blocos principais, que serão analisados a seguir.

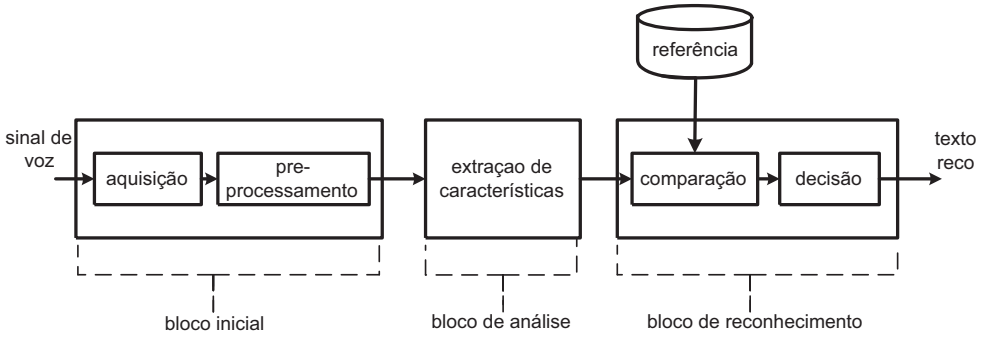

Figura 2.7: Diagrama de blocos geral de um sistema de reconhecimento.

2.4.1 Aquisição e pré-processamento do sinal de voz

A primeira ação que se tem que executar é a aquisição do sinal de voz de entrada ao sistema. Este sinal de voz é transmitido através de ondas de pressão, as quais, antes de passar para o pré-processamento, exigirão sua conversão para onda elétrica, o que será realizado através de microfones e amplificadores, originando um sinal elétrico analógico.

Uma vez amostrado, quantizado e codificado o sinal de voz procedente do microfone, ocorrera o pré-processamento, que visa eliminar o ruído [11], e deixar o sinal de voz tão limpo como o bloco de extração requeira, ou seja, depurar o sinal para robustecer o processo de codificação e eliminar componentes não desejadas, realizando assim um escalado do sinal para reduzir sua margem dinâmica e evitar possíveis erros na quantificação.

O pré-processamento desse sinal inclui as seguintes etapas: pre-ênfase, segmentação, janelamento e transformada de Fourier. A seguir se detalharão os aspectos mais importantes de cada uma delas.

Pre-ênfase

Prévio à segmentação do sinal, é aplicado um filtro digital passa-alta de primeira ordem, a fim de compensar os efeito dos pulsos glotais [12] e ressaltar as frequências dos formantes, esse procedimento justifica-se por duas razões:

- Evitar a perda de dados durante o processo de segmentação, já que a maior parte da informação está contida as frequências baixas.
- Remover a componente DC do sinal, aplainando-o espectralmente.

Sua função de transferência é dada por:

$$H(z) = 1 - \alpha z^{-1} \qquad 0 \leq \alpha \leq 1 \qquad (2\text{-}2)$$

onde α determina a frequência de corte, com valores tipicamente variando ente 0.95 e 0.98.

Segmentação

No reconhecimento de sinais de voz, é preciso determinar com precisão os pontos de início e fim das palavras, quer dizer, distinguir as partes do sinal que têm informação de voz daquelas que não têm, visando reduzir o tempo de cálculo.

Portanto, o sinal de voz é segmentado em quadros relativamente pequenos, nos quais assumem-se características de quasiestacionariedade [13].

Tendo em conta a duração dos fones, o tamanho do quadro geralmente é de 20 a 30ms, com um deslocamento típico de 10 ms entre quadros. Isso impede a perda de representação de um segmento.

Uma vez segmentado o sinal, o quadro é armazenado como um vetor de atributos para o posterior processamento.

Para calcular o número N de amostras que compõem cada quadro, multiplica-se a duração do segmento L_t, pela frequência de amostragem F_s como se apresenta na equação (2-3).

$$N = F_s(amostras/segundo) * L_t(segundos) \qquad (2\text{-}3)$$

Janelamento

Segmentar o sinal de voz traz o problema de descontinuidade ao inicio e ao final de cada quadro, devido ao fato de cada um começar e terminar bruscamente.

É necessário então diminuir este efeito, multiplicando cada quadro por uma janela que seja adequada, visando suavizar as bordas do quadro até chegar a zero, e realçando a parte central para acentuar as propriedades características do segmento, como amostra a Fig. 2.8.

