Capítulo 2

Produção de Fala

Os sinais de fala são compostos por uma sequência de sons ou segmentos fonéticos, regulados pelas regras da língua e pelas características do orador. Para entender, sintetizar, reconhecer ou, de um modo geral, processar os sinais de fala, é necessário perceber o mecanismo da sua produção. Neste capítulo, discutiremos brevemente a produção da fala e caracterizaremos os sons produzidos, restringindo a discussão aos sons do Português Europeu. Uma discussão aprofundada em termos fonéticos e linguísticos está fora dos objectivos deste texto, mas o conhecimento acerca da estrutura do sinal, ou seja, da forma como a informação está inserida no sinal, é importante antes que se proceda ao estudo sobre os modelos de análise e síntese e das suas aplicações em codificação, síntese e reconhecimento de fala.

2.1 O processo de produção de fala

O aparelho fonador humano, apresentado na figura 2.1, é o primeiro bloco na cadeia da comunicação falada. Após a inalação do ar nos pulmões, os sinais de fala são produzidos durante a fase de exalação (a produção de fala durante a fase de inalação é extremamente rara). Este fluxo de ar, depois da eventual vibração das cordas vocais, situadas na laringe, excitam o tracto vocal constituído pela faringe, cavidade bucal, língua, lábios e dentes. Para produção de sons nasalados o véu palatino abre, pelo que o ar depois de passar pelo tracto nasal é radiado pelas narinas.

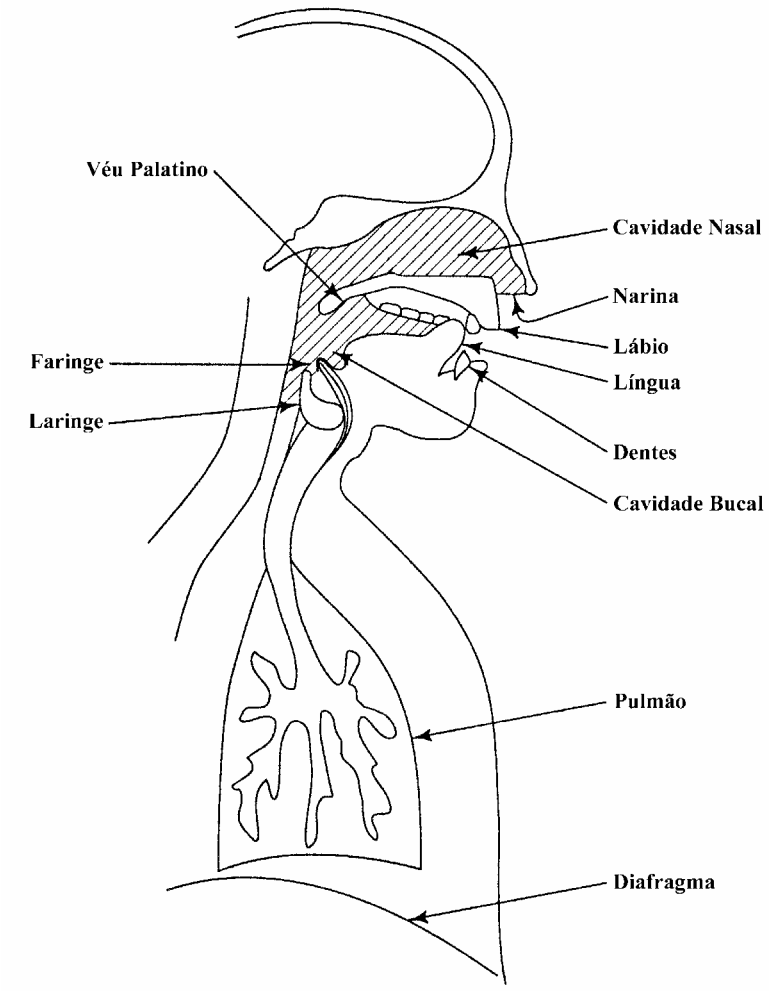

Figura 2.1
Aparelho fonador humano (Adaptado de [Deller (93)]).

2.1.1 Vozeamento

Os sinas de fala são gerados com ou sem vibração das cordas vocais. Os sons produzidos sem vibração das cordas vocais são designados de *não vozeados*, enquanto que os sons produzidos com vibração das cordas vocais, ou seja através da abertura e fecho da glote (espaço entre as cordas vocais), são designados de *vozeados*. Nas zonas vozeadas, à medida que as cordas vocais vibram, estas fazem variar o grau de abertura da glote e consequentemente o volume de ar proveniente dos pulmões que passa através dela. É esta variação periódica na velocidade de volume na glote que vai excitar o tracto vocal, produzindo sons com harmónicas da frequência de vibração das cordas vocais, ou seja, da frequência fundamental ($F0$), habitualmente designada por frequência de *pitch*. Nas zonas não vozeadas a glote mantém-se aberta e o ar proveniente dos pulmões, ao passar com suficiente velocidade por uma constrição do tracto vocal, produz sons com turbulência.

A frequência fundamental depende da dimensão e espessura da glote. Para oradores do género masculino, a gama de vibração das cordas vocais situa-se nos 50-250 Hz, enquanto que para oradores do género feminino essa gama situa-se nos 120-300 Hz, podendo chegar aos 500 Hz para as crianças. Um orador pode ser caracterizado através da sua frequência fundamental média, com variações naturais dependentes da entoação, *stress* e emoção. É normal um orador apresentar uma variação que pode atingir em fala natural uma oitava (*e.g.*, 80-160 Hz para um orador masculino), podendo atingir 2 oitavas no caso de fala forçada ou cantada. Variações mais acentuadas requerem um esforço físico considerável. A figura 2.2 apresenta a forma de onda de um segmento vozeado /e/ e de um segmento não vozeado /s/, ditos por um orador do género masculino e por um orador do género feminino. São

ainda apresentados os respectivos espectrogramas, ou seja gráficos tempo *versus* frequência, em que a intensidade em cada ponto dá informação da energia associada a cada frequência num instante de tempo determinado. Pode-se verificar o maior valor da frequência fundamental para o orador feminino em relação ao orador masculino e a correspondente melhor definição em frequência, já que as harmónicas se encontram mais espaçadas.

