Feedback

Speech Processing, first assignment, November 2017

Marking process

- Lab reports
  - Separate markers for UG and PG
  - All marking was carefully moderated (but not re-marked) by the lecturer
    - Every item: briefly inspected
    - A sample of items across different grade bands: closely inspected

- Literature reviews
  - UG & PG all marked by the lecturer

Moderation process

- Primary goal of the markers is to give you **feedback** on how to do better next time
- Moderation involves shifting and/or scaling of all marks (usually upwards)
  - done separately for the lab report and lit review, and separately for UG & PG
- Your work was marked on hardcopy, and now has two cover sheets attached
  - structured marking scheme
    - raw mark per category
    - sum of raw marks
    - moderated mark (circled, in red)
  - feedback comments
    - 'canned' comments, with those that apply to you circled or highlighted

This year's feedback theme: figures, graphs, tables, diagrams, …

Layout

For example:

'dove' (bird) nn (d uh v)

'dove' (action) vbd (d ou v). (Festival, 2004)

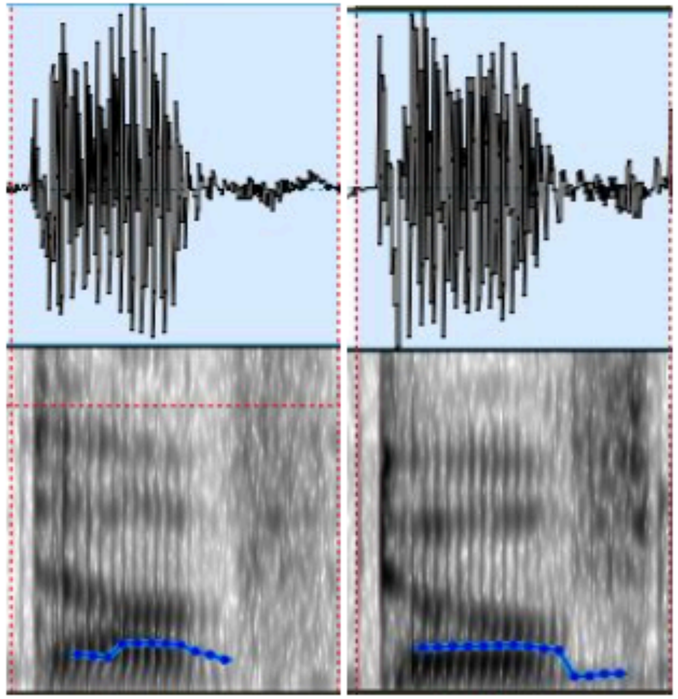


*Figure **2**: On the left, "dove", in the context of "It's a dove."*

On the right, "dove", in the context of "He dove down."

add space here, or 'float' the figure to the top of the page

Layout

don’t allow a page break inside a table, unless it’s more than a page long

“I lost… everything.”

Word	Break

I	NB
lost	NB
eveything	BB

Layout

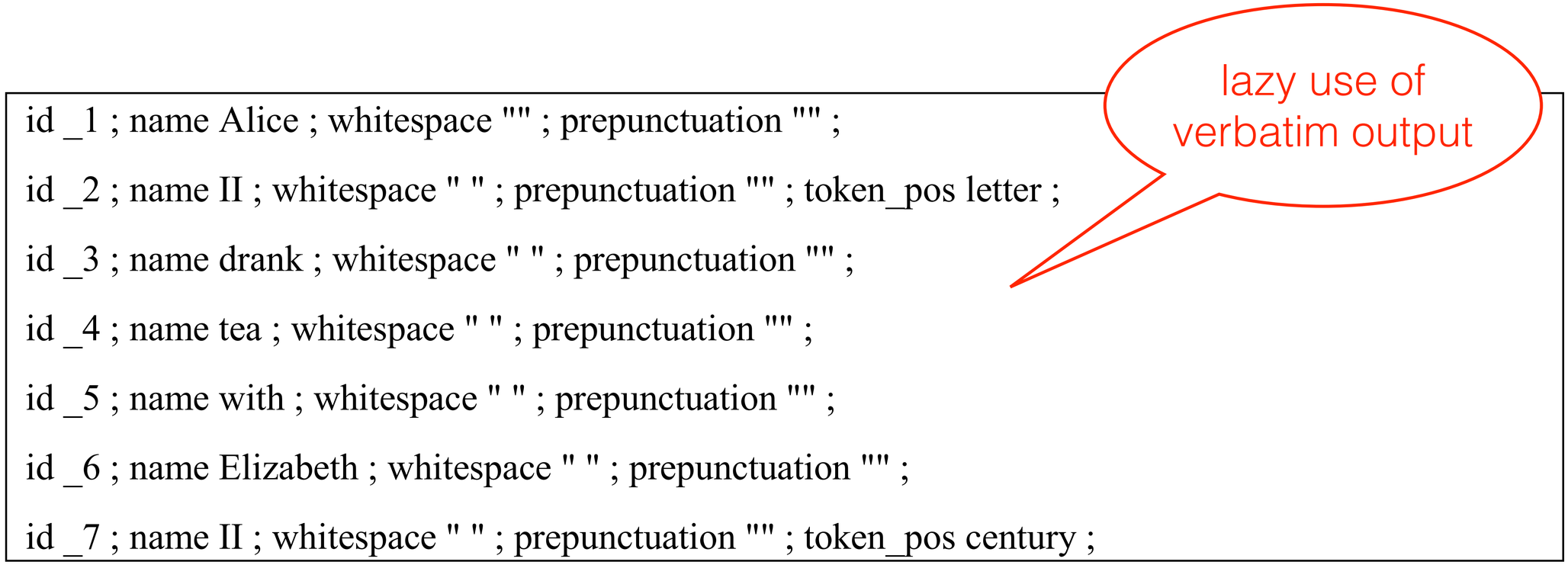


Figure 4: Contents of the (Token) relation after running Token_POS.

Layout

your word processor has capitalised these words

Word	I	Live	In	Edinburgh
POS Tag	nn	jj	in	nnp
POS Tag meaning	noun	Adjective	Preposition	Proper noun
Table 1: Table showing POS tags in the word relation for ‘I live in Edinburgh’.				

not usual to place caption *inside* a table

Layout

phonemes. Here, some of the post-lexical changes that are to be applied later are specified.

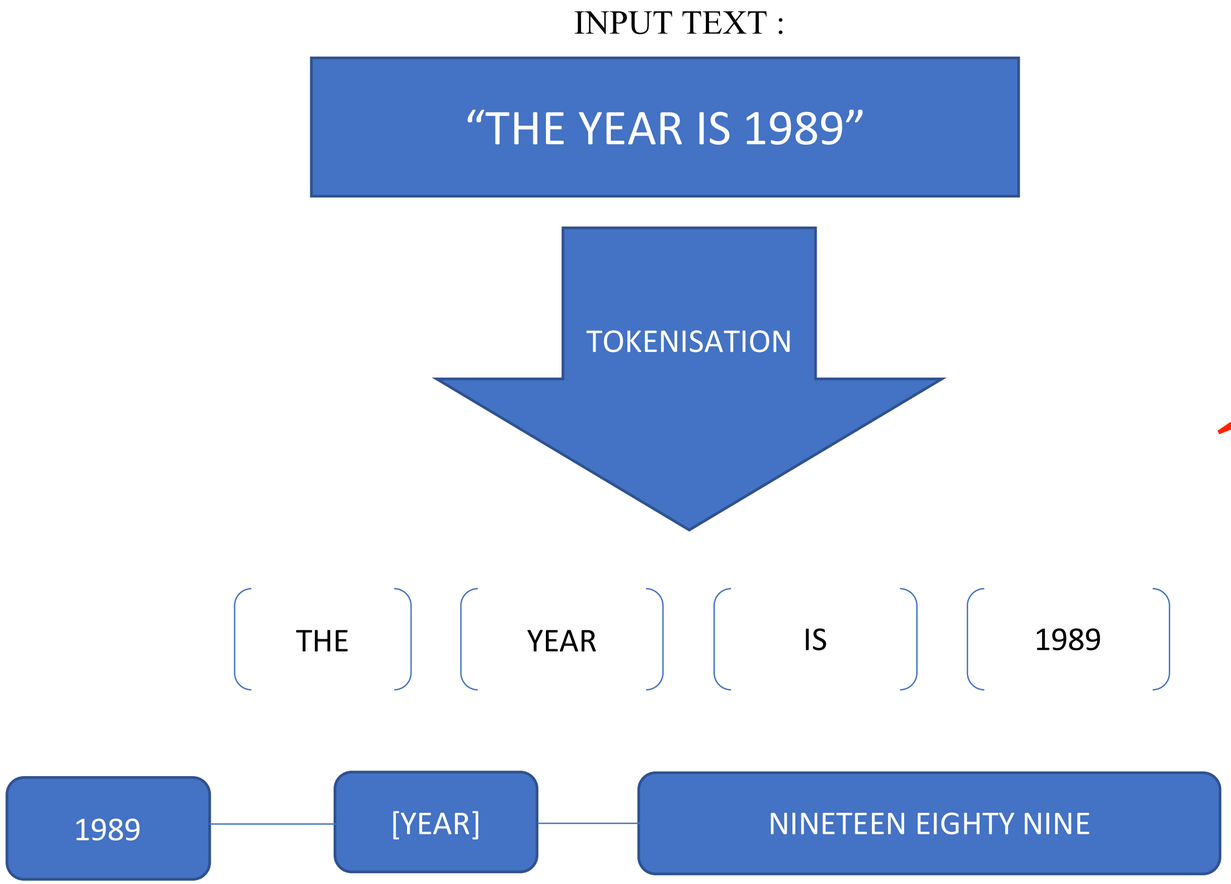
Those changes are derived with the help of hand-crafted rules [6, Slide 30].

Figure 2: Segment relation for the word ‘them’

```
id _28 ; name dh ; end 1.3 ; source_end 0.964563 ;
id _29 ; name e ; reducable 1 ; fullform e ; reducedform @ ; end 1.4 ;
source_end 1.0825 ;
id _30 ; name m ; end 1.5 ; source_end 1.1785 ;
```

caption is closer to body text than to the figure it belongs to

Layout



seems to be
two flowcharts within
one figure

Figure 2.2: *A representation of how Festival tokenises non-word sequences to determine what they are and how to present them for further processing.*

Layout

```
(set! simple_phrase_cart_tree
'
((R:Token.parent.punc in ("?" "." ":"))
  ((BB))
  ((R:Token.parent.punc in ("'" "\"" "," ";"))
   ((B))
   ((n.name is 0)
    ((BB))
    ((NB))))))
```

would be better to **draw** the tree instead (or as well)

Figure 3: CART for phrase break prediction (Black, Taylor, & Caley, 1999)

Layout

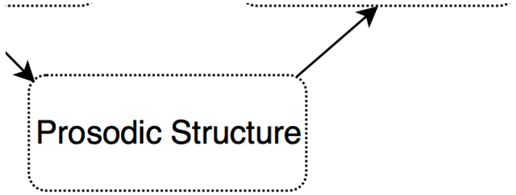


Figure 2.1: Overview of the TTS pipeline

equal spacing makes it hard to see section headings

2.1 Text processing

2.1.1 Tokenisation: (Text)

might be better to vary font size more with heading depth

en down into tokens using simple whitespace tokeni

Token	Punctuation
St	.
Bolbits	

Layout

on the history, they are also dependent on the future words. HMM models could be done bi-directionally. To approach this problem, I checked out the most ambiguous tags that are already reported by [4].

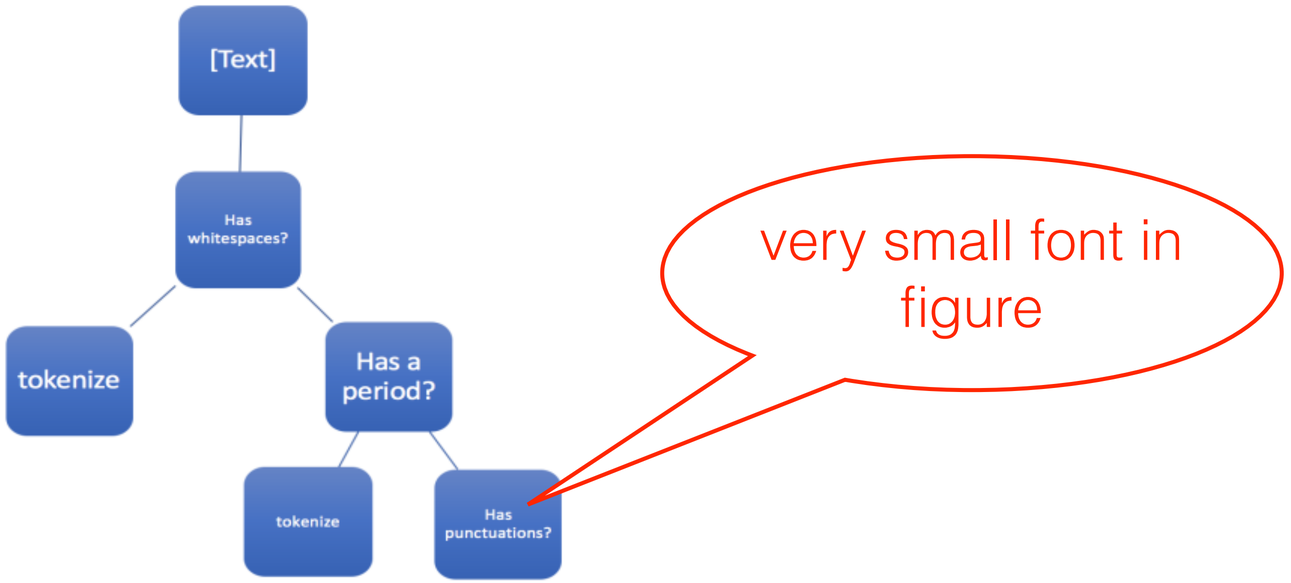


Fig. 1. A Simple CART tree used for text normalization to split tokens [5]

Layout

needs more space
to make start/end of the
example clear

incorrect
quote marks: use
“ ”

Other POS errors occur with homographs such as *read* and *lead*
tag seems to have no effect on the pronunciation retrieved from

”I [rɛd] yesterday” `hg_pos red pos vbd`

”I have [rid]” `hg_pos red pos vbn`

”To [rid] is to live” `hg_pos red pos vb`

The same occurs for *lead*, but here even the POS tag seems to have
could be because its POS dictionary entries are formatted:

Quality of reproduction

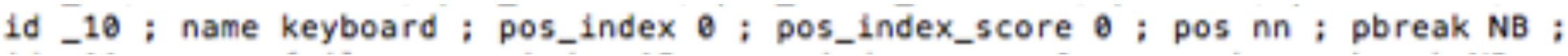


Figure 1.2: POS tagging of “keyboard” in Festival

not just verbatim output, but a **pixellated screenshot of verbatim output**

Quality of reproduction

```
;; Some form of money (pounds or type
(let (amount type currency)
  (cond
   ((string-matches name ".*\\$.*")
    (set! amount (string-after name "
    (set! type (string-before name "
    (set! currency "dollar"))
   ((string-matches name ".*£.*")
    (set! amount (string-after name "
    (set! type (string-before name "
    (set! currency "pound"))
   ((string-matches name ".*#.*")
    (set! amount (string-after name "
    (set! type (string-before name "#"))
    (set! currency "pound"))
   ((string-matches name ".*Y[0-9].*")
    (set! amount (string-after name "Y"))
    (set! type (string-before name "Y"))
    (set! currency "yen"))
   ((string-matches name ".*\\\\.*")
    (set! amount (string-after name "\\"))
    (set! type (string-before name "\\"))
    (set! currency "yen"))
```

"pound"))

name ".*

string-aft

cing-befor

pixel-based image, but at least high resolution

but why not use actual **text** ?

