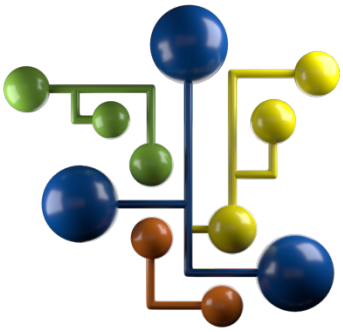


Data Science Academy

www.datascienceacademy.com.br

Introdução à Inteligência Artificial

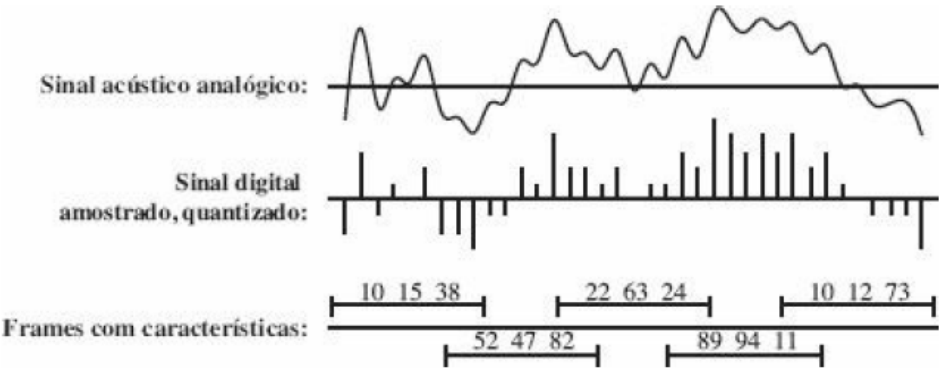
Modelo Acústico e de Linguagem

As ondas sonoras são mudanças periódicas de pressão que se propagam pelo ar. Quando essas ondas incidem no diafragma de um microfone, o movimento de vaivém gera uma corrente elétrica. Um conversor analógico-digital mede o tamanho da corrente — que se aproxima da amplitude das ondas sonoras — em intervalos discretos chamado de taxa de amostragem. Sons de voz, que estão na sua maioria na faixa de 100 Hz (100 ciclos por segundo) a 1.000 Hz, são amostrados tipicamente a uma taxa de 8 kHz (CDs e arquivos de mp3 são amostrados a 44,1 kHz).

A precisão de cada medida é determinada pelo fator de quantização; os reconhecedores de voz normalmente mantêm 8-2 bits. Isso significa que um sistema de baixo custo, com amostragem de 8 kHz, com quantização de 8 bits, iria requerer quase metade de um megabyte por minuto de fala. Uma vez que só queremos saber que palavras foram ditas, não exatamente como soaram, não precisamos manter toda essa informação. Precisamos apenas fazer a distinção entre sons de vozes diferentes. Os linguistas identificaram cerca de 100 sons de voz, ou fonemas, ou segmentos de fala mínimos, que podem ser compostos para formar todas as palavras em todas as línguas humanas conhecidas. Grosseiramente falando, um fonema é o som que corresponde a uma única vogal ou consoante, mas existem algumas complicações: combinações de letras, como “th” e “ng” produzem um fonema único, e algumas letras produzem fonemas diferentes em contextos diferentes (por exemplo, o “a” em rat e rate). A figura abaixo lista todos os fonemas que são usados em inglês, com um exemplo de cada. Um fonema é a menor unidade de som que tem significado distinto para os falantes de uma língua em particular. Por exemplo, o “t” em “stick” soa bem similar ao “t” em “tick” que os falantes de inglês consideram o mesmo fonema. Mas a diferença é significativa no idioma tailandês, por isso há dois fonemas. Para representar o inglês falado precisamos de uma representação que pode distinguir entre fonemas diferentes, mas que não precisa distinguir as variações não fonéticas do som: voz alta ou baixa, rápida ou lenta, masculina ou feminina etc.