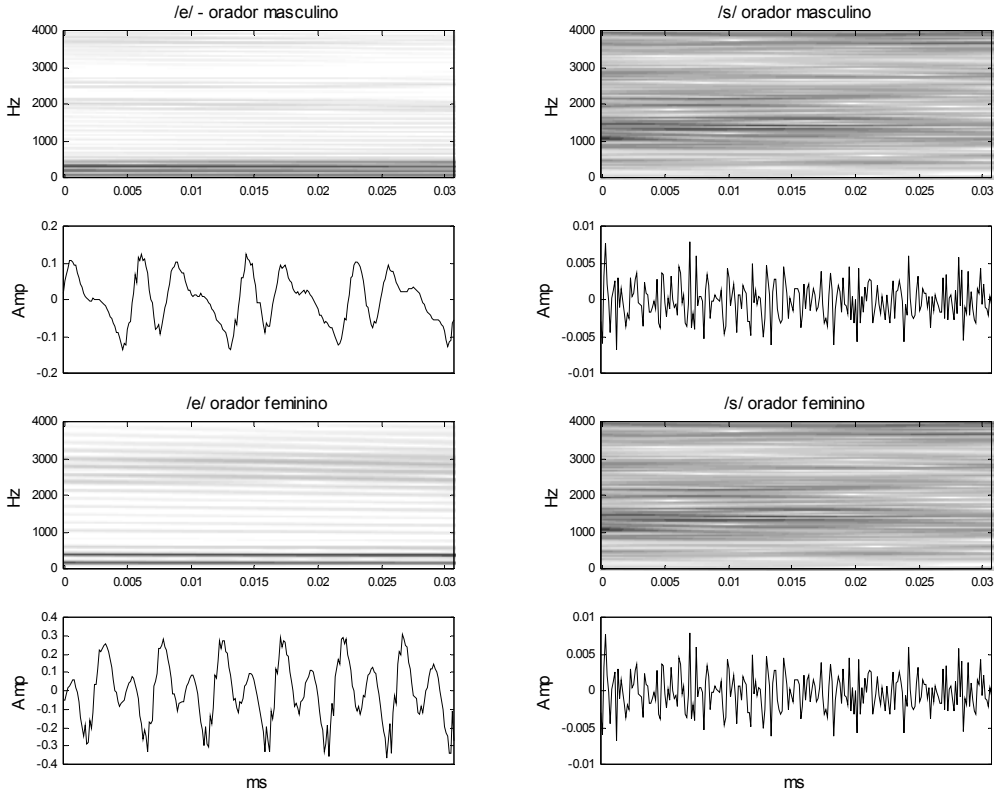

Figura 2.2
Representação temporal e respectivo espectrograma de um segmento fonético vozeado /e/ e de um segmento não vozeado /s/, dito por um orador masculino e outro feminino.

Devido à vibração das cordas vocais que pode ser modelada por um pólo duplo muito perto da frequência zero, as zonas vozeadas têm uma característica passa-baixo. As zonas não vozeadas apresentam em geral maior energia nas altas frequências que as zonas vozeadas.

2.1.2 Formantes

A produção de fala pode ser vista como uma operação de filtragem, na qual uma fonte de som excita o tracto vocal e/ou o tracto nasal. Nas zonas vozeadas a excitação é periódica, sendo do tipo ruidosa e aperiódica nas zonas não vozeadas. Em qualquer dos casos o tracto vocal, actuando como um filtro, amplifica algumas zonas do espectro, atenuando outras. As zonas amplificadas correspondem às zonas de ressonância, definidas por uma frequência central, por uma largura de banda e por uma energia. A frequência central da ressonância é denominada por frequência do formante, ou simplesmente, formante. Os formantes são normalmente representadas por *F*1, *F*2, *F*3,...., começando pela frequência mais baixa. A posição do tracto vocal, especialmente para as vogais, determina os formantes e deste modo o som produzido.

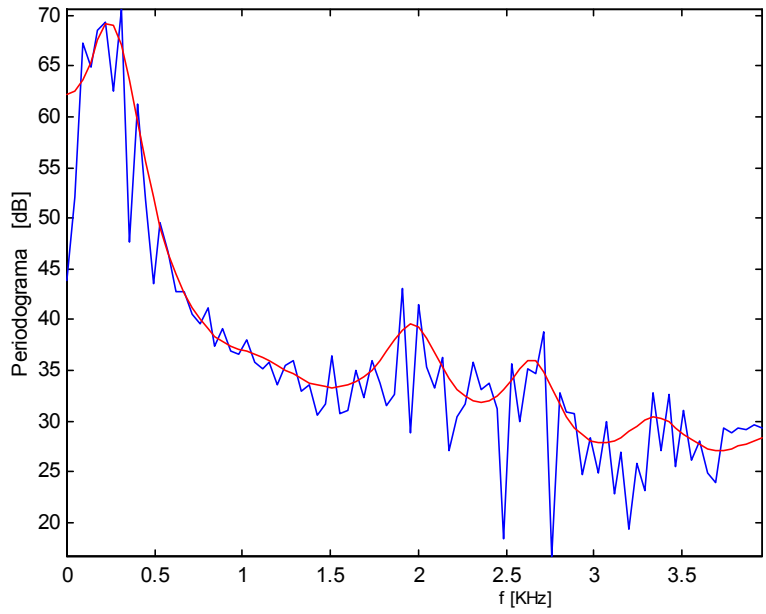

Figura 2.3
Periodograma e respectiva envolvente espectral de uma trama (20 ms) de um segmento fonético correspondente a um /i/, produzido por um orador masculino. (*F*1=266 Hz, *F*2=2044Hz, *F*3=2711Hz, *F*4=3422Hz).

A figura 2.3 ilustra um exemplo do periodograma e respectiva envolvente espectral de um segmento fonético correspondente à vogal /i/, produzido por um orador masculino. Os máximos locais da envolvente espectral correspondem aos formantes, podendo verificar-se na gama de frequências apresentada (0-4 kHz), a ocorrência de 4 formantes. A presença de riscas espectrais (harmónicas da frequência fundamental), embora esbatidas pelo efeito da utilização da janela rectangular de 20 ms utilizada para definir a trama, deve-se à produção deste segmento com vozeamento. É ainda visível o declive espectral que atenua as altas frequências, típico das zonas vozeadas.