Figure 1.1. Festival expands currencies with hard-coded rules. (CSTR, 2015)

Quality of reproduction

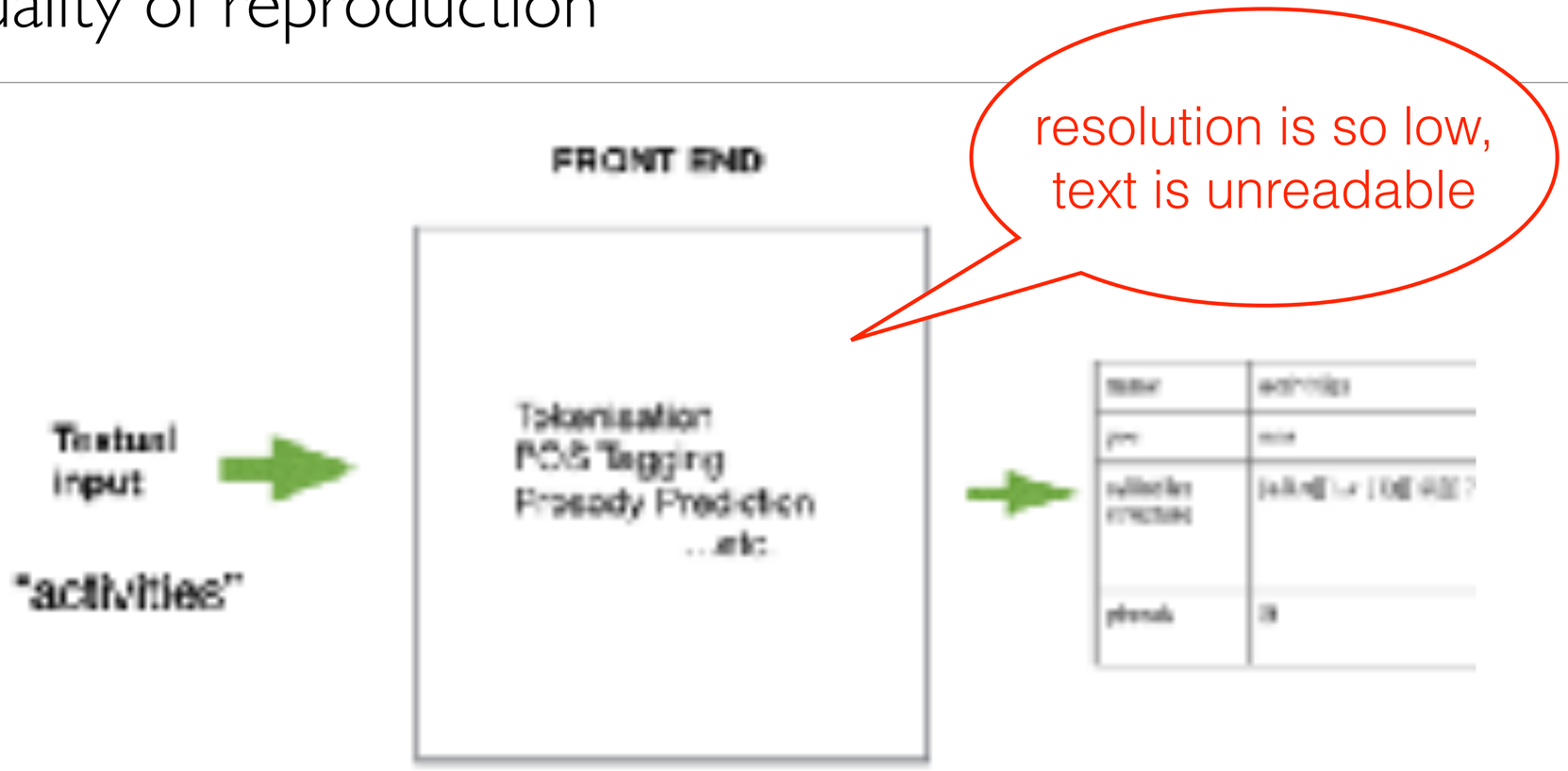


Figure 1: *Different steps in the TTS pipeline.*

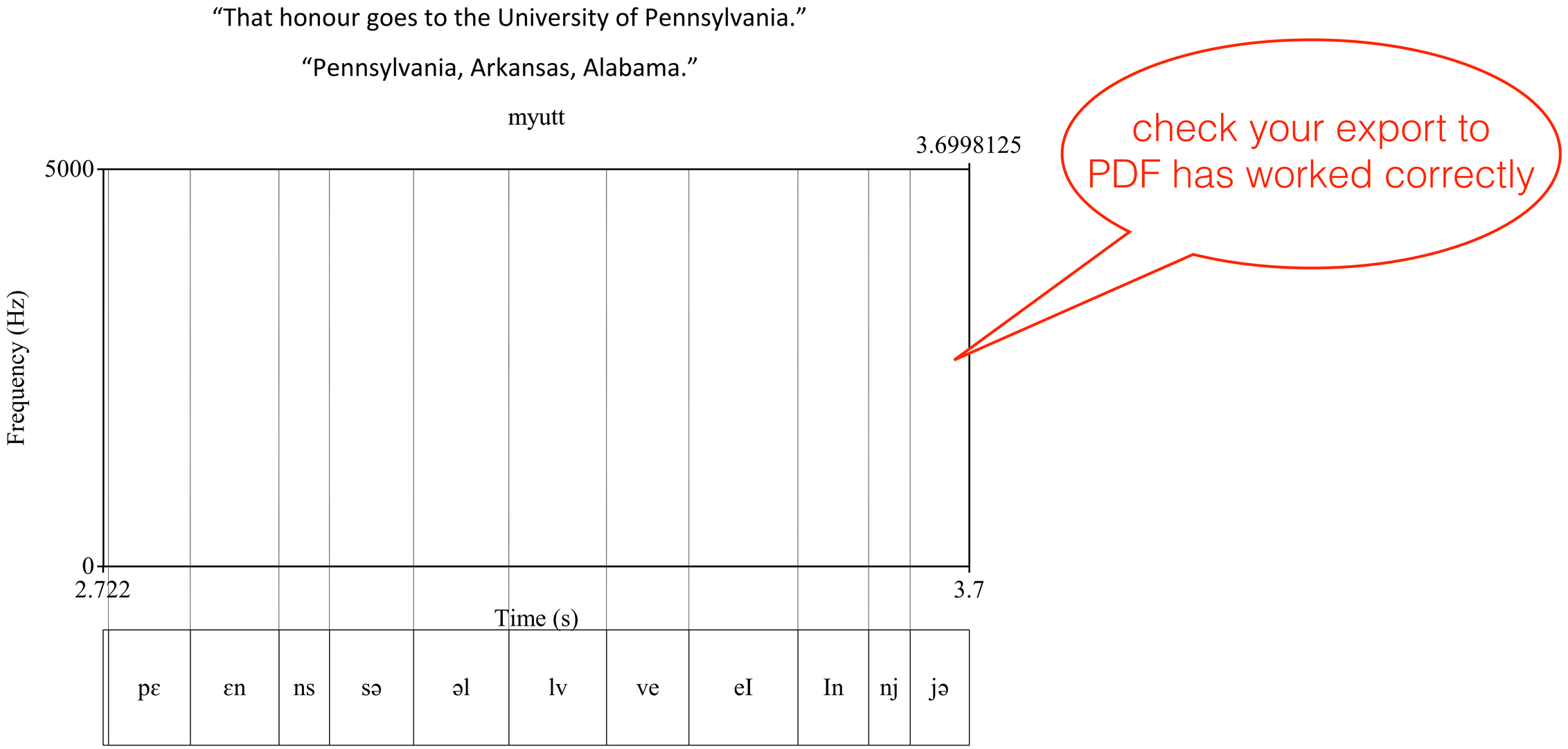
Quality of reproduction

	JJ	NN	NNP	NNPS	RB	RP	IN	VB	VBD	VBN	VBP	Total
JJ	0	**177**	**56**	0	**61**	2	5	10	15	**108**	0	488
NN	**244**	0	**103**	0	12	1	1	29	5	6	19	525
NNP	**107**	**106**	0	**132**	5	0	7	5	1	2	0	427
NNPS	1	0	**110**	0	0	0	0	0	0	0	0	142
RB	**72**	21	7	0	0	16	**138**	1	0	0	0	295
RP	0	0	0	0	**39**	0	**65**	0	0	0	0	104
IN	11	0	1	0	**169**	**103**	0	1	0	0	0	323
VB	17	**64**	9	0	2	0	1	0	4	7	**85**	189
VBD	10	5	3	0	0	0	0	3	0	**143**	2	166
VBN	**101**	3	3	0	0	0	0	3	**108**	0	1	221
VBP	5	34	3	1	1	0	2	**49**	6	3	0	104
Total	626	536	348	144	317	122	279	102	140	269	108	3651

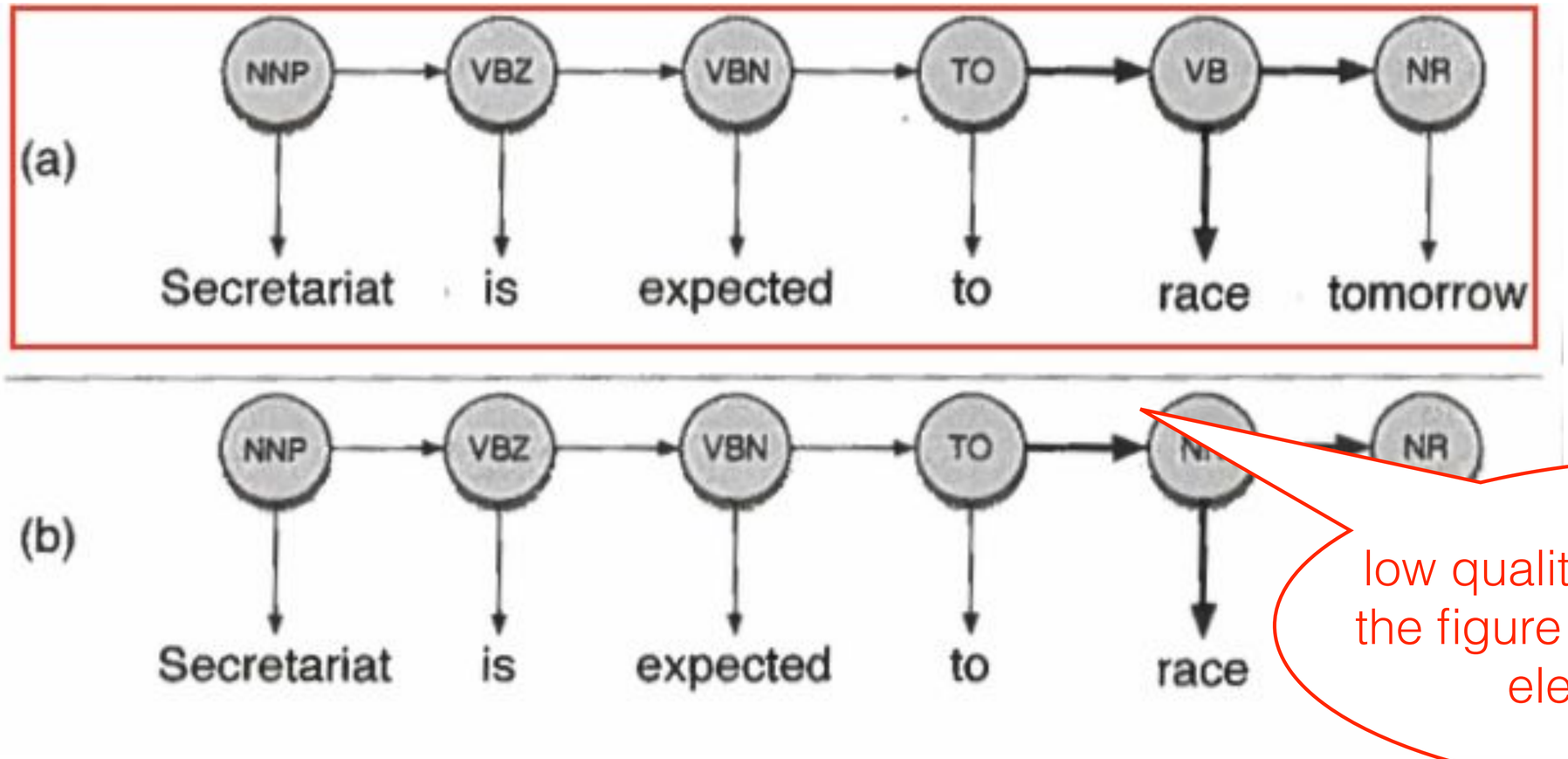
screenshot has been **stretched**

Fig. 2. Number of total disambiguations between part of speech tags [5]

Quality of reproduction



Quality of reproduction



low quality scan - at least copy the figure from a high resolution electronic original

Fig. 1: Two tag sequences for the sentence “Secretariat is expected to race tomorrow”.

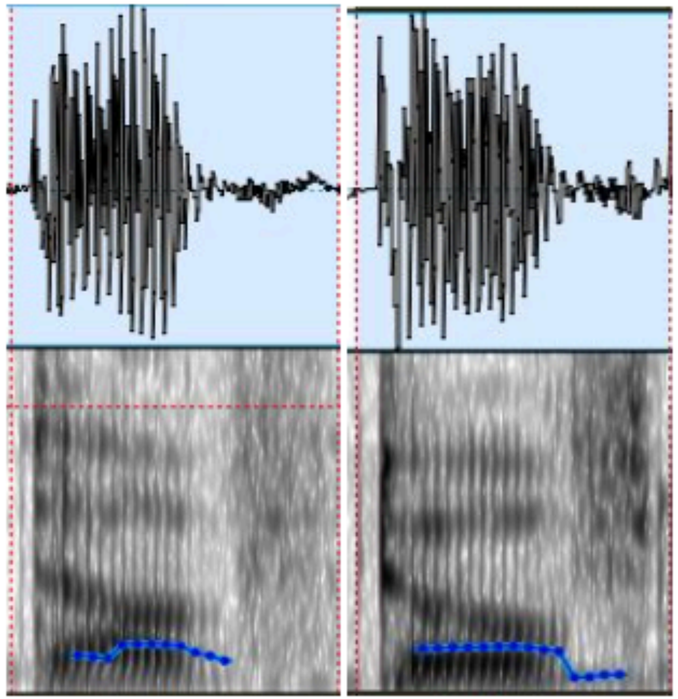
The more probable sentence is highlighted. (Jurafsky and Martin, 2009, 143)

Axes

For example:

'dove' (bird) nn (d uh v)

'dove' (action) vbd (d ou v). (Festival, 2004)



no axes

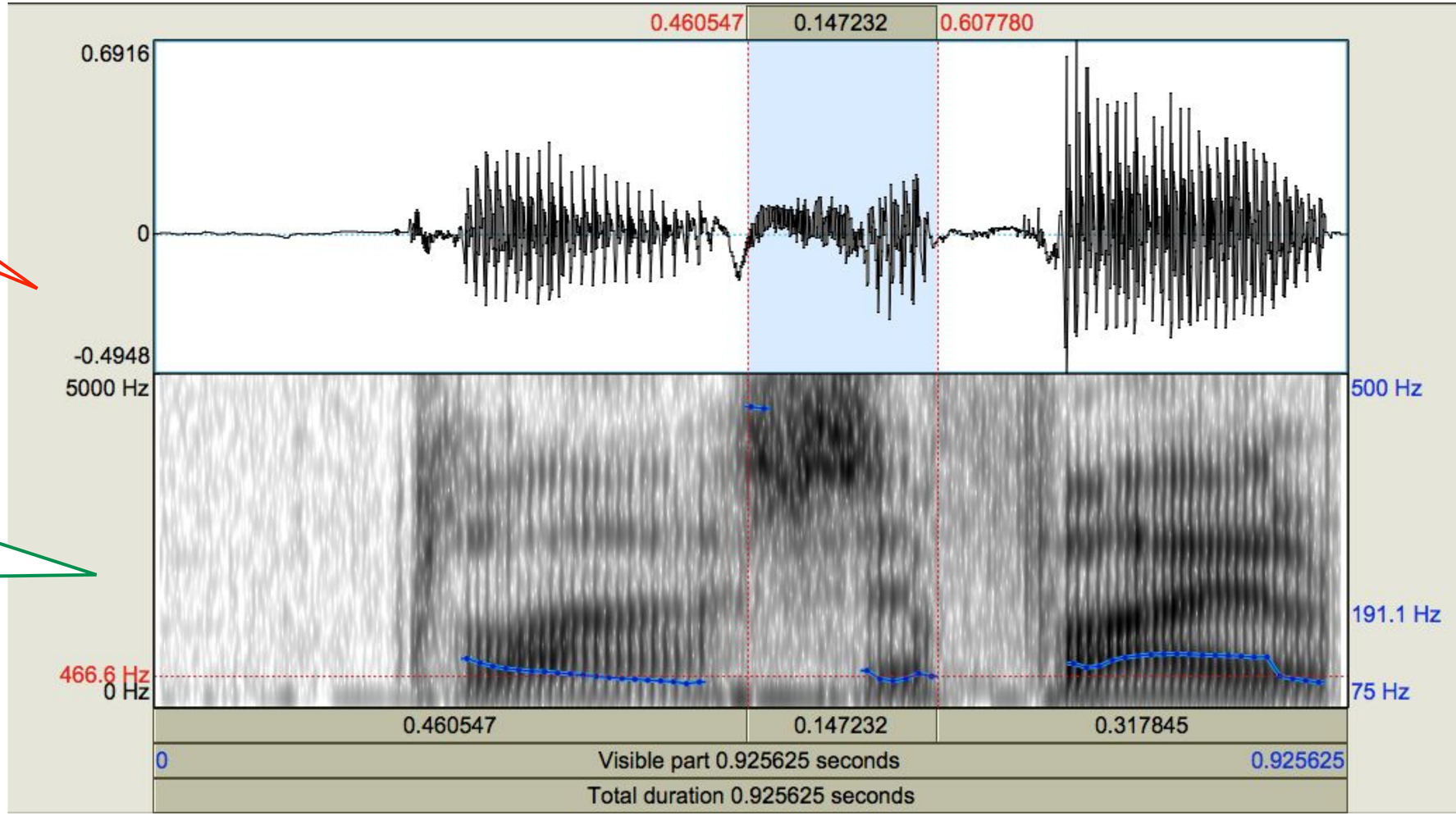
*Figure **2**: On the left, "dove", in the context of "It's a dove."*

On the right, "dove", in the context of "He dove down."