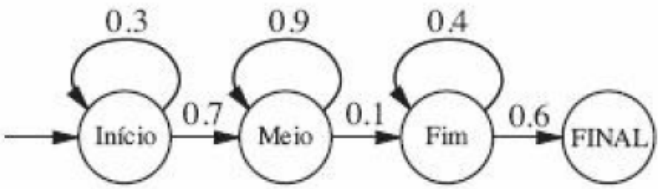
Vogais		Consoantes B–N		Consoantes P–Z	
Telefone	Exemplo	Telefone	Exemplo	Telefone	Exemplo
[iy]	beat	[b]	bet	[p]	pet
[ih]	bit	[ch]	Chet	[r]	rat
[eh]	bet	[d]	debt	[s]	set
[æ]	bat	[f]	fat	[sh]	shoe
[ah]	but	[g]	get	[t]	ten
[ao]	bought	[hh]	hat	[th]	thick
[ow]	boat	[hv]	high	[dh]	that
[uh]	book	[jh]	jet	[dx]	butter
[ey]	bait	[k]	kick	[v]	vet
[er]	Bert	[l]	let	[w]	wet
[ay]	buy	[el]	bottle	[wh]	which
[oy]	boy	[m]	met	[y]	yet
[axr]	diner	[em]	bottom	[z]	zoo
[aw]	down	[n]	net	[zh]	measure
[ax]	about	[en]	button		
[ix]	roses	[ng]	sing		
[aa]	cot	[eng]	washing	[-]	*silence*

Em primeiro lugar, observamos que, embora as frequências de som de voz possam ter vários kHz, as mudanças no conteúdo do sinal ocorrem com muito menos frequência, talvez em não mais de 100 Hz. Portanto, os sistemas de voz resumem as propriedades do sinal em intervalos de tempo chamados quadros. O comprimento de um quadro de cerca de 10 milissegundos (ou seja, 80 amostras em 8 kHz) é curto o suficiente para garantir que poucos fenômenos de curta duração serão perdidos. Os quadros sobrepostos são utilizados para nos certificarmos de não perder um sinal por acontecer de cair sobre um limite do quadro. Cada quadro é resumido por um vetor de características. Escolher as características de um sinal de voz é como ouvir uma orquestra e dizer: “Aqui as trompas estão tocando bem alto e os violinos suavemente.” Apresentaremos uma breve descrição das características de um sistema típico. Primeiro, utiliza se uma transformada de Fourier para determinar a quantidade de energia acústica em cerca de uma dúzia de frequências. Em seguida, calculamos uma medida chamada de coeficiente de frequência mel-cepstral (CFMC) ou CFMC para cada frequência. Calculamos também o total de energia no quadro. Isso fornece 13 características; para cada uma calculamos a diferença entre esse quadro e o anterior, e a diferença entre as diferenças, para um total de 39 características. Esses são de valor contínuo; a maneira mais fácil de adequá-los à estrutura MOM é discretizar os valores (também é possível estender o MOM para manusear misturas contínuas de gaussianas). A figura abaixo mostra a sequência de transformações do som em estado natural para uma sequência de quadros com características distintas.



Vimos como ir do sinal acústico em estado natural a uma série de observações, et. Agora, temos de descrever os estados (não observáveis) do MOM e definir a transição do modelo, P(Xt | Xt-1), e o modelo de sensor, P(Et | Xt). O modelo de transição pode ser quebrado em dois níveis: palavra e fonema. Vamos começar pela parte inferior: o modelo do fonema descreve um fonema como três estados, o início, o meio e o fim. Por exemplo, o fonema [t] tem início em silêncio, um pequeno rompante explosivo de som ao meio, e (geralmente) um assobio no final. A figura abaixo mostra um exemplo para o fonema [m]. Observe que, na fala normal, um fonema médio tem duração de 5-10 milissegundos ou 5-10 frames. Em cada estado, os autolaços permitem variações nesse período. Com muitos autolaços (especialmente no estado do meio), podemos representar um longo som "mmmmmmmmmm". Ignorar os autolaços produz um som "m" curto.