2.2 Classificação fonética

Os segmentos fonéticos, para além de se distinguirem pela presença ou ausência de vozeamento, são ainda diferenciados por classes (vogais, glides, oclusivas, fricativas, nasais e líquidas), dependendo do modo de articulação. Dentro de cada classe os segmentos fonéticos distinguem-se ainda pelo ponto de articulação no tracto vocal. Para representar cada um dos segmentos fonéticos é utilizado um alfabeto fonético, sendo o mais conhecido o alfabético fonético internacional (IPA - *International Phonetic Alphabet*). Este alfabeto utiliza no entanto caracteres normalmente não imprimíveis, pelo que utilizaremos o alfabeto fonético SAMPA (SAM *Phonetic Alphabet*) adoptado pelo projecto SAM (*Speech Assessment Methods*) [SAM (92)] e utilizado nomeadamente para transcrever a versão para o Português Europeu do sub-*corpus*[1] de fala EUROM.1 [Ribeiro (93)]. Na tabela 2.1 são apresentados os subconjuntos dos alfabetos IPA e SAMPA necessários para representar o Português Europeu.

[1] *corpus*: base de dados de sinais de fala, utilizado na investigação e desenvolvimento das aplicações em processamento de fala.

Vogais e Glides

Classe	símbolo IPA	símbolo SAMPA	Altura da elevação da língua	Posição da língua na cavidade bocal	palavra	transcrição SAMPA
Vogais	ɐ	6	média	média	cama	k6m6
	a	a	baixa	média	cara	kar6
	e	e	média	anterior	pêra	per6
	ɛ	E	baixa	anterior	sete	sEt@
	ɨ	@	alta	média	que	k@
	i	i	alta	anterior	fita	fit6
	o	o	média	posterior	dou	do
	ɔ	O	baixa	posterior	corda	kOrd6
	u	u	alta	posterior	mudo	mudu
	ɐ̃	6~	média	média	manta	m6~t6
	ẽ	e~	média	anterior	menta	me~t6
	ĩ	i~	alta	anterior	pinta	pi~t6
	õ	o~	média	posterior	ponta	po~ta
	ũ	u~	alta	posterior	mundo	mu~du
Glides	w	w	alta	posterior	pau	paw
	j	j	alta	anterior	pai	paj
	w̃	w~	alta	posterior	cão	k6~w~
	j̃	j~	alta	anterior	mãe	m6~j~

Consoantes

Classe	símbolo IPA	símbolo SAMPA	Presença de Vozeamento	Ponto de articulação	palavra	transcrição SAMPA
Oclusivas	p	p0,p	não	bilabial	pai	p0paj
	t	t0,t	não	apicodental	tia	t0ti6
	k	k0,k	não	velar	casa	k0k6za
	b	b0,b	sim	bilabial	bar	b0bar
	d	d0,d	sim	apicodental	data	d0dat6
	g	g0,g	sim	velar	gato	g0gatu
Fricativas	f	f	não	labiodental	férias	fEri6S
	s	s	não	apicodental	selo	selu
	ʃ	S	não	palatal	chave	Sav@
	v	v	sim	labiodental	vaca	vak6
	z	z	sim	apicodental	azul	6zul~
	ʒ	Z	sim	palatal	agir	6Zir
Nasais	m	m	sim	bilabial	meta	mEt6
	n	n	sim	apicodental	neta	nEt6
	ɲ	J	sim	palatal	senha	s6J6
		N	sim			
Líquidas	l	l	sim	apicodental	lado	ladu
	ɫ	l~	sim	apicodental	sal	sal~
	ʎ	L	sim	palatal	folha	foL6
	R	R		velar	carro	kaRu
	ɾ	r		apicodental	caro	karu
Silêncio		sil				

Tabela 2.1
Alfabetos IPA e SAMPA de descrição do Português Europeu e caracterização dos respectivos segmentos fonéticos pela presença de vozeamento, tipo e posição de articulação no tracto vocal.

Naturalmente estes segmentos fonéticos não ocorrem com a mesma frequência. Ribeiro estima a frequência de ocorrência de cada um dos segmentos fonéticos a partir de um *corpus* de sinais de fala de 32 minutos, correspondente a 8 oradores. Os valores estimados das frequências de ocorrência são apresentados na tabela 2.2, sendo o segmento fonético com maior ocorrência a vogal /6/ com 8%, seguido do segmento /r0/ com 5%. O segmento menos frequente é o /L/ com apenas 0,2%.

SF	FrqOcurr	SF	FrqOcu
6	0,0847	e~	0,0151
r0	0,0515	e	0,0150
t	0,0490	l	0,0147
t0	0,0487	Z	0,0143
a	0,0406	E	0,0129
i	0,0403	@	0,0114
u	0,0380	O	0,0112
r	0,0379	o~	0,0104
d0	0,0369	D	0,0101
d	0,0331	f	0,0094
s	0,0325	R	0,0094
S	0,0312	b0	0,0081
m	0,0294	w	0,0078
k0	0,0294	l~	0,0076
k	0,0292	i~	0,0068
N	0,0262	g0	0,0053
p0	0,0249	J	0,0053
p	0,0248	w~	0,0050
n	0,0191	j~	0,0047
j	0,0188	u~	0,0046
v	0,0179	b	0,0043
6~	0,0170	g	0,0038
z	0,0163	L	0,0023
o	0,0155		

Tabela 2.2
Estimativas [Ribeiro (2000-a)] das frequências das ocorrências (FrqOcu) dos segmentos fonéticos (SF), obtidas em 32 minutos de fala, correspondentes a 8 oradores (4 masculinos e 4 femininos).

2.2.1 Vogais

Os sons correspondentes às vogais são normalmente vozeados e produzidos com o tracto vocal numa forma fixa. Existem em Português Europeu 9 vogais não nasais (/6/, /a/, /e/, /E/, /@/, /i/, /o/, /O/, /u/) e 5 vogais nasais (/6~/, /e~/, /i~/, /o~/, /u~/). As vogais têm normalmente uma duração maior do que as glides e consoantes e uma melhor definição em frequência. Em Português Europeu, contudo, assiste-se frequentemente ao fenomeno denominado de redução vocálica, caracterizado pela diminuição de energia e duração, ou mesmo supressão, de um segmento vocálico.