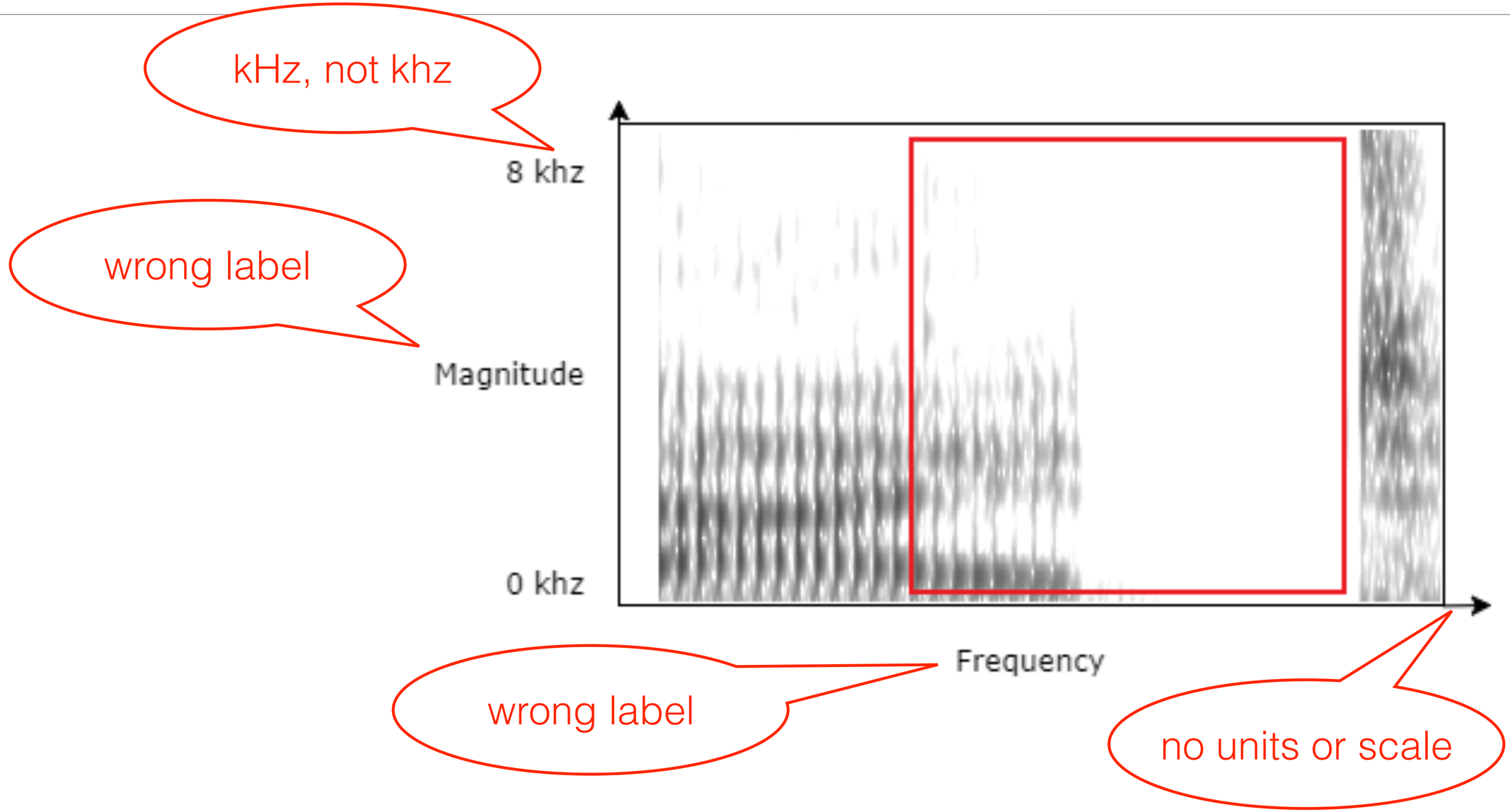
Axes

but you still need to **label** the axes

Praat shows the units and scale



Axes


kHz, not khz
8 khz
wrong label
Magnitude
0 khz
wrong label
Frequency
no units or scale

Axes

better to actually draw them on the figure

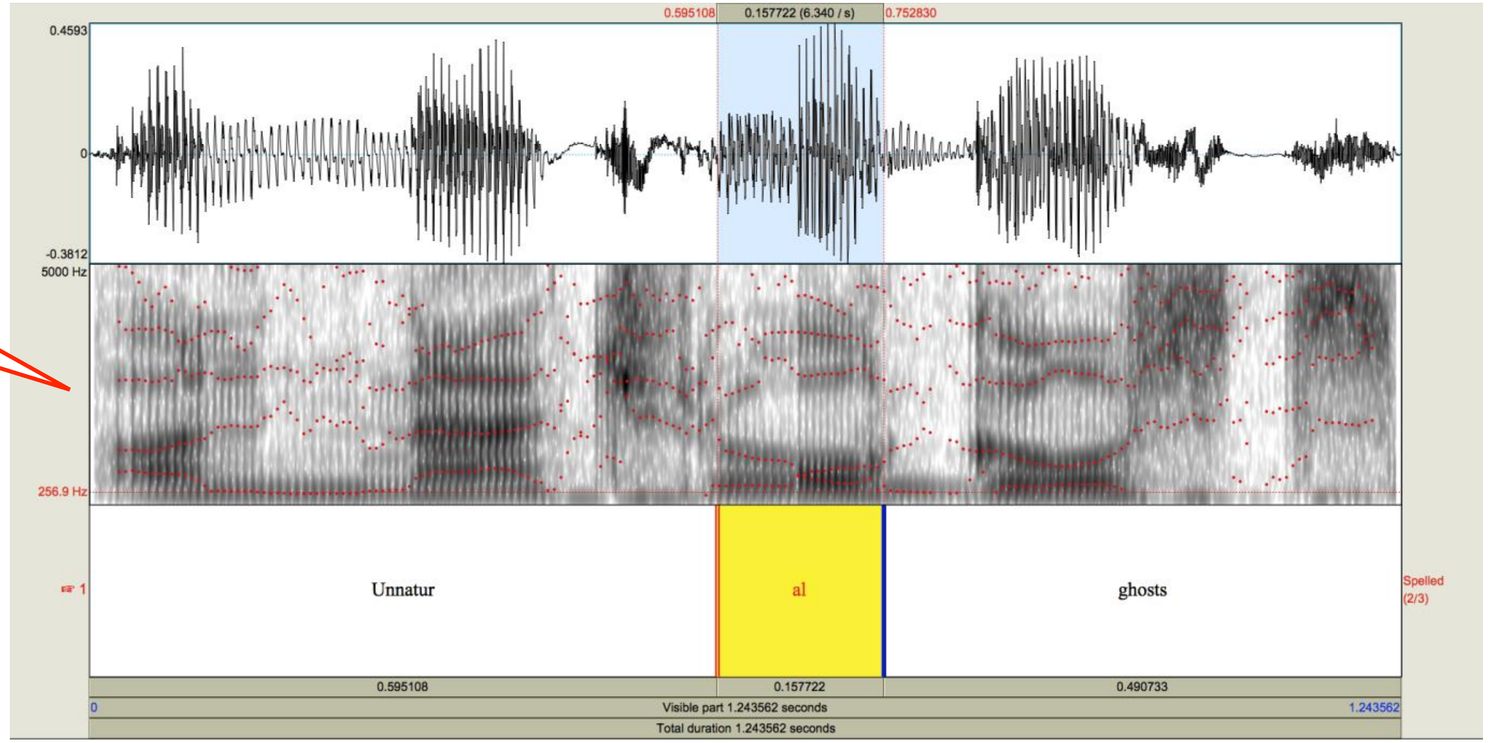


Figure 2: Waveform (y axis is amplitude) and spectogram (y axis is frequency) of 'unnatural ghosts', with the highlighted pa[illegible]ontaining the error. Notice the sudden increase in amplitude in the waveform

it's acceptable to describe the axes in the caption

Axes

"Group" is not explained

tables sometimes have "axes" too - here the vertical direction is presumably time

	Group 1	Group 2	Group 3
Previous Tag	DET	VBB	NOUN
Current Tag	NOUN	PRON	PUNC
Next Tag	VERB	PUNC	End Mark
Break	NB	B	BB

Table 2.2 It's hypothetical table which shows the break results collected from training data. PUNC means punctuation. End mark means symbol likes </s> to indicate the end of the sentence. But this is not the actual one from Festival

Axes

again, this table has axes, but they are not made clear

l_{i-1}	l_i	l_{i+1}	l_{i+2}	**PHONEME**
-	t	i	m	[t]
s	t	r	i	[t]
a	t	c	a	[t]
a	t	c	h	[ch]
a	t	h	e	[th]
i	t	c	h	[ch]

Figure 2.2: Toy training dictionary for LTS CART

Axes

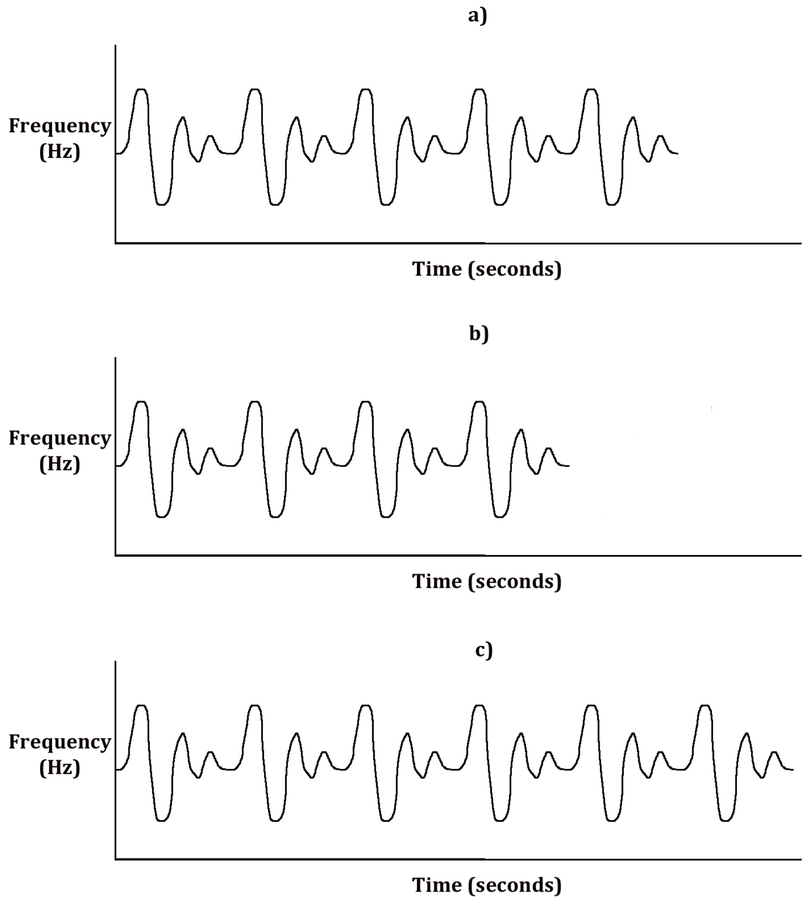
not clear if this is two tables stacked together, or what

Token	Word1	Word2	Word3
Break mark	NB	NB	BB
Token	Word4	Word5	Word6
Break	NB	NB	B

Table 4. Training data for phrase break prediction

Axes

wrong label

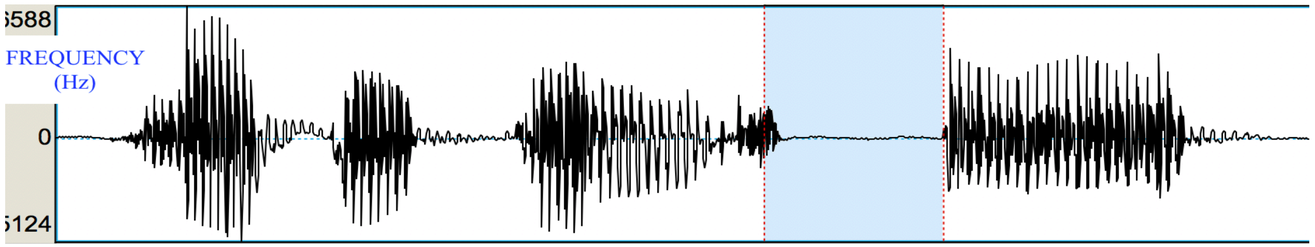


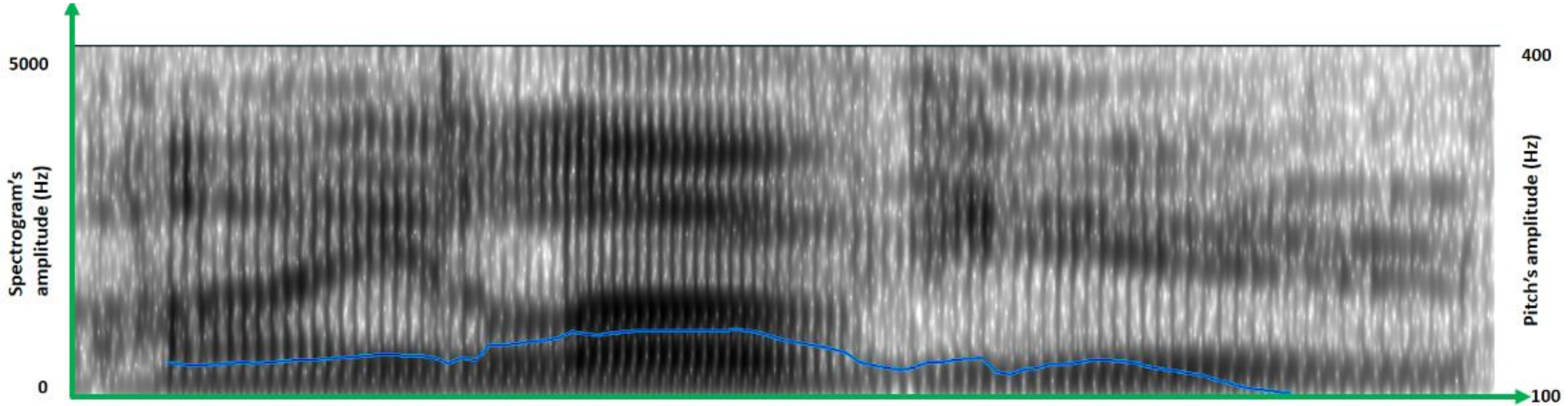
if you use **automatic** numbering and **cross-referencing**, this won't happen

Figure 4: Using TD-PSOLA to alter duration. Figure 5a is the signal being modified. Figure 5b shows a decrease in duration, Figure 5c shows an increase in duration.