No reconhecimento de voz, existem diferentes tipos de janelas, no entanto, a mais utilizada é a janela Hamming [14].

Matematicamente, a janela Hamming está representada pela seguinte equação:

$$w(n) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\Pi n}{N-1}\right) & \text{para} \quad 0 \leq n \leq N-1, \\ 0 \quad \text{para} \quad \text{caso contrário.} \end{cases} \tag{2-4}$$

Desta forma, o sinal de voz segmentada e sem perdas de de informação devido à descontinuidade entre quadros é definido pela multiplicação das amostras de cada quadro pela janela de Hamming 2-5

$$x[n] = N * w(n) \tag{2-5}$$

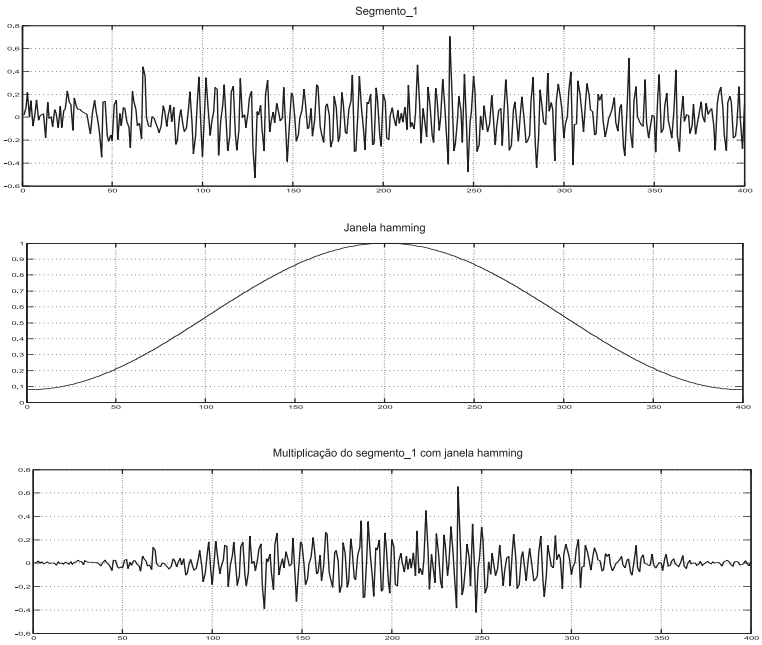

Figura 2.8: Segmento janelado com hamming

Transformada de Fourier

Enquanto uma função no domínio do tempo indica como a amplitude do sinal muda no tempo, sua representação no domínio da frequência permite saber quantas vezes essas mudanças ocorrem.

Para o reconhecimento de voz, o sinal é transformado em suas componentes frequenciais, conseguindo assim diferenciar as vozes de diferentes locutores e determinar as palavras que foram ditas [15].

Devido ao fato do sinal de voz não ser estacionário extrai-se o espectro de potência de cada um dos quadros janelados usando a transformada discreta de Fourier (DFT) [16], matematicamente representada pela equação

$$X(k) = \sum_{k=0}^{N-1} x[n] e^{\frac{-j2\Pi kn}{N}} \quad (2\text{-}6)$$

onde N é o total de amostras do quadro.

2.4.2 Extração de características

Basicamente o objetivo deste bloco é representar o sinal de voz de forma adequada para o reconhecedor através de conjuntos de vetores de n componentes que representam o espectro de cada segmento de voz.

A partir deste conjunto de vetores, obtêm-se consegue uma compressão do sinal, suprimindo a informação irrelevante para sua posterior análise fonética dos dados pré-processados. Esse conjunto de vetores pode ser representado de diversas formas, utilizando parâmetros que caracterizem diferentes aspectos do sinal, e cuja interpretação física seja imediata. Algumas delas serão apresentadas no capítulo 3.

Deve-se considerar que o número de parâmetros tem que ser pequeno para não saturar a base de dados, já que quanto mais parâmetros tenha o vetor, menos confiáveis serão os resultados e mais custosa a implementação.

2.4.3 Comparação e Decisão

O sistema de reconhecimento em seu bloco final compõe-se de três subestruturas fundamentais que visam misturar e comparar os vetores de características com os padrões de referência. Estas referências representam os diferentes objetos a reconhecer, que podem ser sílabas, fonemas ou palavras dependendo do modelo de linguagem e da arquitetura do reconhecedor.