A figura 2.4 ilustra o gráfico do valor médio do primeiro formante (*F*1) função do valor médio do segundo formante (*F*2), para cada vogal não nasal em Português Europeu, obtidos de nove palavras lidas por nove oradores [Martins (88)]. O triângulo correspondente às vogais /a/, /i/, /u/, é normalmente designado por triangulo das vogais.

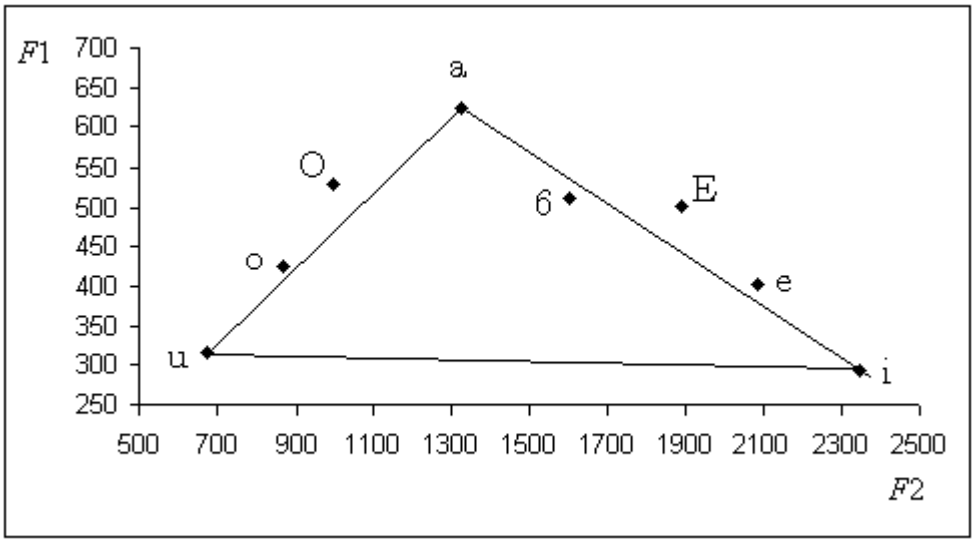

Figura 2.4
O triangulo das vogais.
Gráfico de *F*1 em função de *F*2, para as vogais em Português Europeu.

A tabela 2.3 lista os valores médios dos formantes para as vogais, já ilustrados na figura 2.4, e respectivos desvios padrão. Estes valores podem ser comprovados através dos espectrogramas apresentados na figura 2.5, nos quais são visíveis formantes que correspondem às zonas mais escuras ao longo do tempo. Como se verifica, esta é uma caracterização importante das vogais.

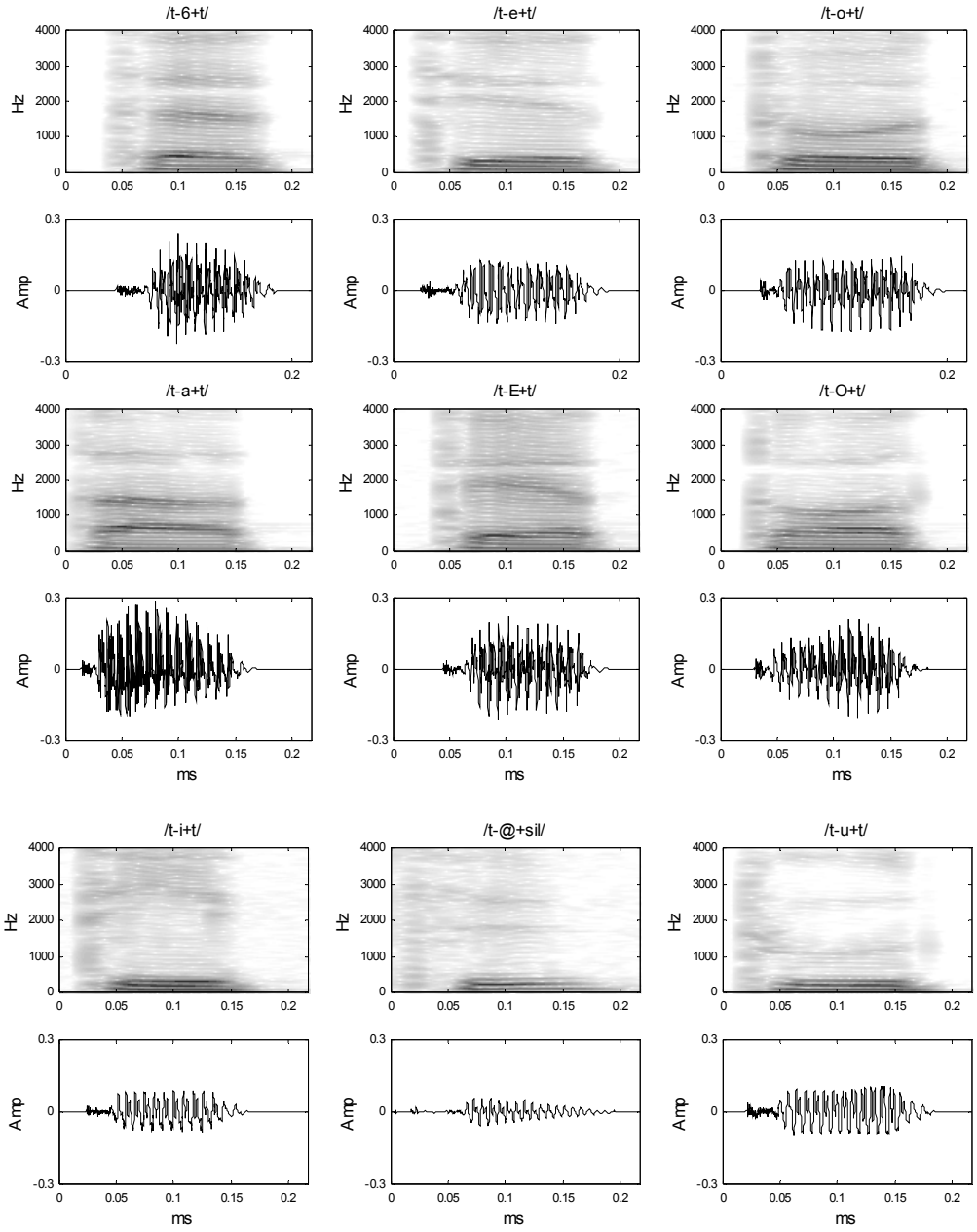

Figura 2.5
Espectrogramas e ondas acústicas das vogais em P.E.
no contexto /t-vogal-t/ (@ no contexto /t-@+sil/).