Axes

more wrong labels





Axes

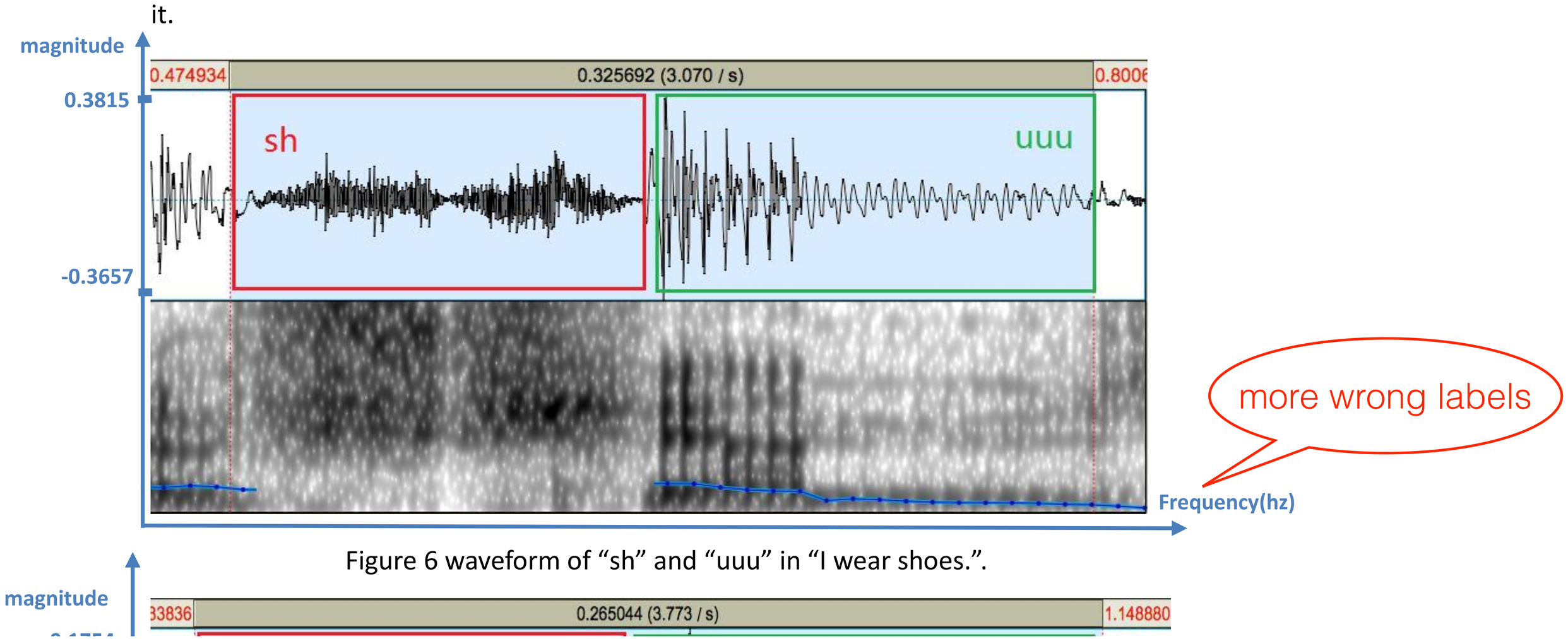


Figure 6 waveform of “sh” and “uuu” in “I wear shoes.”.

Axes

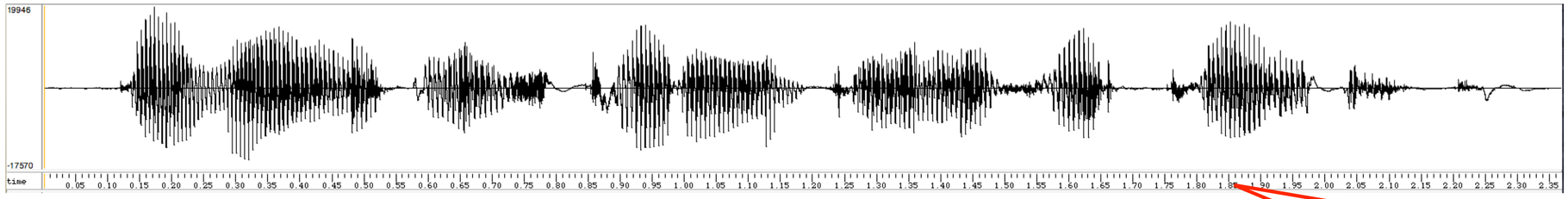


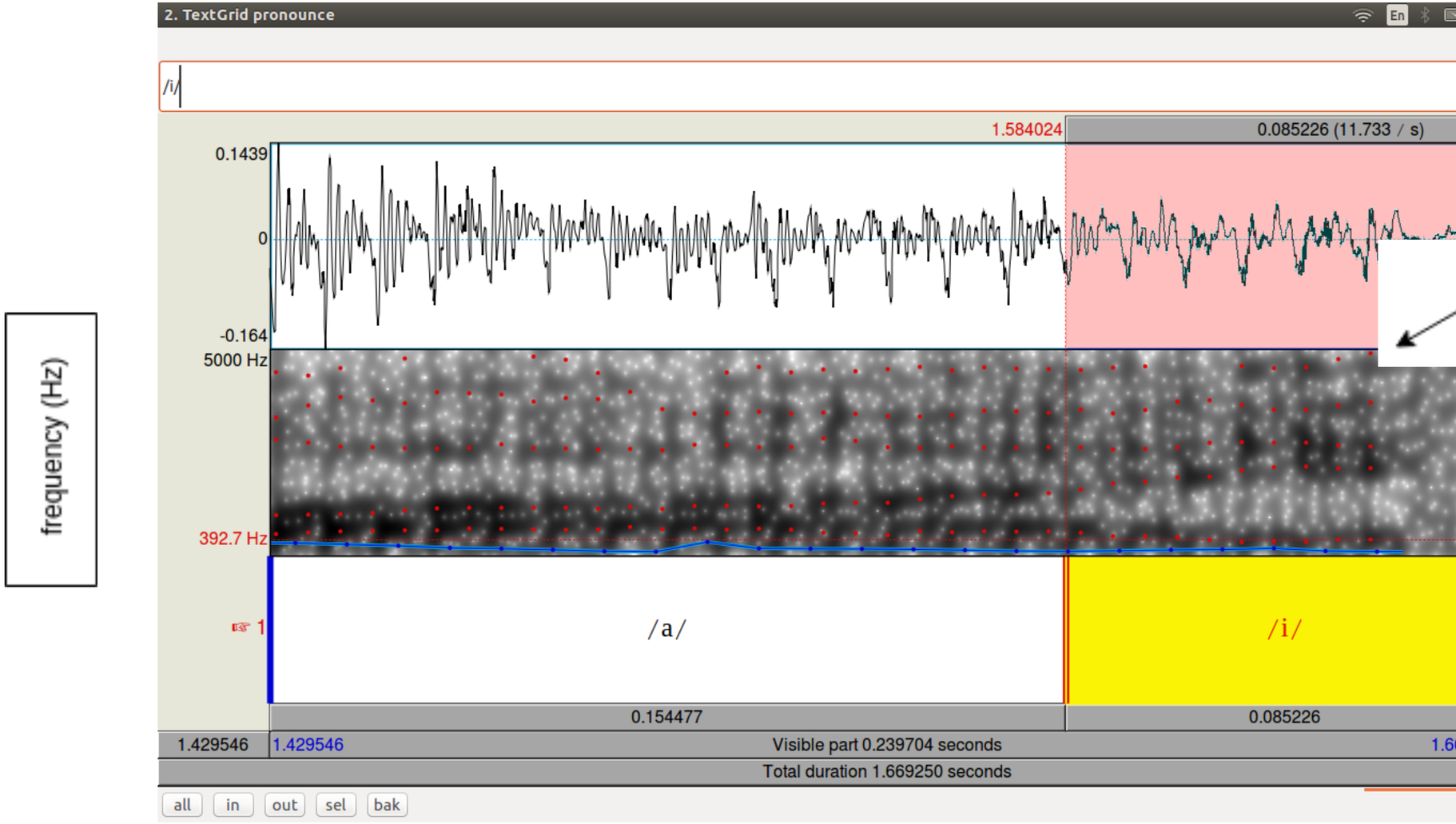
Figure 2: ”NLP is currently a hot topic” waveform

unreadable font size

The first bit of silence that can be seen in the sound wave in figure number 2 is between L and P whereas one would expect it to be between P and is. Moreover, the p_iii diphone is missing so Festival

Axes

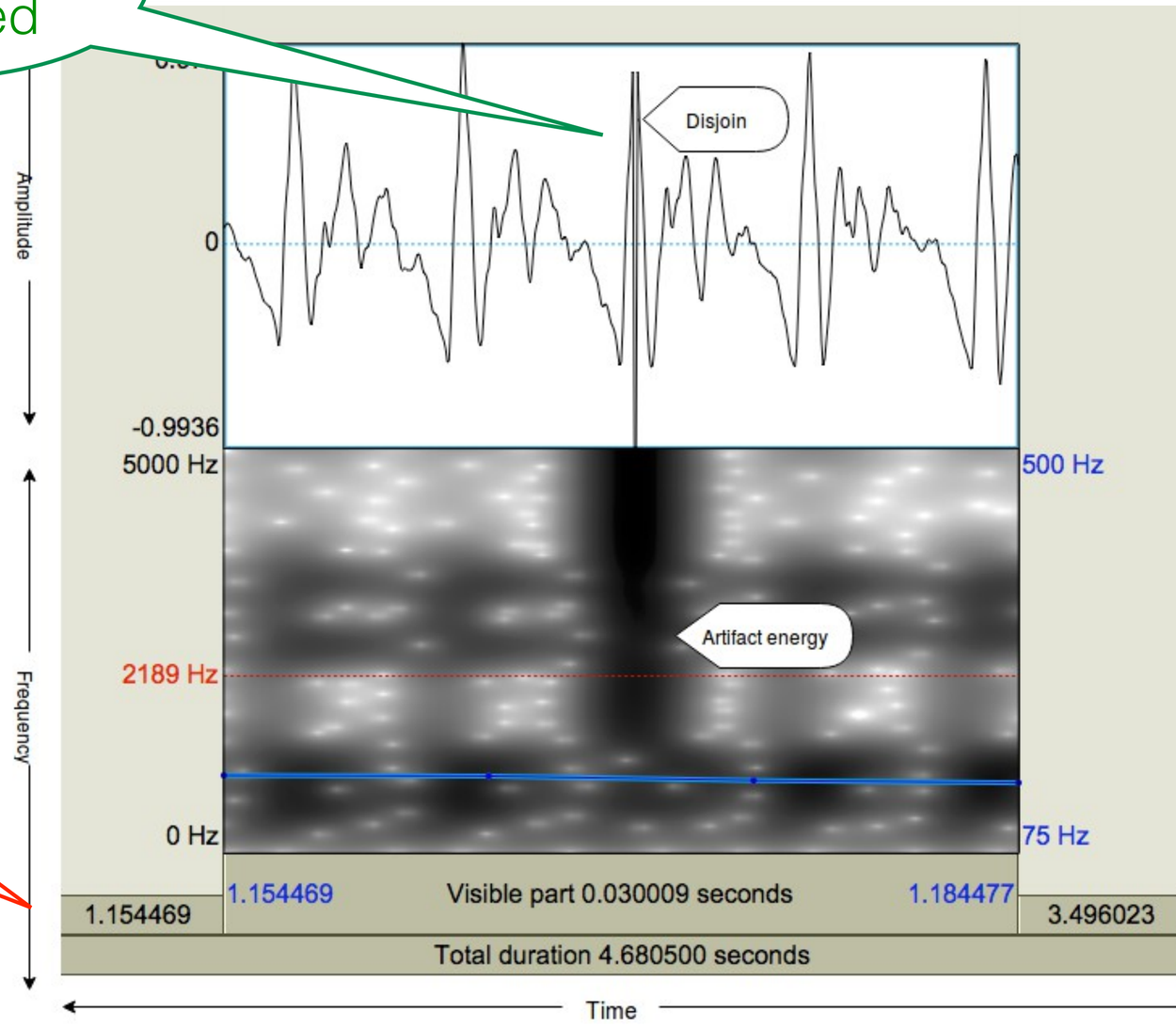
multiple plots in one - so need multiple axes



Axes

nicely annotated

arrow should be in one direction only, and align precisely with the spectrogram



Tables

ID	Type	Input
1.a	Metric	“This battery is 5**V**”
1.b	Tittle	“Kennedy **Jr.**”
1.c	Currency	“**$100bn**”
1.d	Degree Honour	“I achieved a **2.1**”
1.e	Degree title	“**BA** Law”

compact list of examples, with identifiers so they can be referred to in the text

but make sure you **do** refer to them

Tables

OK to have very brief column headers to achieve a compact format

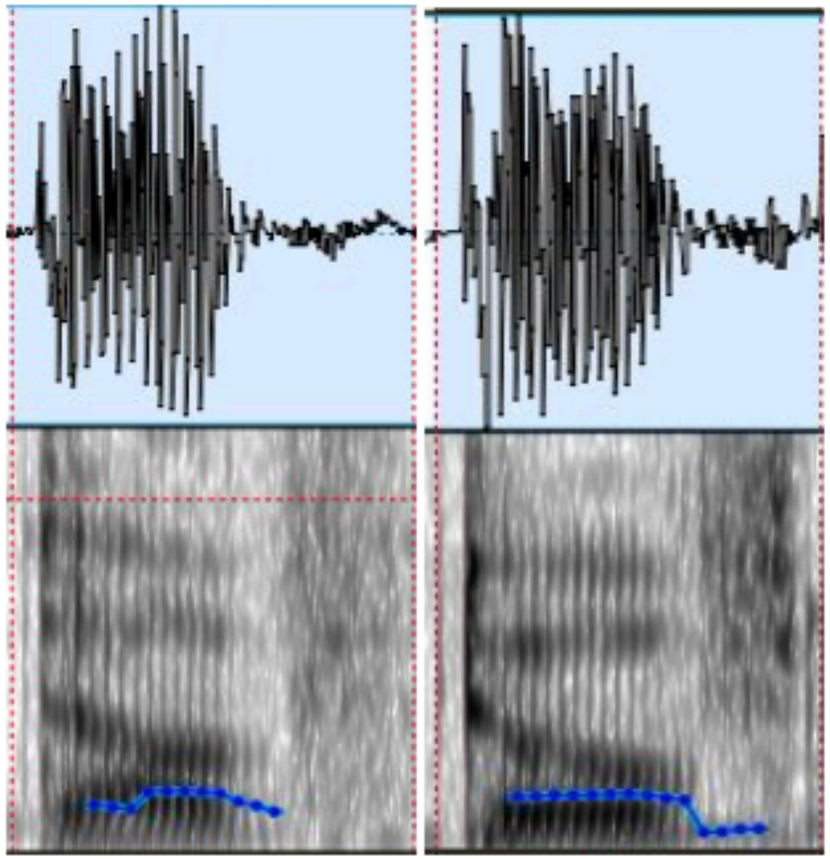
good use of a table

NSW	**Expected output**	**Festival output**	**Error?**
£3.45	*three pounds forty five*	*three pounds* ***dot*** *forty five*	Yes
$3.45	*three dollars forty five*	*three dollars forty five*	No

Figure 2. *Expected and Festival output for two NSWs differing only i* *bol. Errors are* ***boldfaced.***

but should explain such headers in the caption (here, the acronym "NSW")

Captions



descriptive caption

*Figure **2**: On the left, "dove", in the context of "It's a dove."*

On the right, "dove", in the context of "He dove down."