Depois de obter o vetor de características e os padrões de referência, se faz a comparação entre as referências e as frases a reconhecer. Esta forma de comparação está ligada ao projeto do sistema de reconhecimento, o qual precisa estabelecer um modelo eficiente para identificar uma palavra entre várias. Uma das técnicas mais utilizadas nos últimos tempos é a mistura de gaussianas para a representação e construção dos modelos de classificação [17].

Por último, são calculadas as probabilidades de que, dado um dos modelos de unidade fonética, representado por um HMM correspondentes às palavras contidas no dicionário, a observação de entrada tenha sido produzida por esse modelo.

Escolhendo-se por fim a palavra mais provável, gera-se à saída o texto reconhecido.

2.5 Modelos ocultos de Markov - HMM aplicados ao reconhecimento de voz contínua

As técnicas mais utilizadas e eficazes para o reconhecimento automático de fala até agora têm sido os Modelos Ocultos de Markov (HMM)[18] [19].

O sucesso destas estruturas deve-se, principalmente, à sua capacidade de modelar tanto as variabilidades acústicas como temporais do sinal de fala, e também por permitir a construção hierárquica dos modelos acústicos das sentenças .

A introdução dos HMMs no campo da voz é usualmente creditada aos trabalhos independentes na Carnegie Mellon University [20] e na IBM [21]. Nesses trabalhos, foi percebida a necessidade de utilizar técnicas de modelamento estatístico que afrontaram o problema de variabilidade da voz, a qual aumenta significativamente quando a complexidade e o tamanho do vocabulário aumenta.

Nesta seção, serão apresentados os conceitos básicos dos HMMs, sua topologia e os problemas básicos dos HMMs.

2.5.1 Conceitos básicos dos HMMs

O Modelo Oculto de Markov (HMM) é uma técnica de modelagem probabilística que inclui dois componentes básicos:

- Uma cadeia de Markov de estados finitos;
- Um conjunto finito de distribuições de probabilidade de saída.

Estes modelos normalmente permitem a modelagem das unidades fonéticas, que podem ser palavras para pequenos vocabulários, enquanto para grandes vocabulários, como sentenças completas ou até mesmo um parágrafo, são empregadas subpalavras, ou seja, fones, difones, etc. [6]

Em geral, os HMMs podem ser considerados como um conjunto de estados ligados por transições com probabilidades associadas a cada transição, como ilustra a Fig. 2.9. O modelo começa com o estado inicial, e em cada passo de tempo discreto, ocorre uma transição a um novo estado, e um símbolo de

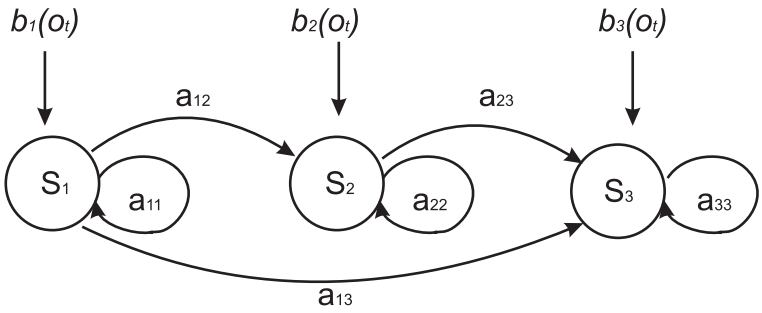

Figura 2.9: Representação de esquerda a direita do HMM

saída é gerado. A transição e o símbolo de saída são aleatórios, sendo regidos por distribuições de probabilidade.

Os estados do processo de Markov estão "ocultos" devido ao fato de que os HMMs encerram processos estocásticos que não são observáveis, mas que afetam a sequência de símbolos emitidos.

Matematicamente, um HMM está caracterizado pelos seguintes elementos [22] [23]:

$S = \{ s_i \}, \quad i = 1, 2..., N$: Que é um conjunto de todos os estados possíveis interligados entre si.

$A = \{ a_{ij} \}$: Matriz de transição de probabilidade de estados, onde a_{ij} é a probabilidade de ocorrer a transição do estado i ao estado j, representado por:

$$a_{ij} = P\{ q_{t+1} = s_j \mid q_t = s_i \} \qquad i, j = 1, 2..., N. \tag{2-7}$$

onde q é a sequência oculta de estados $\{ q_1, q_2, ..., q_T \}$

$\{ V \}$: um conjunto V com M símbolos de observação $V = \{ v_1, v_2, ..., v_M \}$. Estas observações correspondem à saída física do sistema a ser modelado.