Formante	Segmento fonético⇒	6	a	e	E	i	o	O	u
F1	Valor Médio	511	624	403	501	294	426	531	315
	Variância	56	78	40	46	37	46	57	45
F2	Valor Médio	1602	1325	2084	1893	2344	864	994	678
	Variância	205	157	187	155	139	111	81	124

Tabela 2.3
Valor médio e desvios padrão das frequências dos formantes para as vogais em Português Europeu. (Adaptado de [Martins (88)])

Nestes gráficos, desde que o espectro localizado seja calculado com suficiente resolução, é possível verificar quer a estrutura harmónica quer a posição (grosso modo) dos formantes.

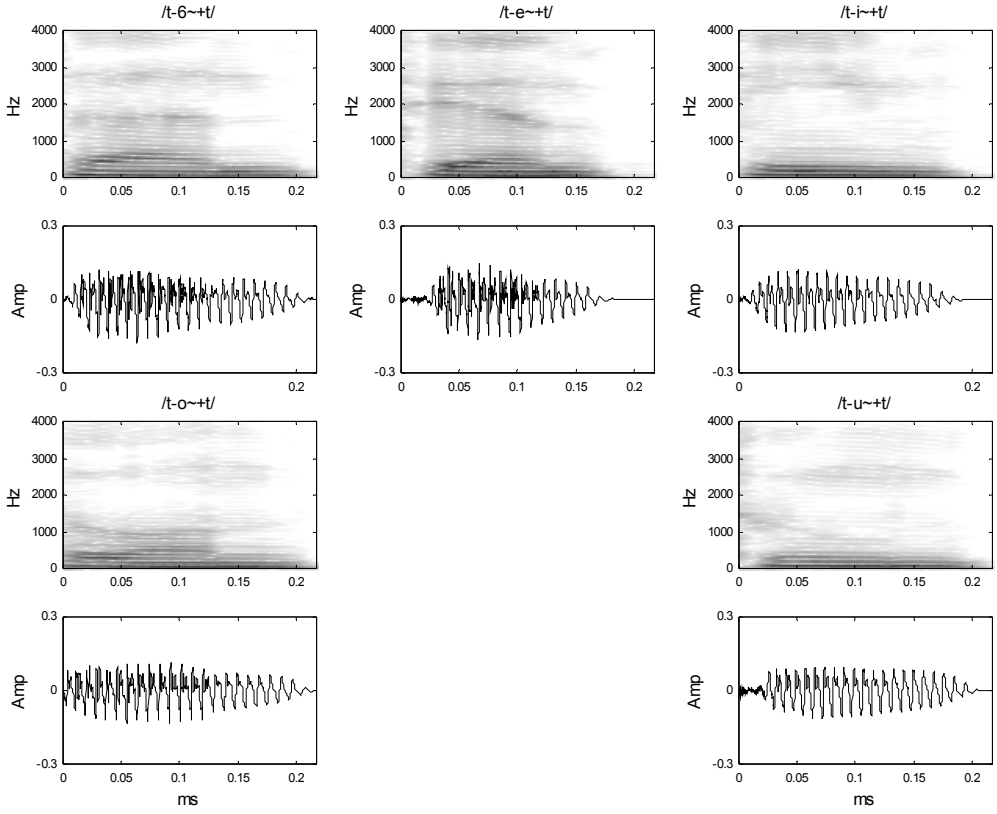

Figura 2.6
Espectrogramas e ondas acústicas das vogais nasais em P.E.
no contexto /t-vogal-t/.

2.2.2 Glides

As glides ou semi-vogais, /w/ e /j/, e os respectivos sons nasalados /w~/ e /j~/, ocorrem em Português Europeu simultaneamente com uma vogal que lhe precede ou procede, formando ditongos, em que há transição dos formantes entre dois valores, correspondentes aos dois sons do ditongo. As glides podem ser vistas como vogais com maior constrição e menor duração que as vogais respectivas (/w/:/u/, /j/:/i/).

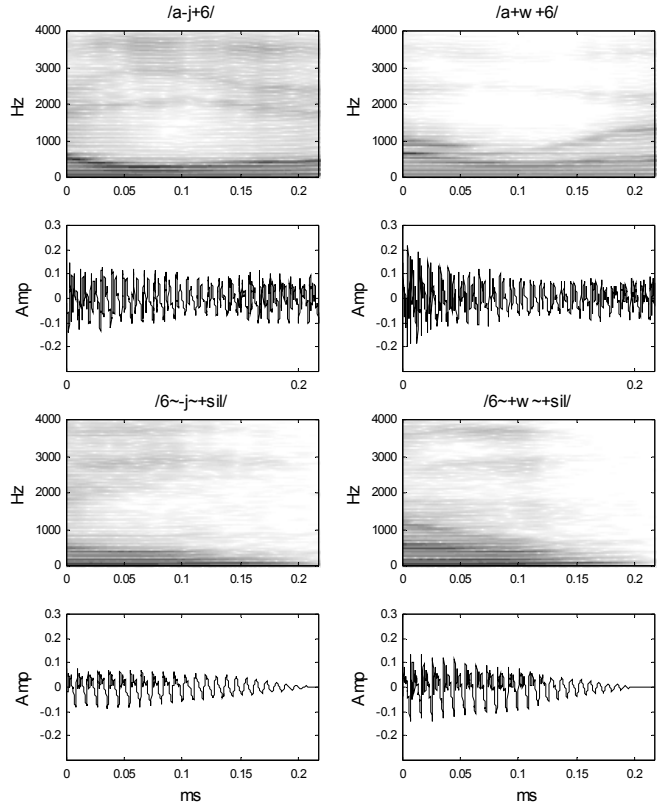

Figura 2.7
Espectrogramas e ondas acústicas das glides em P.E.
no contexto /a-glide-6/ ou /a~-glide nasal-sil/.