Captions

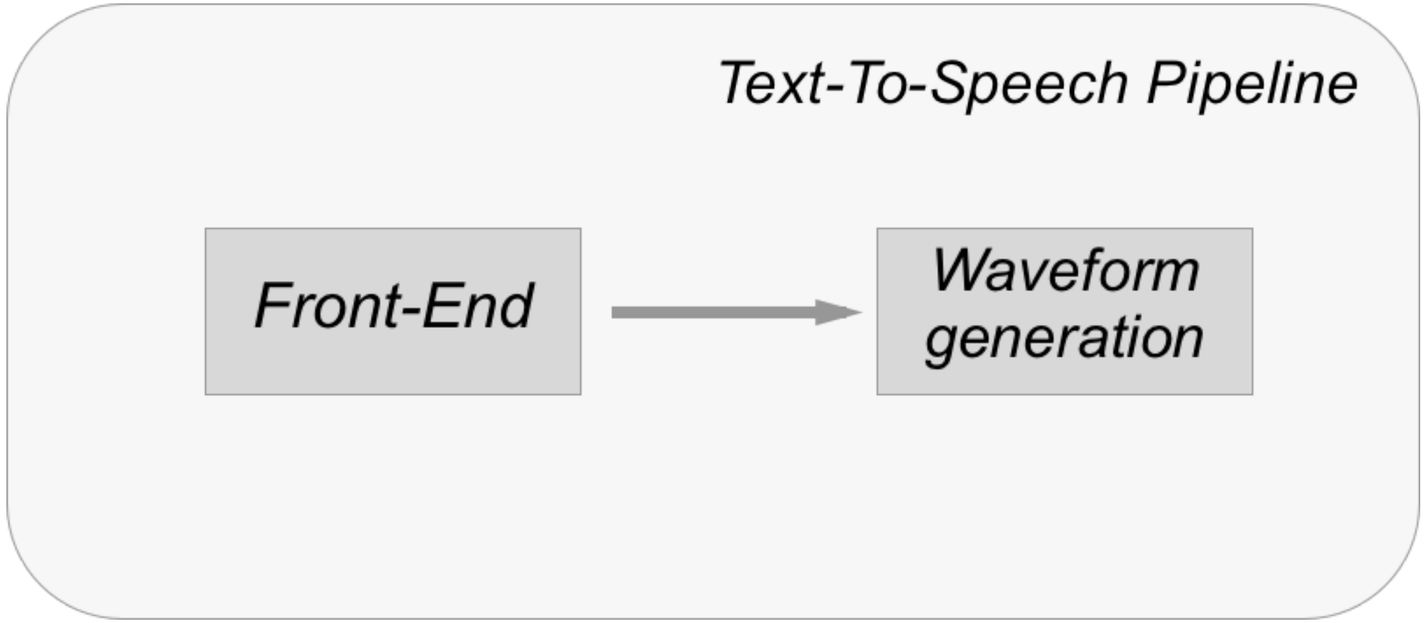
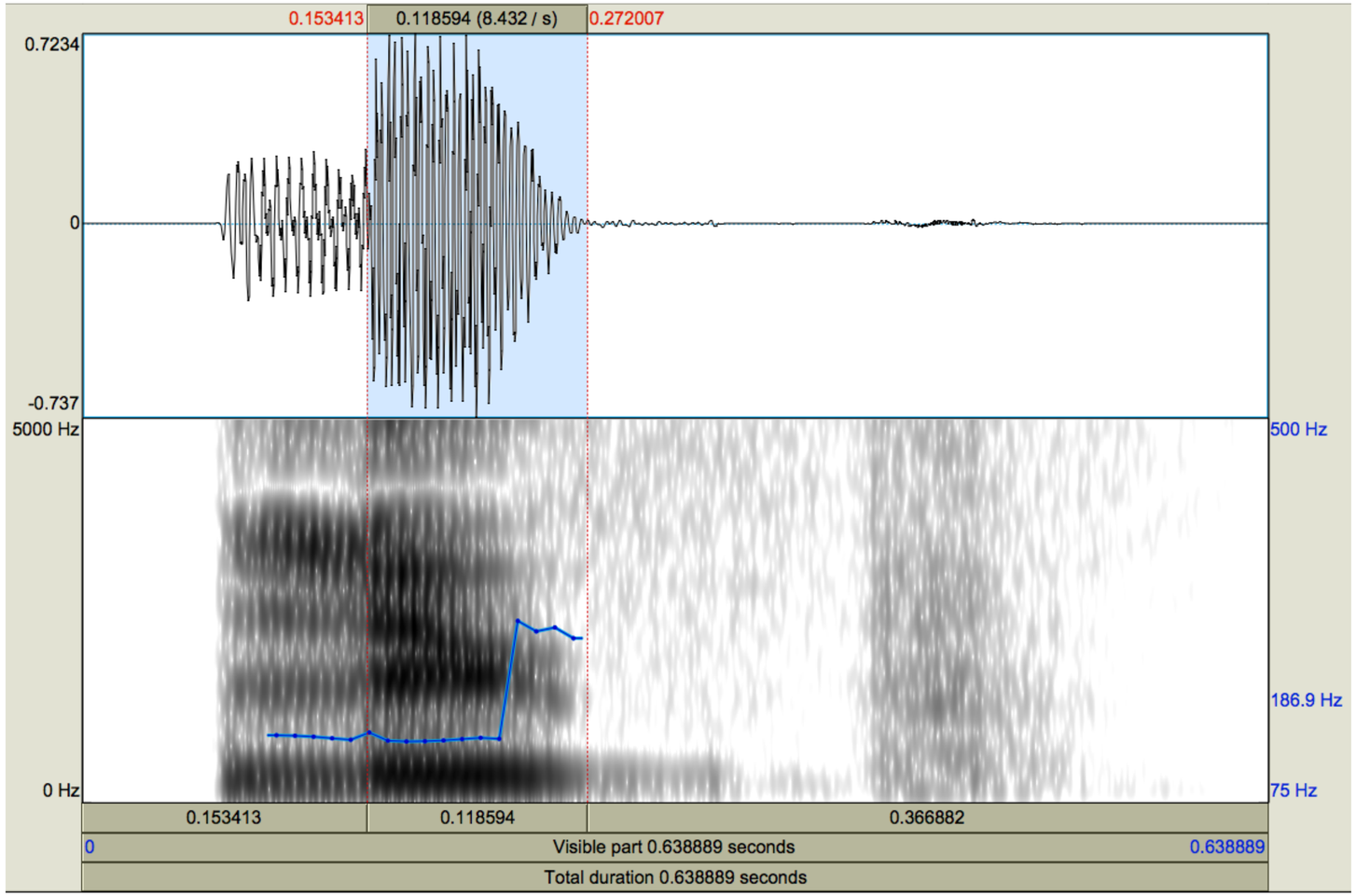


Figure 1

no caption

Captions



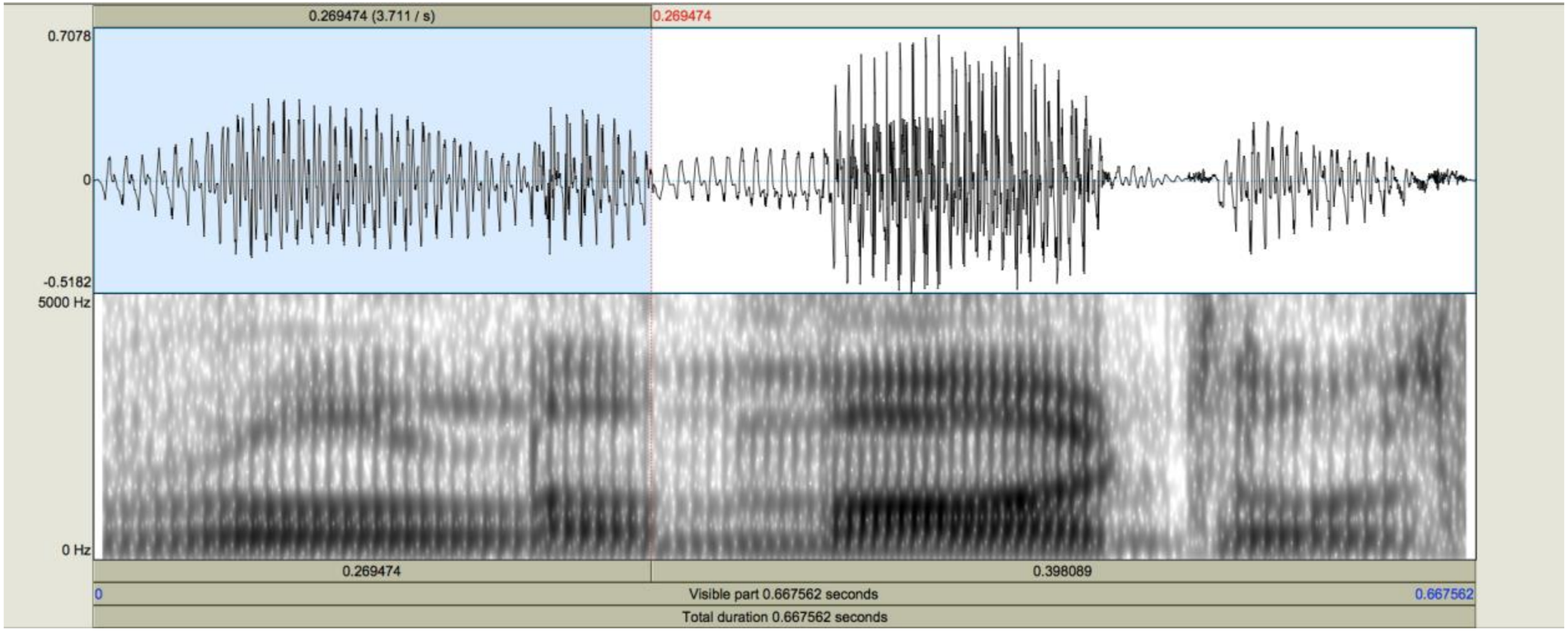
no caption

Figure 6:

Captions

caption too brief
and cryptic

Fig.2 “role model”



Captions

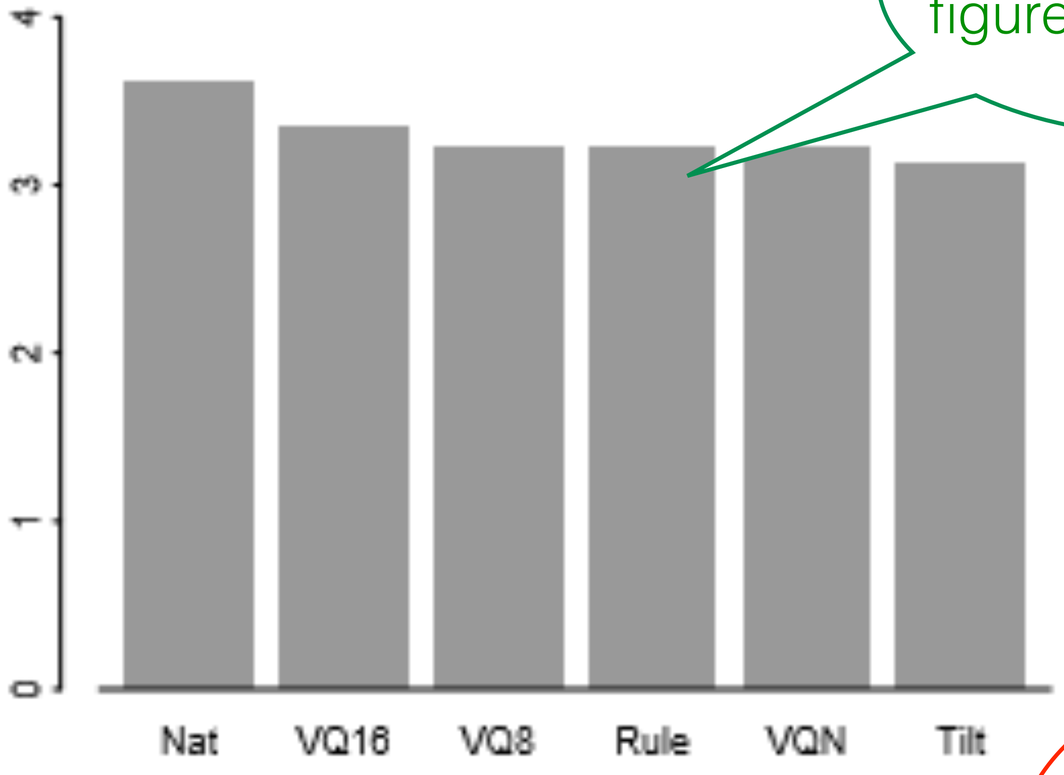


Figure 1: Mean Ratings of Prosody

it's OK to quote a figure from a paper

but provide your own caption, numbered correctly in sequence, and in the caption cite the source of the figure

Captions

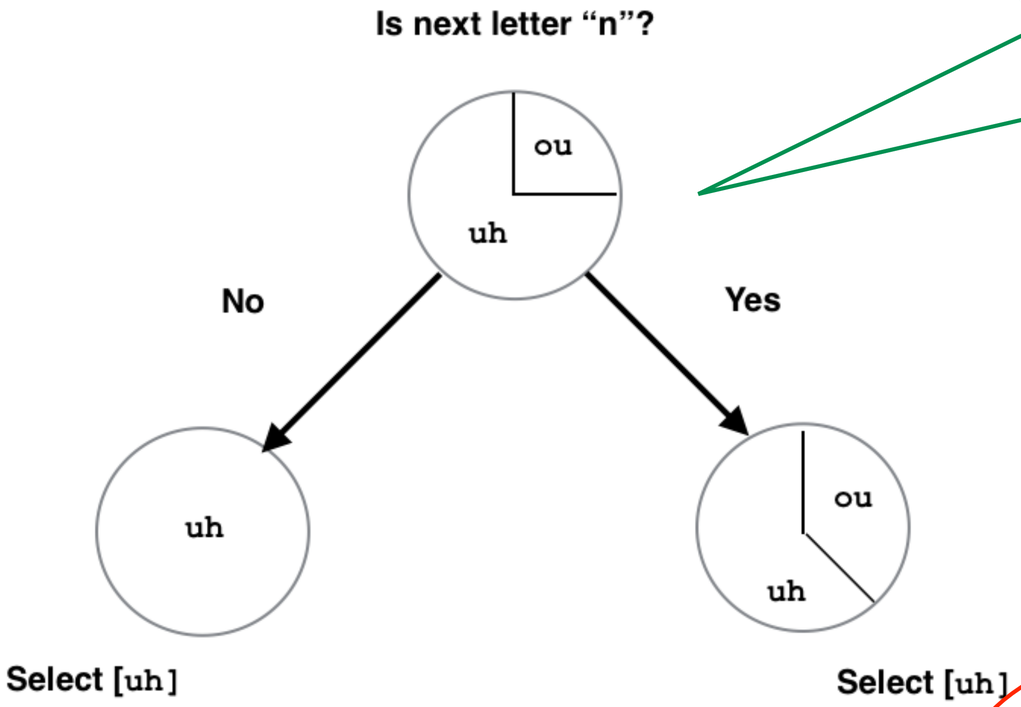
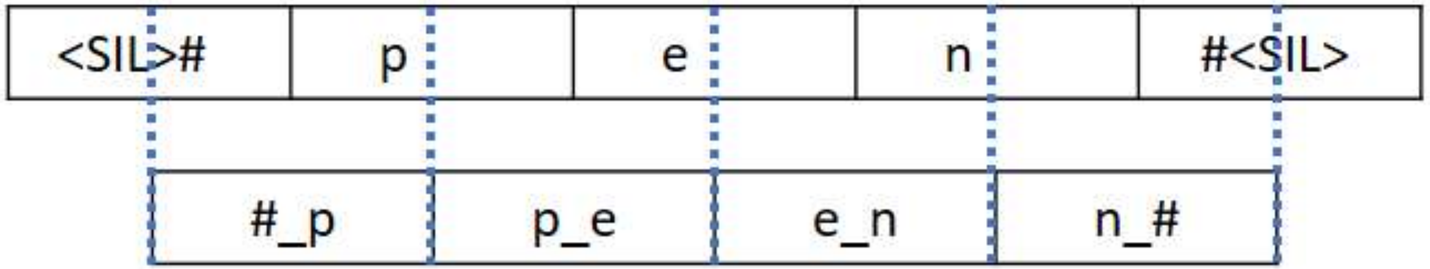


Figure 3: Example of training data for the letter “o”

a great figure,
demonstrates understanding of
how training data are used by a
CART

but use the caption to
explain the notation (pie charts,
phoneme symbols)

Captions



good diagram

Figure modelling the structure of diphones from phones.