$B = \{ b_j(o) \}$: Matriz de probabilidade dos símbolos de observação no estado j, ou distribuição de saída associada ao estado j, representada por:

$$b_j(k) = P\{ v_k \mid q_t = s_j \} \qquad j = 1, 2..., N \quad k = 1, 2..., M. \tag{2-8}$$

onde M é a quantidade de símbolos observáveis

De acordo à natureza da matriz B das distribuições de probabilidade das saídas, os HMMs podem-se classificar em diversos tipos, um deles são os modelos contínuos (CHMMs), onde $b_j(k)$ definem-se em espaços de observações

contínuas, passando a ser uma função densidade de probabilidade (fdp), cujos valores são proporcionais à probabilidade de ocorrência.

Essa fdp tem que representar bem variáveis aleatórias que tendem a se concentrar ao redor de um valor específico. Uma das melhores distribuições é a fdp Gaussiana(ou normal). No entanto uma única distribuição normal não representa bem o comportamento do sinal de voz, é por isso que essas distribuições precisam algum procedimento para que o número de parâmetros do sistema seja adaptável, e as re-estimações sejam constantes, esse procedimento baseia-se em combinar as um conjunto de gaussianas numa nova distribuição. Com esse principio $b_j(k)$ pode ser bem modelado, através da combinação linear de M distribuições gaussianas com pesos (p_{mj}) da seguinte forma

$$b_j(k) = \sum_{m=1}^{M} p_{mj} \frac{l}{(2\pi|\Lambda_{mj}|)^{\frac{l}{2}}} exp\left\{ -\frac{1}{2}(O-\mu_{mj})^T \Lambda_{mj}^{-1}(O-\mu_{mj}) \right\} \qquad (2\text{-}9)$$

$$b_j(k) > 0 \Longrightarrow p_{mj} > 0 \qquad (2\text{-}10)$$

onde Λ é a matrix de covariância da fdp gaussiana, μ é o vetor de média e l é a dimensão do vetor O.

$\Pi = \{\ \pi_i\}$: vetor com as probabilidades de estados inicias. Nos modelos *esquerda a direita*, normalmente é assumido $\pi_1 = 1$ $\pi_i = 0$ para todo $i \neq 1$

$$\pi_i = P\{\ q_i = s_j\} \quad i = 1, 2..., N. \qquad (2\text{-}11)$$

Uma vez que a e b são ambos medidas probabilidades, devem satisfazer às seguintes propriedades:

$$a_{ij} \geq 0, \qquad b_j(k) \geq 0, \qquad \forall i, j, o \qquad (2\text{-}12)$$

$$\sum_{j=1}^{N} a_{ij} = 1, \qquad i = 1, 2, ..., N \qquad (2\text{-}13)$$

$$\sum_{k=1}^{M} b_i(k) = 1, \qquad i = 1, 2, ..., N \qquad (2\text{-}14)$$

Assim, qualquer modelo de Markov passa a ser caracterizado pelo conjunto de parâmetros

$$\lambda = (a_{ij}, \pi_i, p_{mj}, \mu_{mj}, \Lambda_{mj}) \tag{2-15}$$

onde a sequência de símbolos que gera o modelo, $O = (o_1, o_2, ..., o_T)$, é denominada *observação* e a sequência de estados que fica oculta, $Q = (q_1, q_2, ..., q_T)$ é denominada percusso.

No caso de reconhecimento de voz, os HMMs são utilizados ao serem consideradas duas hipóteses [23]:

– A fala pode ser segmentada e 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.

– A probabilidade de observação de que um vetor de características seja gerado depende apenas do estado atual, e de nenhum símbolo gerado anteriormente.

2.5.2 Topologia dos HMMs

Os HMMs seguem a utilização de uma topologia adequada para os modelos, já que assim melhora-se o rendimento do sistema de reconhecimento, devido ao fato de que os algoritmos de treinamento baseiam-se na seleção correta desta topologia.