MOM para fone de [m]:

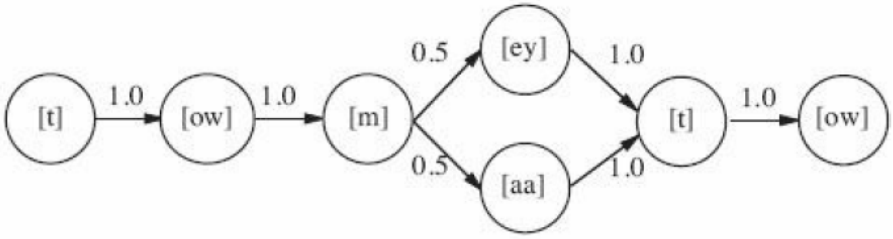


Probabilidades de saída do MOM do fone:

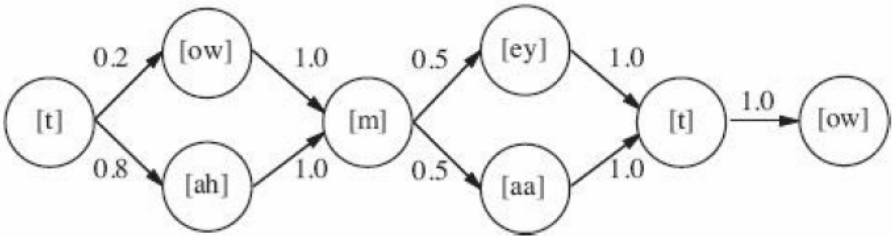
Início:	Meio:	Fim:
C_1: 0.5	C_3: 0.2	C_4: 0.1
C_2: 0.2	C_4: 0.7	C_6: 0.5
C_3: 0.3	C_5: 0.1	C_7: 0.4

Na figura abaixo, os modelos de fonema são amarrados para formar um modelo de pronúncia para uma palavra. De acordo com Gershwin (1937), você diz [t ow m ey t ow] e eu digo [t ow m aa t ow]. A figura (a) abaixo mostra um modelo de transição que prevê essa variação de dialeto. Cada um dos círculos nesse diagrama representa um modelo de fonema.

(a) Modelo de palavra com variação de dialeto:



(b) Modelo de palavra com coarticulação e variações de dialeto:



Além da variação de dialeto, as palavras podem ter variação de coarticulação. Por exemplo, o fonema [t] é produzido com a língua na parte superior da boca, enquanto o [ow] tem a língua perto da parte de baixo. Ao falar rapidamente, a língua não tem tempo para ficar em posição para o [ow], e acabamos com [t ah] em vez de [t ow]. A figura (b) acima apresenta um modelo para "tomato", que leva esse efeito de coarticulação em conta. Modelos de fonemas mais sofisticados levam em conta o contexto dos fonemas circundantes. Pode haver variação substancial na pronúncia de uma palavra. A pronúncia mais comum de "because" é [b iy k ah z], mas isso representa apenas cerca de um quarto das utilizações. O outro quarto (aproximadamente) substitui [ix], [ih] ou [ax] pela primeira vogal, e o restante substitui [ax] ou [aa] pela segunda vogal, [zh] ou [s] pelo final [z] ou tira o "b" inteiramente, deixando "cuz".

Para propósitos gerais de reconhecimento de voz, o modelo de linguagem pode ser um modelo ngrama de texto instruído a partir de um corpus de sentenças escritas. No entanto, a linguagem falada tem características diferentes da linguagem escrita, por isso é melhor obter um corpus de transcrições de linguagem falada. Para a tarefa específica de reconhecimento de voz, o corpus deve ser uma tarefa específica: a construção de um sistema de reservas aéreas, obtenção de transcrições de chamadas anteriores. Ele também ajuda a ter vocabulário de tarefas específicas, como uma lista de todos os aeroportos e cidades atendidas, e todos os números de voo. Parte do projeto de uma interface de usuário de voz é para coagir o usuário a dizer coisas de um conjunto limitado de opções, para que o reconhecedor de voz lide com uma distribuição de probabilidade mais compacta. Por exemplo, a pergunta "A qual cidade você quer ir?" provoca uma resposta com um modelo de linguagem altamente restrito, enquanto a pergunta "Como posso ajudálo?", não.

Você não achou que seria fácil processar linguagem natural a partir da voz, não é? Mas fique tranquilo, pois temos muitas ferramentas que nos auxiliarão neste trabalho!

Referências:

Livro: Inteligência Artificial
Autor: Peter Norvig