Figure should be numbered so you can refer to it in the text.

Say which is which

Captions

it's clear which is which here

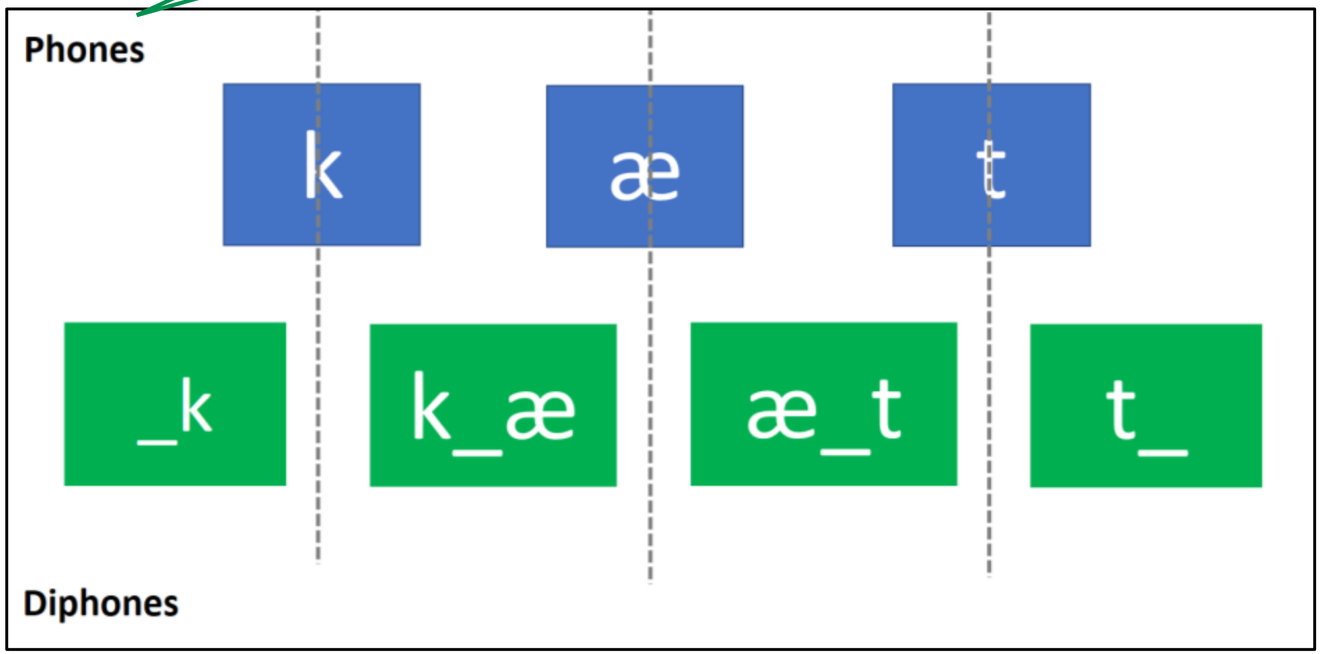


Figure 2: Diphones

too short

Captions

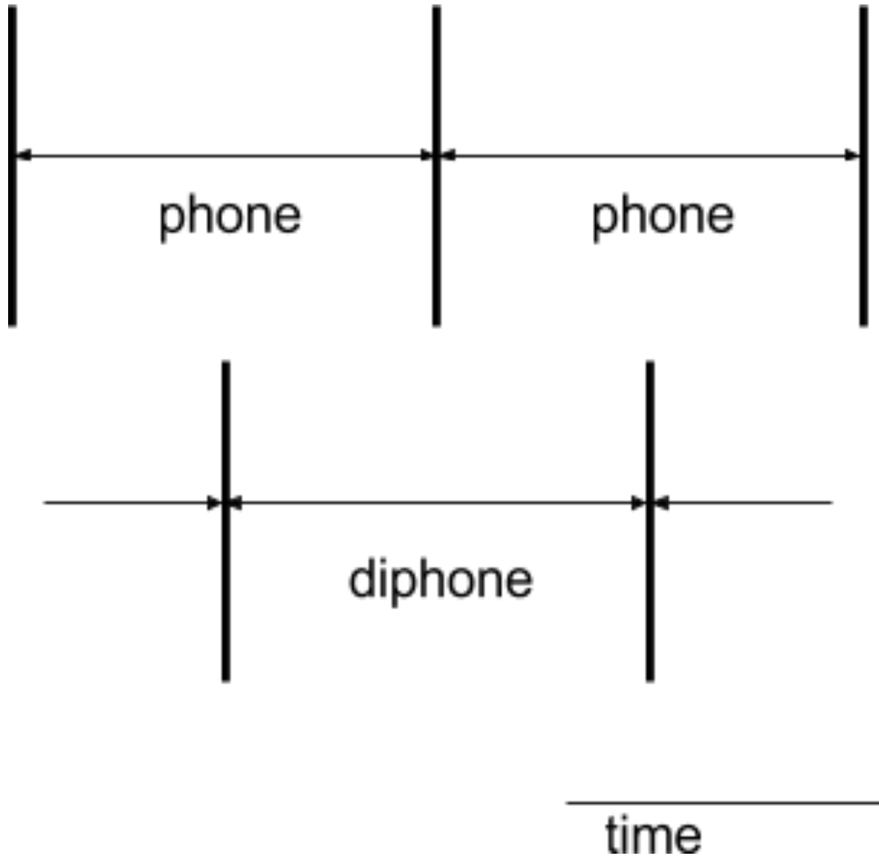


Figure 3. Diphone units.

very similar to a figure from lecture slides - should cite the original

caption rather terse

Captions

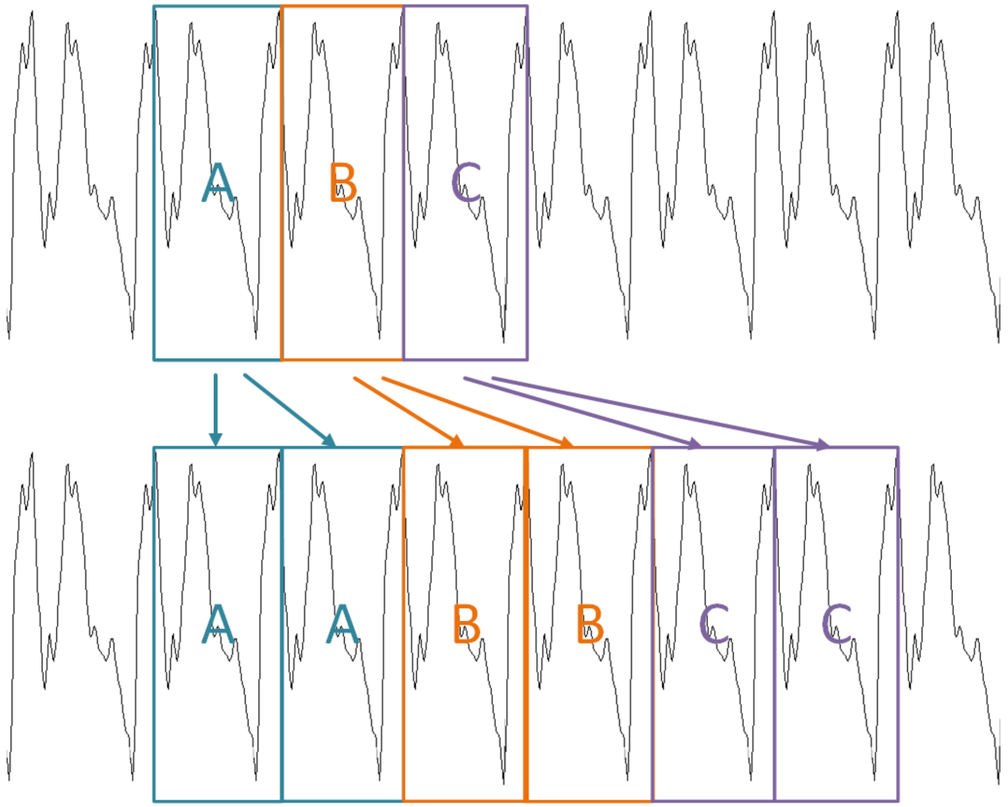
$$\vec{t}_1^n = argmax \prod_{i=1}^{n} P(w_i|t_i)P(t_i|t_{i-1})$$

Equation 1: HMM equation for POS tagging (Jurafsky and Martin 2009, 5.5)

equations are usually just numbered (on the right), rather than having captions

correct to cite original, although is “5.5” a section or an equation number?

Captions



correct to cite the original

except this is not quite the same, so should have said "based on…"

Figure 2: Example given by Jurafsky and Martin (2009, p. 310)) to increase duration of a diphone by adding "pitch-synchronous frames" (Jurafsky and Martin (2009, p. 310)).

Captions

Considering the following examples:

(1) "She is from Northern Ireland."

(2) "She wears leathern bracelets."

One would expect letter sequence *'-thern'* in *'Northern* same pronunciation in the dictionary. This is not th received /r n!/ as the last sequence, and *'leathern'* /dh causes Festival to back off in the waveform generati incorrect pronunciation. This could be fixed by altering entry.

good idea to number examples

3.4.2. Letter-to-sound

(1) *"He ate a ciabatta"*

(2) *"He said ciao and left."*

but potentially confusing to reset numbering - better to continue in sequence, or prepend the section number, etc

Spectrograms vs waveforms

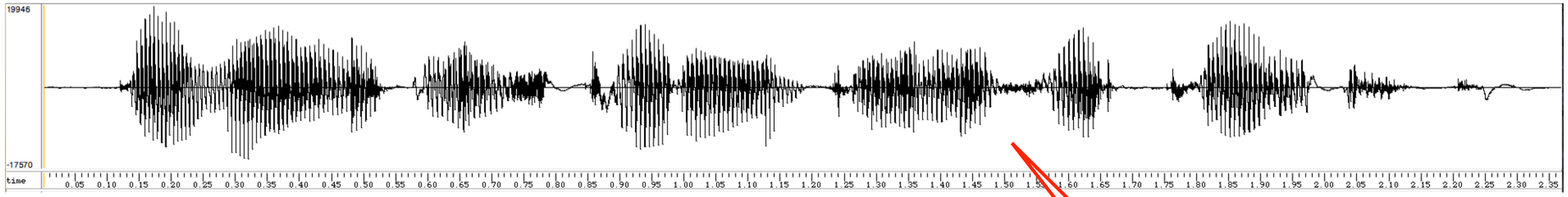


Figure 2: ”NLP is currently a hot topic” waveform

no annotations

Spectrograms vs waveforms

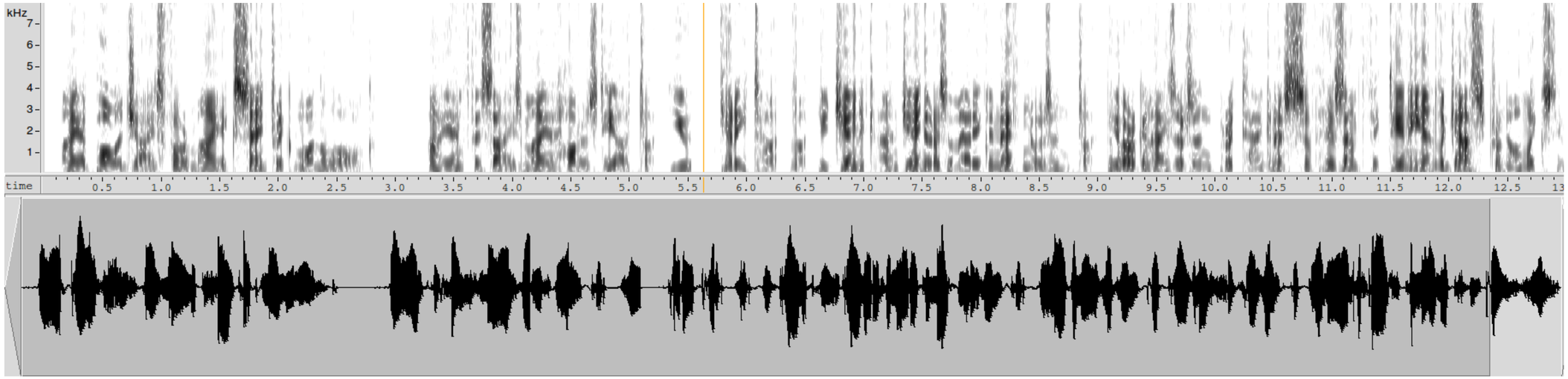


Figure 3: Corresponding spectogram and waveform for sentence 1.

what point
is being made?