2.2.3 Oclusivas

As oclusivas são sons produzidos pela constrição total do tracto vocal (zona de oclusão), seguida da libertação da pressão acumulada (zona de explosão). As diferentes oclusivas são distinguidas através do ponto em que se dá a oclusão e da presença (/b/, /d/, /g/) ou ausência (/p/, /t/, /k/) de vozeamento. Estas últimas apresentam uma zona de oclusão com um silêncio quase total, enquanto que os segmentos oclusivos vozeados mantêm a periodicidade dos segmentos vizinhos. Uma vez que as zonas de oclusão e de explosão exibem características bastante distintas, o alfabeto SAMPA foi estendido de modo a distingui-las, sendo a zona de oclusão definida colocando um '0' após o símbolo que representa a explosão (*e.g.*, /p0/ para a zona de oclusão e /p/ para a zona de explosão).

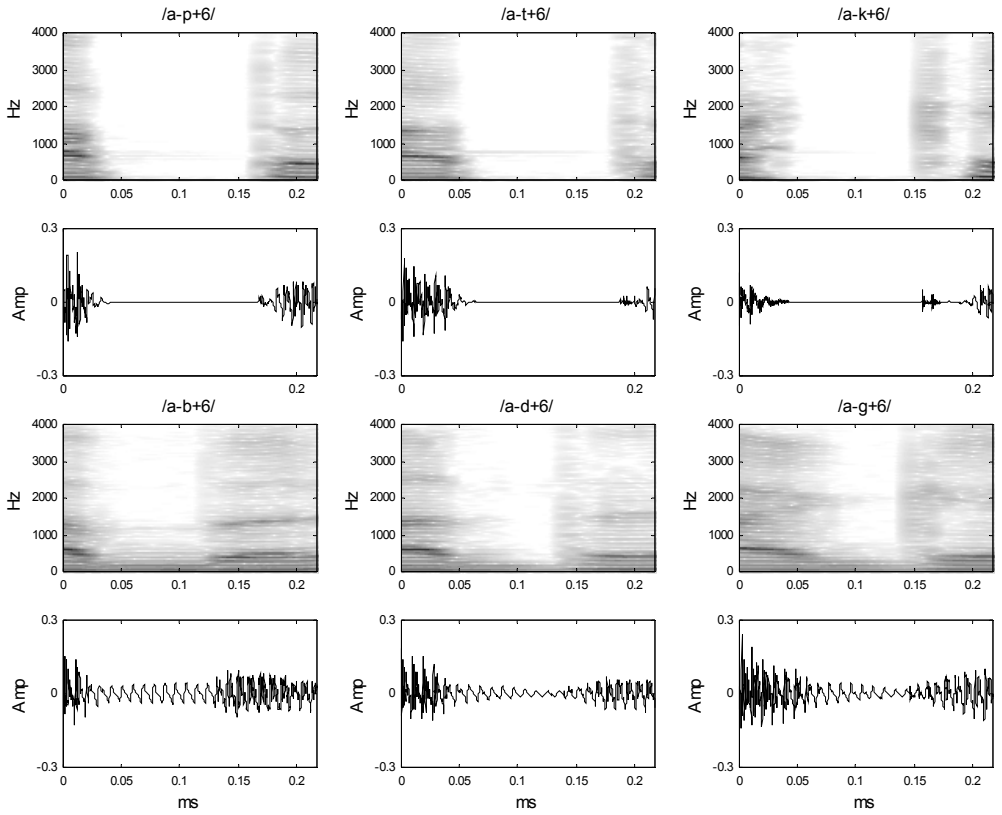

Figura 2.8
Espectrogramas e ondas acústicas das oclusivas em P.E. no contexto /a-oclusiva-6/

2.2.4 Fricativas

As fricativas são produzidas com uma constrição do tracto vocal, que dá origem a turbulência. As fricativas podem ser distinguidas através do ponto de constrição e da presença (/v/, /z/, /Z/) ou ausência (/f/, /s/, /S/) de vozeamento. No entanto as fricativas vozeadas, devido à presença de turbulência, têm na realidade uma componente não periódica, sendo consideradas como tendo excitação mista. Uma das características das fricativas, contrariamente à maioria das outras classes fonéticas, é a grande energia contida nas altas frequências, pelo que podem perder a inteligibilidade quando filtradas passa-baixo (*e.g.*, através de um canal telefónico). Tal como as oclusivas, as fricativas têm uma intensidade bastante mais baixa que as vogais.