Uma topologia adequada à natureza sequencial do sinal de voz é o modelo de Bakis [8], mais conhecido como *esquerda-direita*, ilustrado na Fig. 2.9. Neste modelo, os estados estão ordenados e só é permitida a transição de um estado s para ele mesmo, ou para um posterior $s_{i+\Delta}$, onde Δ pode ter valores entre 0 e um valor de salto máximo.

2.5.3 Problemas básicos dos HMMs

Nos HMMs são identificados três problemas fundamentais para seu projeto [8]: Avaliação da probabilidade de uma sequência de observações dado um HMM (onde cada unidade fonética, seja ela palavra ou fonema, será modelada por um HMM característico), determinação da melhor sequência dos estados do modelo e ajuste dos parâmetros dos modelos. Segundo [24], estes três problemas são resumidos em duas etapas: Treinamento e reconhecimento.

A seguir apresenta-se uma explicação de cada um dos problemas e as possíveis soluções que podem resolvê-los visando obter modelos úteis em

aplicações reais.

Etapa 1 Treinamento: Dada uma sequência de observação de treinamento, como treinar o HMM para representar essas observações?.

Este problema pode ser interpretado como a busca pela forma de estimar a matriz de probabilidade de transição A, as distribuições de probabilidade B e as distribuições dos estados inicias Π a partir, de uma sequência de observações.

Etapa 2 Reconhecimento: Consiste na avaliação do problema, quer dizer, dado um HMM treinado, como se encontra a probabilidade do modelo ter produzido uma determinada sequência de observação.

Para resolver estes tipos de problemas, tanto no treinamento quanto no reconhecimento, são utilizadas duas metodologias.

– ***Método de reestimação Baum-Welch*[25]** É uma técnica de maximização das probabilidades, conhecido também como *maximização de esperança* (EM), que utilizando as probabilidades para frente e para atrás do algoritmo *forward-backward*, permite determinar, para a sequência de observação O, a probabilidade de que o modelo λ gere essa observação, isto é $P(O/\lambda)$.

O algoritmo *forward-backward* [26] [15] está dividido em duas partes:

A primeira consiste em definir a variável denominada forward $\alpha_i(i)$, expressa como:

$$\alpha_i = P(O_1, O_2, ...O_t, q_t = S_i \mid \lambda) \tag{2-16}$$

que representa a probabilidade de observar a sequência parcial $O_1, O_2, .., O_t$, até o instante t, e estar no estado S_i naquele instante t, para o modelo λ.

A variável *fordward* pode ser calculada através do seguinte algoritmo:

$$1) \quad \text{Inicialização} \qquad \alpha_1(i) = \Pi_i b_i(O_1), \tag{2-17}$$

$$2) \quad \text{Indução} \qquad \alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij} \right] b_j(O_{t+1} \tag{2-18}$$

$$3) \quad \text{Finalização} \qquad P(O \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i) \tag{2-19}$$

Uma vez inicializada a probabilidade para a frente através da equação 2-17, para a probabilidade dos estados s_i e a observação inicial O_1, implementa-se a etapa de indução (equação.2-18), que é o ponto chave do algoritmo, que permite calcular as variáveis forward no instante $t+1$ a partir das variáveis no instante t, das probabilidades de transição e das probabilidades de observação.

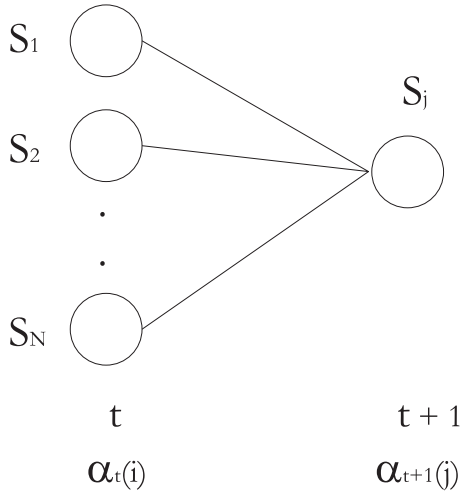

Figura 2.10: Sequência de operações para qualquer variavel $\alpha(i)$ para frente (forward).

Na Fig. 2.10 pode-se ver como o estado s_j é alcançado no tempo $t+1$, partindo-se de N possíveis estados s_i, $1 \leq i \leq N$, no tempo t.

O cálculo da equação 2-18 se faz para todos os estados j, $1 \leq j \leq N$ para um t dado e os cálculos são iterados para $t = 1, 2, ..., T-1$.