Spectrograms vs waveforms

axes and units correct

but no scale

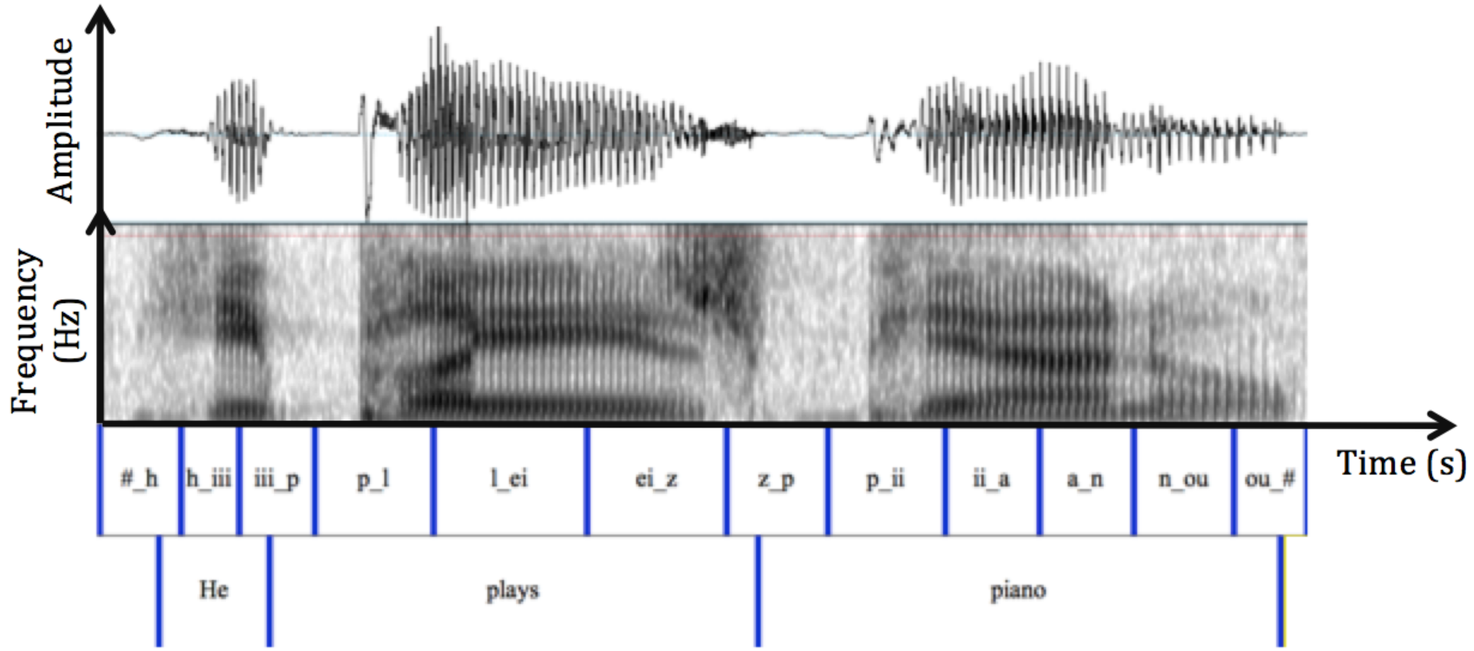
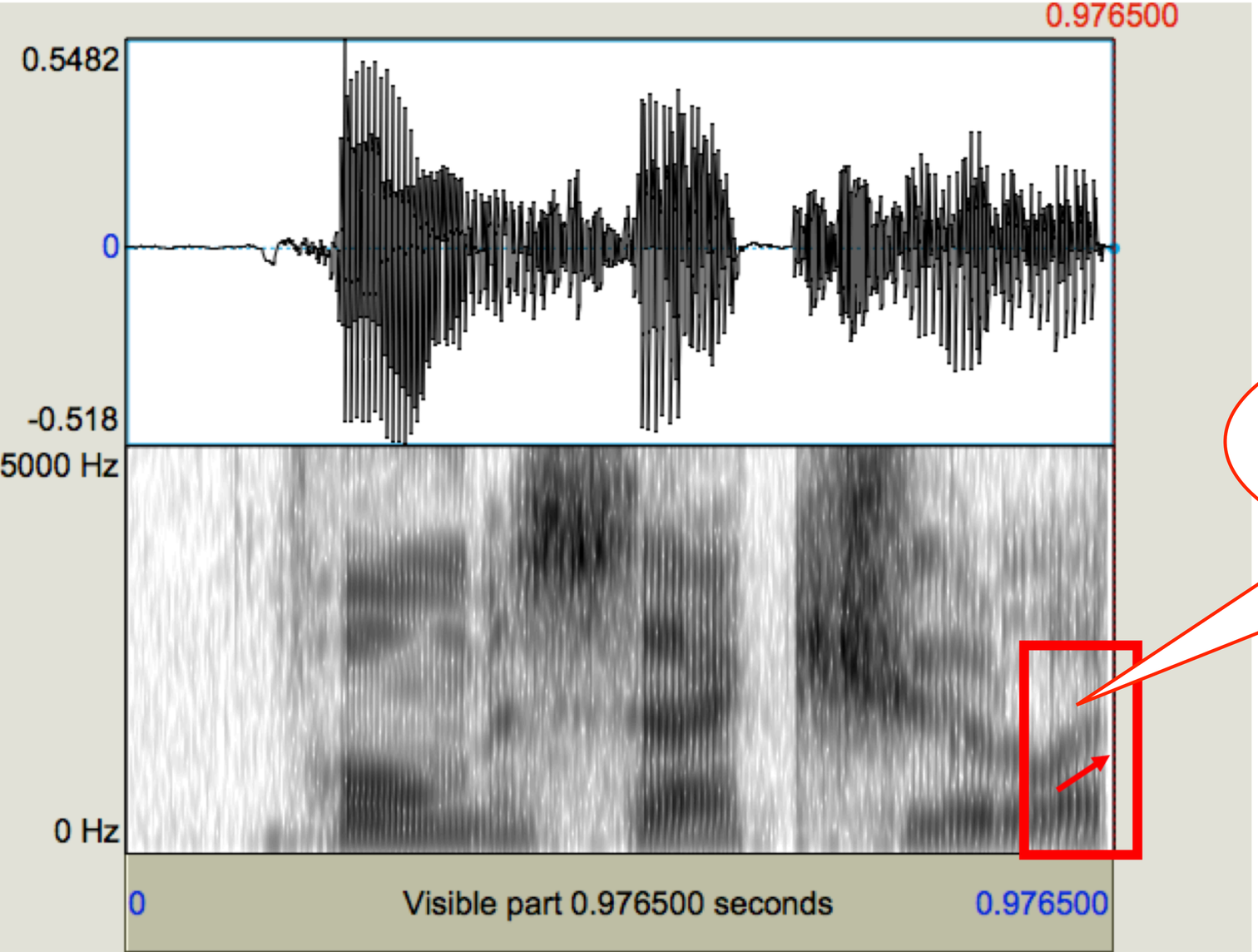


Figure 2: A waveform and spectrogram of the example sentence, "He plays piano", annotated into words and diphones.

good caption

Spectrograms vs waveforms

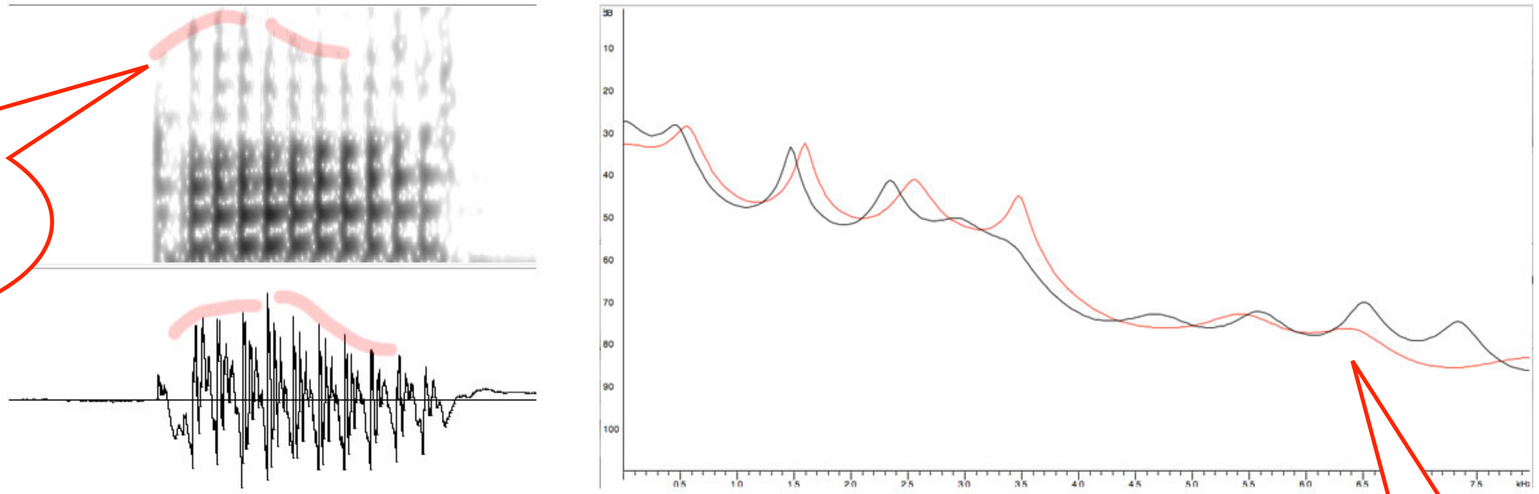


could have tweaked spectrogram settings to make formant more obvious

Spectrograms vs waveforms

potentially a great figure

but what are these lines?



and what's the difference between the red and black spectra?

Fig. 2. Left graph showing the waveform. right graph the LP gram.

Spectrograms vs waveforms

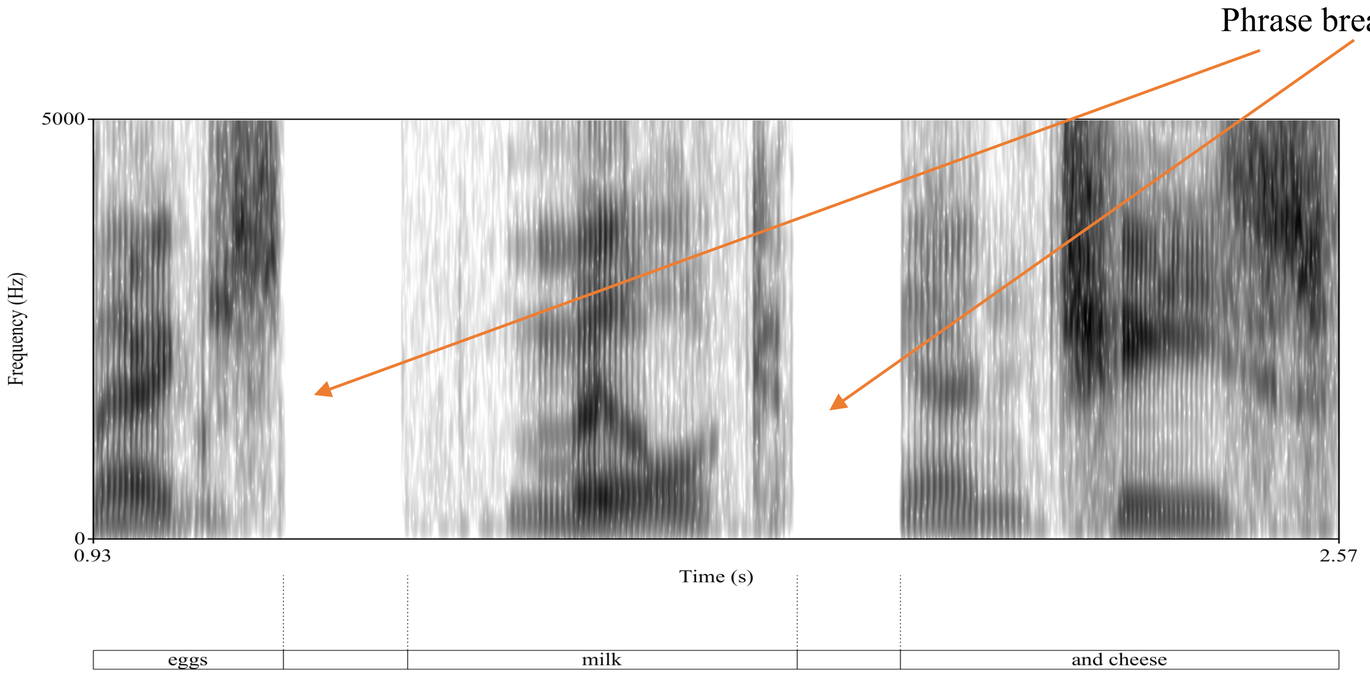
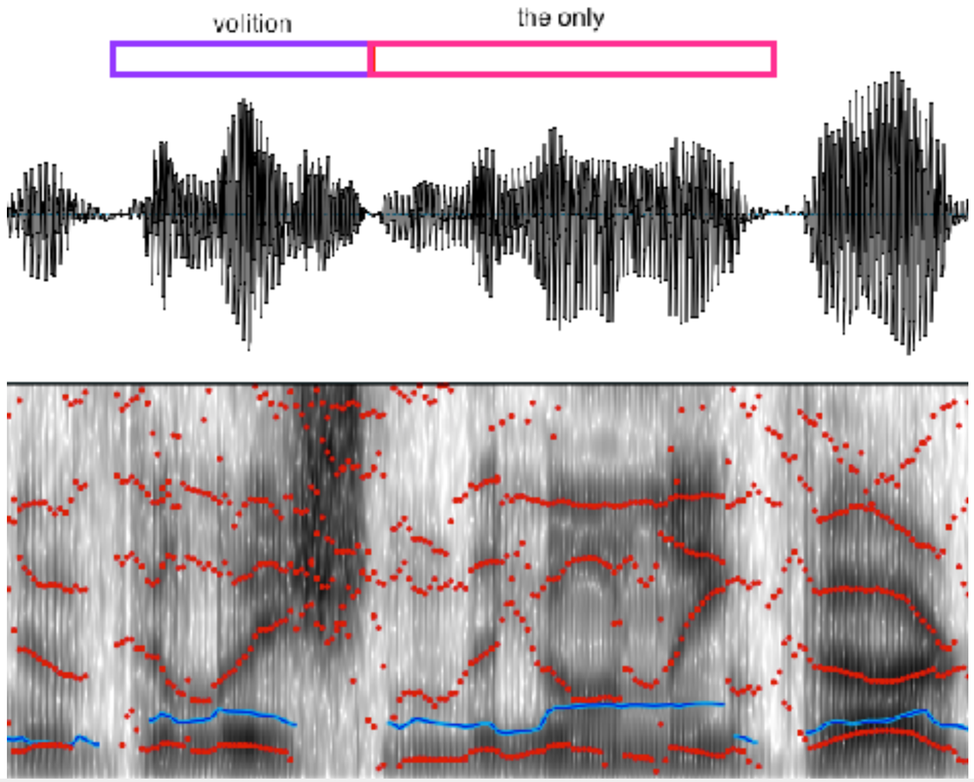


Figure 3: Spectrogram for Festivals output of "eggs, milk, and cheese"

is all of this really necessary just to show where the phrase breaks are?

Spectrograms vs waveforms



Voila! […] violation of volition! The only verdict […]

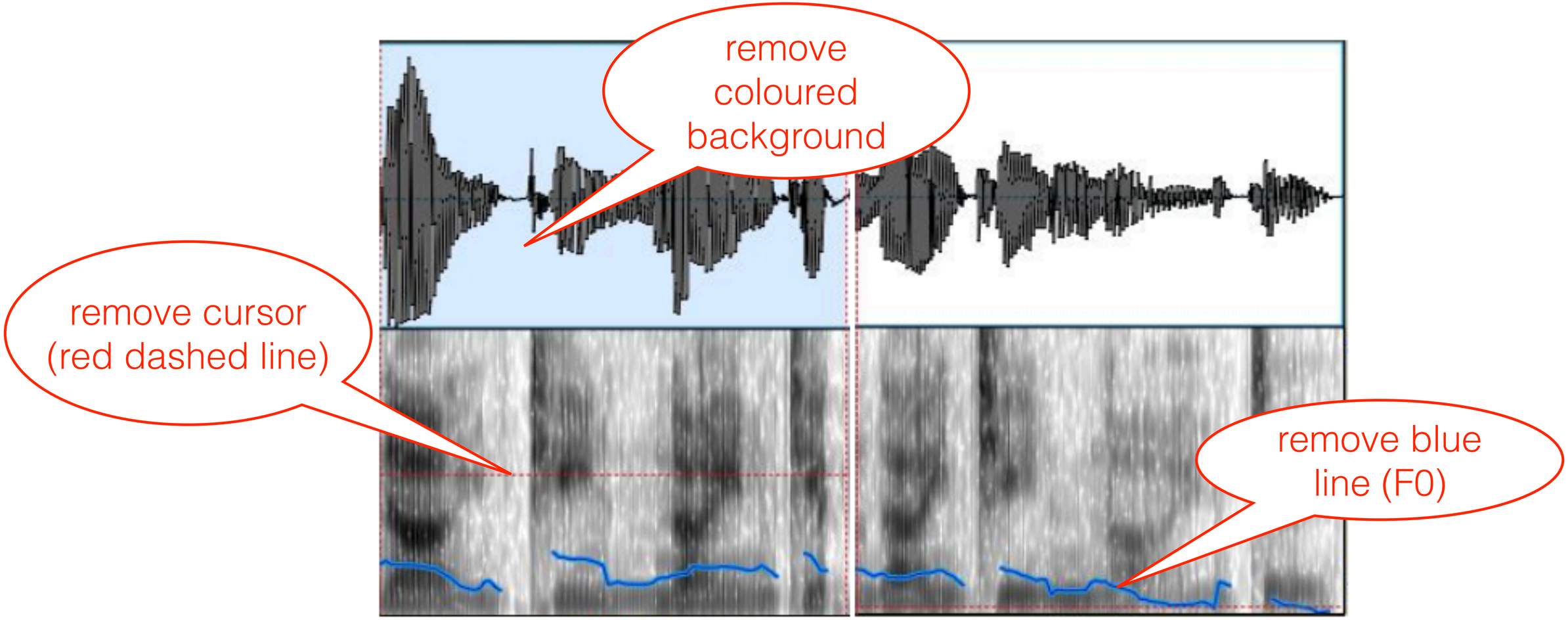
“Voila […] violation of volition the only verdict […]”

FIG. 07

A speech from the 2005 movie *V for Vendetta*, once fed into Festival, came out pretty accurately, except that Festival disregarded an exclamation mark altogether (Fig. 07). This is a deeper problem, though, as from the context, this

does this really demonstrate that “*Festival disregarded an exclamation mark*” ?