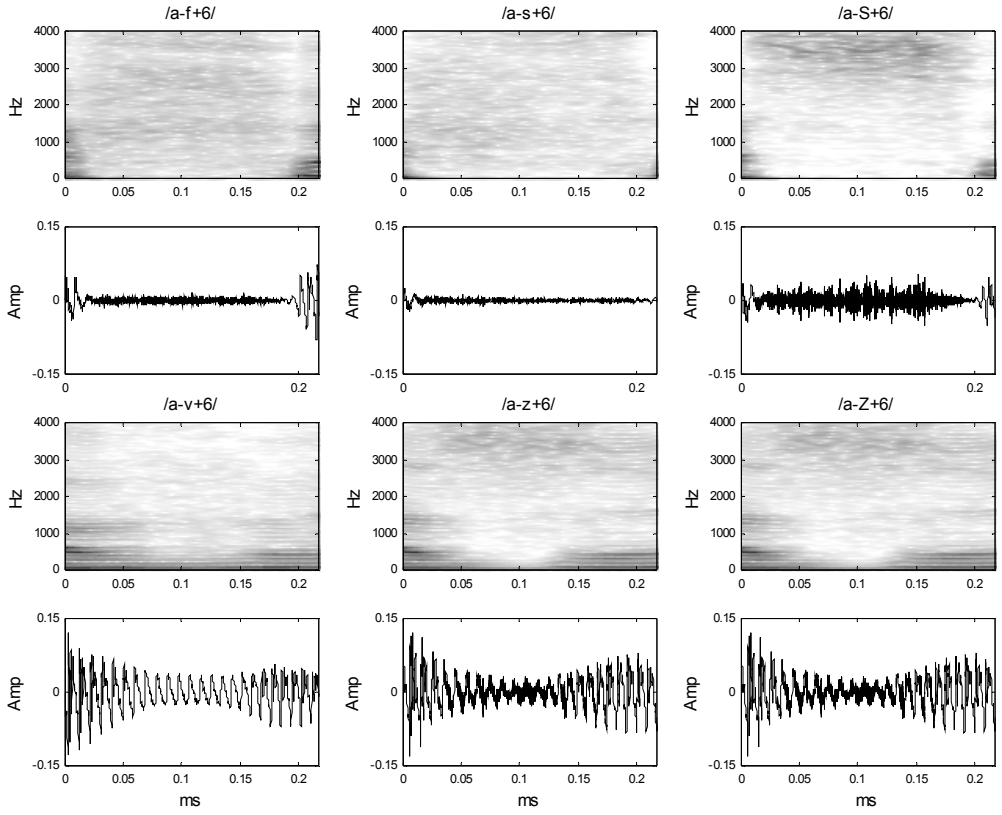

Figura 2.9
Espectrogramas e ondas acústicas das fricativas em P.E. no contexto /a-fricativa-6/

2.2.5 Nasais

As nasais /m/, /n/, /J/ são produzidas com vibração das cordas vocais e com o tracto vocal totalmente fechado num ponto ao longo da cavidade bucal. Adicionalmente o véu palatino baixa e, consequentemente, o ar proveniente dos pulmões é radiado através das narinas. A cavidade bucal embora fechada mantêm-se acoplada à faringe e à cavidade nasal, resultando uma anti-ressonância, ou seja um zero em termos espectrais, muitas vezes dominante e cuja frequência é inversamente proporcional à dimensão da constrição da cavidade bucal, ocorrendo a uma frequência menor para o /m/ e maior para o /J/. Dada a oclusão do tracto vocal, estes segmentos são também designados de oclusivos nasais.

Quando um segmento fonético nasalado, quer este seja uma consoante nasalada ou seja uma vogal ou glide nasalada, precede uma oclusiva, a nasalidade pode-se prolongar para a zona de oclusão. Uma extensão do alfabeto SAMPA utiliza o símbolo /N/ para marcar esta variante da oclusão, com características diversas de uma zona de oclusão sem a nasalidade activa. É também normal que durante uma vogal que preceda uma oclusiva nasal o véu palatino baixe, causando a nasalidade da vogal.

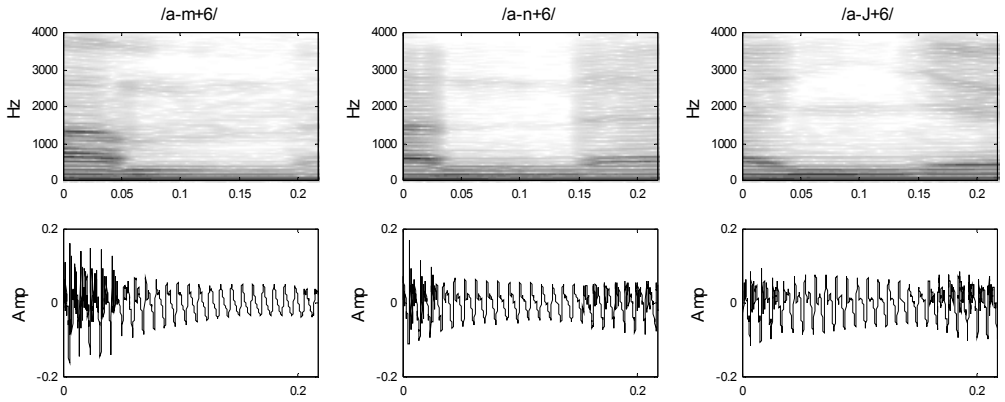

Figura 2.10
Espectrogramas e ondas acústicas das nasais em P.E. no contexto /A-nasal-6/

2.2.6 Líquidas

As líquidas têm espectros que tal como as vogais têm uma estrutura marcada de formantes, embora com uma menor energia. Estas dividem-se em laterais (/l/, /l~/ e /L/), e vibrantes (/r/ e /R/). As laterais são pronunciadas com obstrução do fluxo de ar no tracto vocal provocada pela língua, com o ar a passar por ambos os seus lados. As líquidas /l/ e /l~/ (l-velarizado) têm o mesmo ponto de articulação, mas o /l~/ ocorre apenas em final de sílaba. A vibrante /R/ (r múltiplo) é produzida com a língua a vibrar, atingindo repetidamente o velo. No caso do /r/ (r simples), este é produzido com apenas um toque da língua nos alvéolos dentários. Estes segmentos têm contudo uma grande variabilidade, podendo ou não ser vozeados e fricatizando em alguns casos.

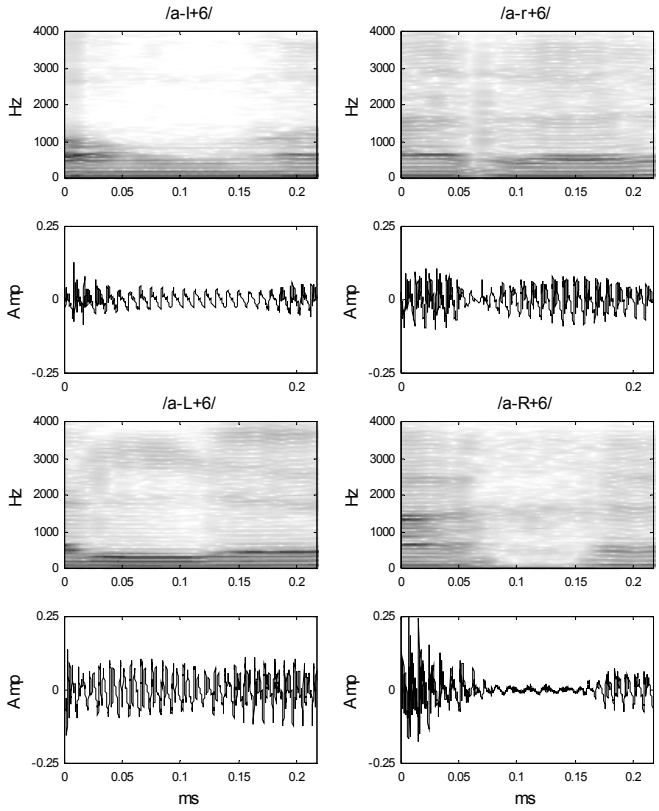

Figura 2.11
Espectrogramas das líquidas em P.E. no contexto /A-líquida-6/.