Finalmente, a equação 2-19 gera o cálculo desejado de $P(O \mid \lambda)$, como a soma no terminal da variável forward.

Assim reduz-se a complexidade e o custo computacional, já que realiza uma contagem bem mais simples, mudando a complexidade de $2TN^T$ no calculo direito de $P(O)$ em cada uma dos modelos de palavras, para a complexidade TN^2 do calculo forward.

Da maneira similar, pode-se utilizar a variável *Backward* expressada como:

$$\beta_i = P(O_{t+1}, O_{t+2}, ...O_T \mid q_t = S_i, \lambda) \tag{2-20}$$

que representa a probabilidade conjunta de se observar a sequência parcial $O_{t+1}, O_{t+2}, ...O_T$ desde o instante $t+1$ até o instante T, e estar no estado S_i no instante t.

A explicação do método *Baum-Welch* encontra-se no Apêndice A.

– ***Algoritmo de Viterbi*** É uma técnica que obtém a sequência mais provável de estados, para uma dada sequência emitida pelo HMM [27].

Este algoritmo é considerado mais rápido, pois em vez de considerar todas as combinações de transições de estado possíveis, como é feito no algoritmo *forward-backward*, considera somente a sequência de estados com maior probabilidade de produzir a sequência de observações.

Expressada formalmente, seu procedimento é definido por:

$$1) \quad \text{Inicialização} \quad = \begin{cases} V_1(j) = b_j(O_1)\Pi_j \\ B_1(j) = 0. \end{cases} \tag{2-21}$$

$$2) \quad \text{Indução} \quad = \begin{cases} V_t(j) = b_j(O_t)\max_{1\leq i\leq N}[V_{t-1}(i)a_{ij}] \\ B_t(j) = \max_{1\leq i\leq N}[V_{t-1}(i)a_{ij}] \end{cases} \tag{2-22}$$

$$3) \quad \text{Finalização:} \quad \widehat{s_T} = \max_{1\leq i\leq N} B_T(i) \tag{2-23}$$

$$4) \quad \text{Backtracing} \quad = \begin{cases} \widehat{s_t} = B_{t+1}(\widehat{s_{t+1}}), t = T-1, T-2, ..., 1 \\ \widehat{s} = \{\widehat{s_1}, \widehat{s_2}, ..., \widehat{s_T}\} \end{cases} \tag{2-24}$$

A demostração do algoritmo de *viterbi* encontra se no Apêndice B.

2.6 Reconhecimento de voz contínua

O sistema de reconhecimento de voz possui três enfoques distintos, de acordo com o número de palavras que o sistema pode reconhecer e da forma como são faladas.

– Palavra isolada: o falante pronuncia apenas uma palavra e o reconhecedor tentará identificá-la numa lista de palavras conhecidas.

A principal exigência deste tipo de reconhecedor é que as palavras a serem reconhecidas devem ser pronunciadas com pausas de 200 ms entre elas, de modo que seja possível localizar as fronteiras entre elas, facilitando a tarefa de reconhecimento.

– Palavra conectada: o falante pronuncia de forma fluida utilizando uma linguagem reduzida, sem pausas entre palavras.

  Para a tarefa de reconhecimento de palavras conectadas, é conveniente decompor o sistema em dois níveis: nível de frases (gramático) e nível intra-palavra.

  O nível intra-palavra pode ser um HMM da palavra inteira, ou uma representação da palavra formada pela concatenação de modelos HMM de sub-unidades acústicas, como monofones, bifones e trifones.

  O nível gramático é representado por uma rede gramática (de acordo com o modelo de linguagem). Estas representações vão desde redes simples, com poucas restrições sintáticas (por exemplo, gramáticas bigrama ou trigrama), a redes gramáticas altamente complexas e restritivas (por exemplo, gramáticas sensíveis a contexto).

  A utilização desses níveis gera por definição o terceiro tipo de sistema de reconhecimento, o qual é apresentado a seguir.