2.3 Coarticulação

Como se pode verificar nas zonas vozeadas ilustradas na figura 2.2, os períodos glotais não são exactamente iguais, sendo as variações da forma de onda causadas quer pela evolução lenta do tracto vocal, quer por diferenças de energia. A fala não é na realidade uma sequência de sons bem definidos, com uma mudança brusca entre estes, mas antes a transição entre um par de segmentos fonéticos produz-se de forma gradual, exibindo o sinal pequenas variações das características de um som para o do som procedente, efeito denominado de *coarticulação*. De notar, contudo, que a fala contêm outra informação para além da simples sequência de sons e respectiva coarticulação, uma vez que os ouvintes podem inferir a identidade do orador, o seu género e idade, estado de alegria ou tristeza e as suas emoções.

2.4 Transcrição Fonética

Para o desenvolvimento de sistemas de processamento de fala é frequentemente necessário traduzir uma onda acústica nos sons produzidos, processo denominado de transcrição fonética. Este processo produz a sequência de símbolos fonéticos e respectivas marcas temporais, utilizando para tal um alfabeto fonético. Como exemplo, a figura 2.12 apresenta uma forma de onda da frase "*e a chuva não bate assim*", a que corresponde a transcrição fonética utilizando o alfabeto SAMPA /sil j 6 S u v 6 n 6~w~ N b0 b a t0 t 6 s i~ sil/.

Transcrever foneticamente uma frase é uma tarefa de realização difícil, devendo ser efectuada manualmente por um especialista em fonética, recorrendo à análise da onda acústica, ao espectrograma e à audição do trecho correspondente. A marcação das fronteiras pode no entanto ser auxiliada por um reconhecedor fonético que force o

alinhamento entre a onda acústica e a sequência fonética [Ribeiro (96)]. Na maioria das vezes resta ao transcritor manual apenas introduzir pequenas correcções nas fronteiras entre segmentos.

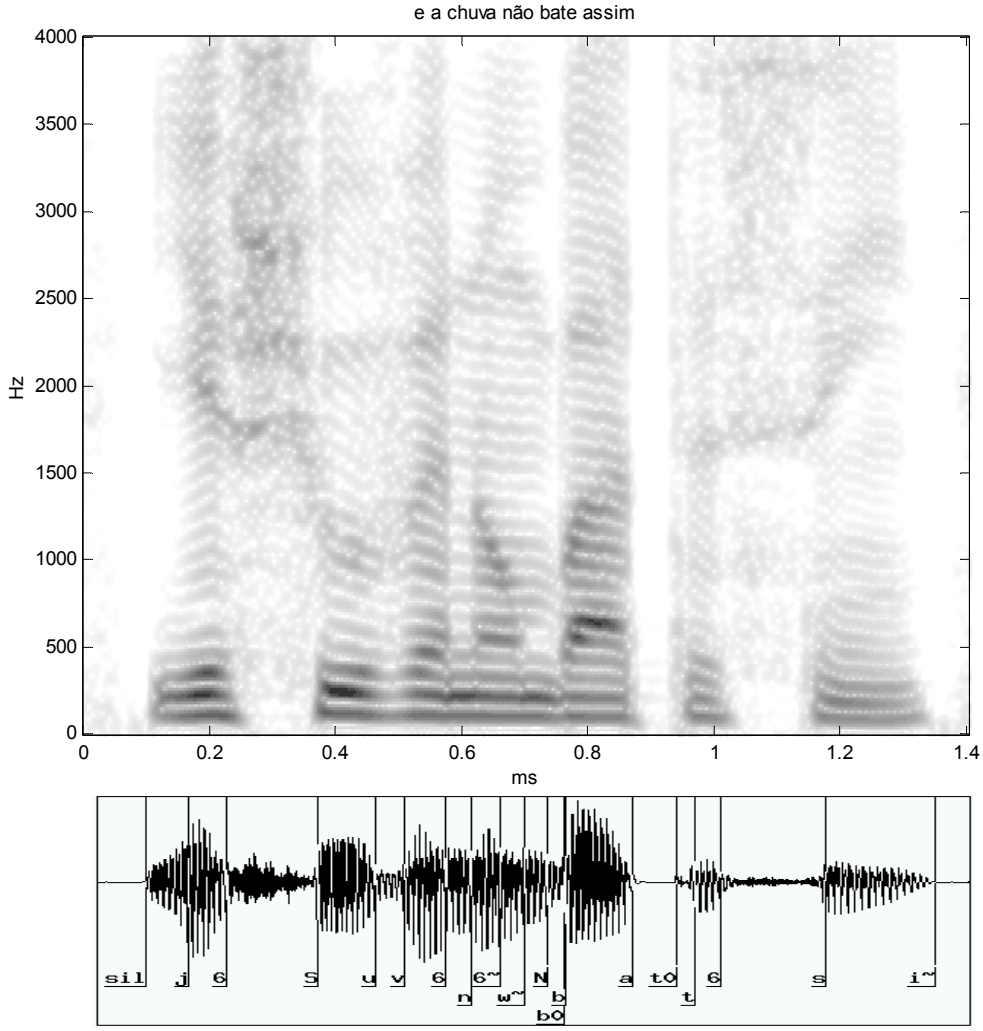

Figura 2.12
Onda acústica e respectiva segmentação e transcrição fonética, correspondente à frase "*e a chuva não bate assim*", produzida por um orador masculino.

Um outro nível de anotação mais simples é a transcrição fonética larga, derivada apenas da transcrição ortográfica de determinada frase, não tendo associada uma onda acústica. O termo *larga* provém do facto de a sequência fonética produzida corresponder muito de perto à ortografia, ocorrendo variações para determinada realização, nomeadamente devido à coarticulação e à redução vocálica.