– Voz contínua: neste sistema o falante faz uma pronuncia natural e o reconhecimento se faz atendendo, geralmente, as unidades menores que as palavras, sobre as sentenças emitidas, isso é, sem necessidade de silêncios entre as palavras que a conformam. Este sistema de reconhecimento é consideravelmente mais complexo do que os sistemas já explicados, por causa de três características da fala contínua, que são:

  – **Limites das palavras:** no reconhecimento de voz contínua não são claros e são difíceis de encontrar. Isto impede a divisão da fala em unidades que sejam notavelmente diferenciadas que pudessem ser tratadas individualmente, como o caso da palavra isolada, no qual os limites são conhecidos, restringindo a pesquisa e melhorando as taxas de acerto.
  – **Efeitos de coarticulação:** embora existam no reconhecimento de palavra isolada, são mais fortes na fala contínua, aparecendo novos efeitos de coarticulação entre as palavras que compõem uma frase, além dos que têm lugar no interior de cada palavra, quando é aumentada a velocidade de pronúncia.
  – **Entonação:** Esta característica acrescenta um novo fator de variabilidade, ao depender da pronunciação de cada palavra e sua situação na frase. Por exemplo, palavras significativas como nomes, verbos, etc têm um realce maior na entonação do que preposições, pronomes, conjunções, etc.

No caso de reconhecimento de voz contínua, a concatenação de CHMMs, como no caso de palavras conectadas, não apresenta resultados viáveis, já que se o vocabulário é muito grande, seria trabalhoso demais gerar centenas ou milhares de modelos. Em vez disso, cada modelo será representado por subunidades de palavra (fones), capazes de serem combinadas entre si para gerar todo o vocabulário. Através desta nova abordagem não importa mais o tamanho do vocabulário. Já que a quantidade de modelos sera fixa, devido a que agora existe um HMM para cada fone. Assim só basta saber quais são os fones de sua pronúncia e então conectar os modelos para formar novos CHMMs que representem as palavras.

No entanto na realidade um fone não é completamente independente de seu vezinho, já que cada um deles sofre influências do fone anterior e do posterior, é daí que sai o conceito de trifone: um fone que é caracterizado pelo seu antecessor e pelo sucessor.

Além disso no reconhecimento de voz contínua não basta utilizar apenas as características da voz, já que podem ser geradas frases sem sentido lógico, é necessário ter certo conhecimento do idioma do locutor e um modelo de linguagem para identificar quais as frases fazem mais sentido, dados os fones pronunciados. Este procedimento é feito através da estatística da ocorrência de palavra [28].

O reconhecimento de voz contínua procurará, então, determinar a sentença mais provável, W, que consiste de uma sequência de palavras, $W = w_1, w_2, ..., w_n$, dado o vetor acústico observado O, isto é, a probabilidade da sequência de palavras dados os vetores de característica do sinal de voz. Esse valor é obtido usando o teorema de Bayes:

$$P(w \mid O) = \frac{P(w)P(O \mid w)}{P(O)} \tag{2-25}$$

Assim, a frase reconhecida $\widehat{w}$ será, finalmente aquela que fornece o maior valor à equação 2-25, ou seja

$$\widehat{w} = arg\max_{w} \frac{P(w)P(O \mid w)}{P(O)} = arg\max_{w} P(w)P(O \mid w) \tag{2-26}$$

onde $P(O)$ é a probabilidade de ocorrer uma determinada observação. Seu valor é constante e independente de W, de modo que é removido do processo de maximização.

Os elementos a avaliar são, portanto,

– A probabilidade a *priori* $P(W)$ de que apareça a sequência W, chamada de modelo de linguagem, que constitui-se de todas as possíveis $P(W)$, ou seja, das probabilidades de ocorrência de todas as palavras, a qual utiliza um conjunto de regras gramaticais para determinar a melhor sequência de palavras, como mostrado na seção 2.2.4.

– A probabilidade da evidência acústica de cada frase $P(O \mid w)$, isto é a probabilidade de que a transcrição W tenha a representação acústica O, chamada de modelo acústico, que seria calculado simplesmente unindo os HMMs das palavras de W utilizando o algoritmo *forward-backward* apresentado na seção 2.5.3.

Por último, devido a que cada modelo de palavra foi formado por HMMs conectados, a rede inteira passa a ser um grande HMM [15]. Logo, a busca pela melhor sequência de palavras $\widehat{w}$ passa a ser a busca pela melhor sequência de estados $\widehat{s}$, ou seja,

$$\widehat{s} = arg\max_{s} P(O, s) = arg\max_{s} P(s)P(O \mid s) \tag{2-27}$$

Nesta rede, o percusso que tiver a maior probabilidade será aquele associado à sequência de palavras mais provável. Este cálculo pode ser feito pelo algoritmo de Viterbi, apresentado na seção 2.5.3.