Extraneous information detracts from the point you're making



Extraneous information detracts from the point you're making

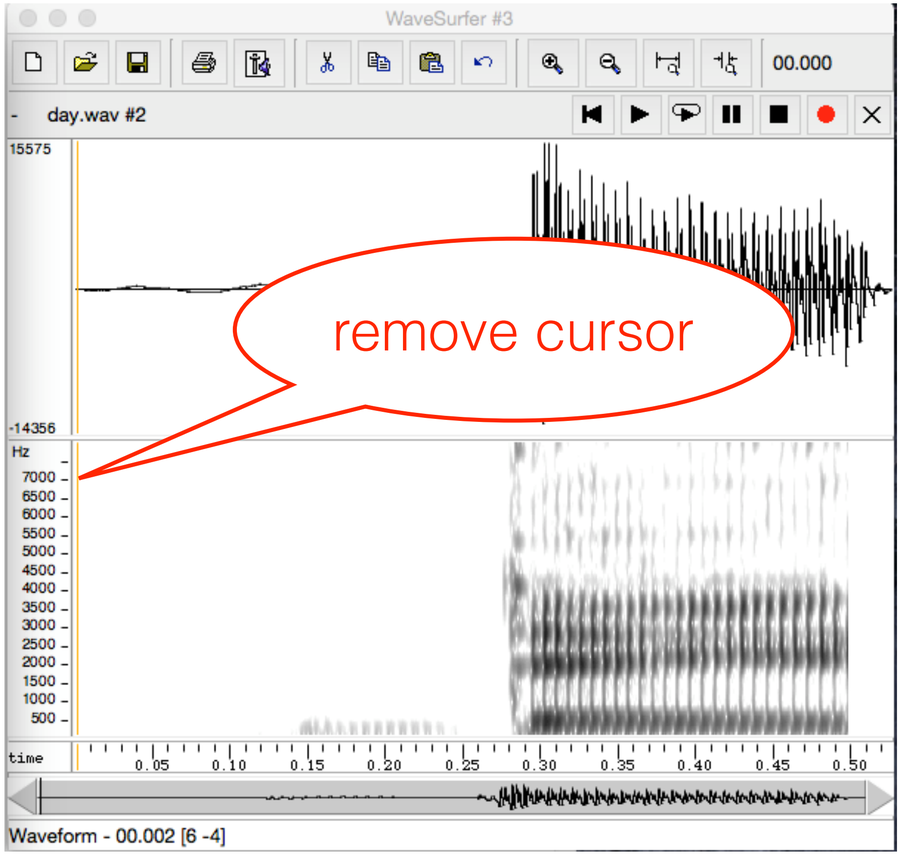


Figure 1: Waveform of "day"

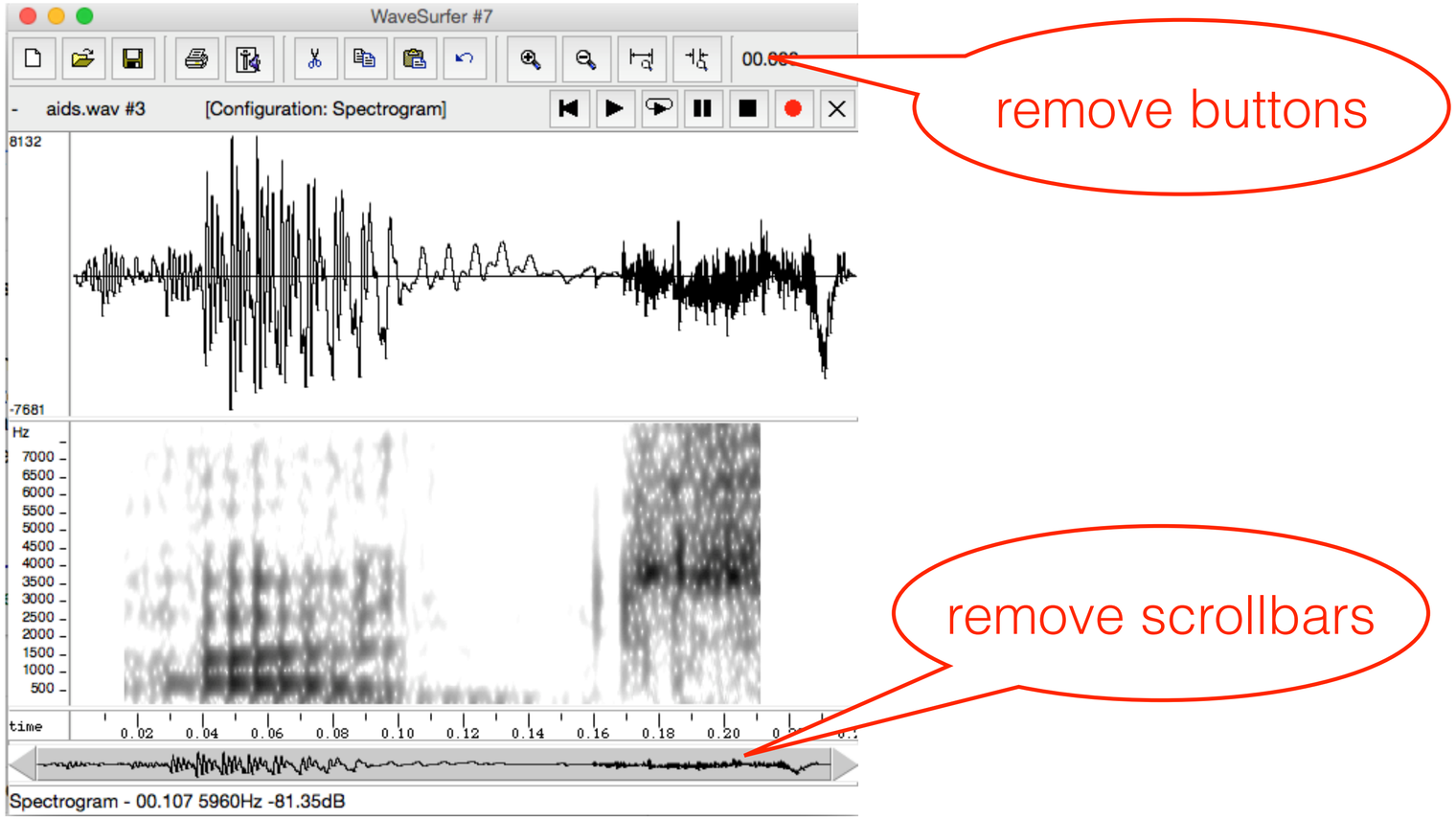


Figure 2: Waveform of "AIDS", as produced by Festival

Extraneous information, detracts from the point you're making

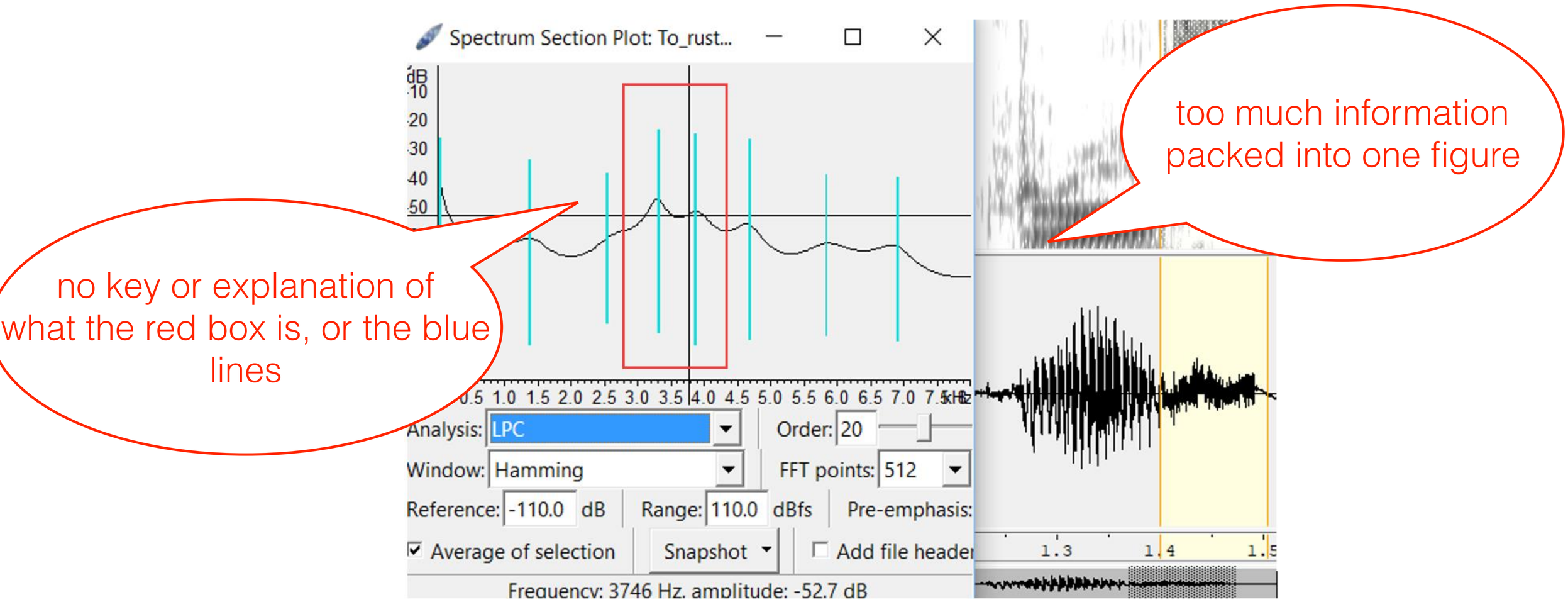
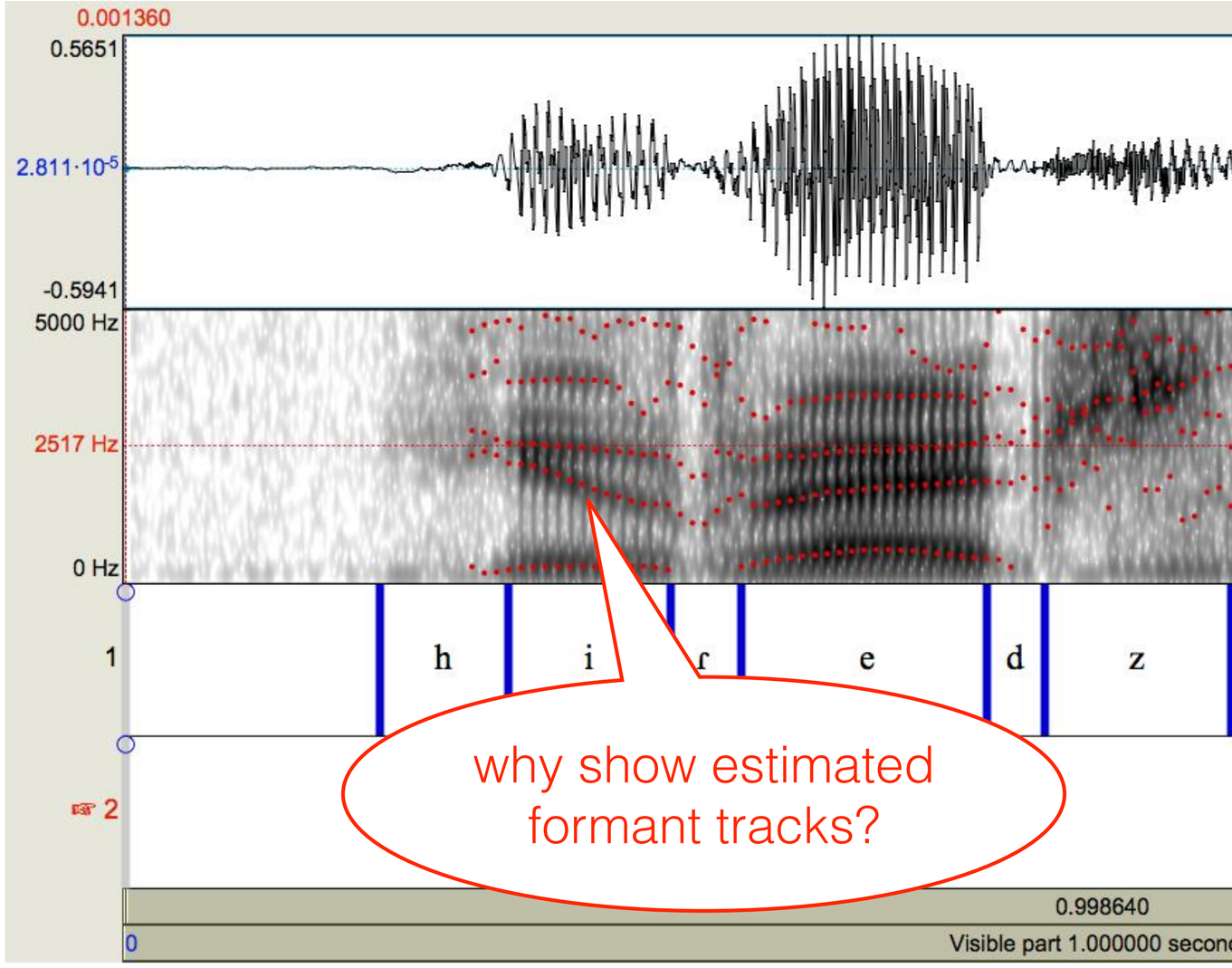


Figure 2.a: Highlighted section of "rust" had discontinued F0 in the spectrum slice.

Extraneous information, detracts from the point you're making

what are all the numbers?



Extraneous information, detracts from the point you're making

```
festival> (set! myutt7 (SayText "Mum works 24/7"))
  inserting pause after: z.
  Inserting pause
()
id _4 ; name Mum ; pos_index 5 ; pos_index_score 0 ; pos jj ; pbreak NB ;
id _5 ; name works ; pos_index 15 ; pos_index_score 0 ; pos vbz ; pbreak NB
id _6 ; name twenty ; pos_index 7 ; pos_index_score 0 ; pos cd ; pbreak NB
id _7 ; name four ; pos_index 7 ; pos_index_score 0 ; pos cd ; pbreak NB ;
id _8 ; name seventh ; pos_index 5 ; pos_index_score 0 ; pos jj ; pbreak NB
id _9 ; name 's ; pos nnp ; pos_index 2 ; pos_index_score
()
id _43 ; name # ;
id _12 ; name m ;
id _13 ; name uh ;
id _14 ; name m ;
id _16 ; name w ;
id _17 ; name @@r ;
id _18 ; name r ;
id _19 ; name k ;
id _20 ; name s ;
id _22 ; name t ;
id _23 ; name w ;
id _24 ; name e ;
id _25 ; name n ;
id _27 ; name ? ;
id _28 ; name ii ;
id _30 ; name f ;
id _31 ; name our ;
id _32 ; name r ;
id _34 ; name s ;
id _35 ; name e ;
id _37 ; name v ;
id _38 ; name n! ;
id _39 ; name th ;
id _41 ; name @ ;
id _42 ; name z ;
id _44 ; name # ;
Missing diphone: n!_th
 diphone still missing, backing off: n!_th
 backed off: n!_th -> n_th
#<Utterance 0x1d3e080>
```

all of this verbatim output, just to illustrate one small point

Seven after slash in 24/7 expanded as seventh's

figure has the phone sequence annotated, but this is an example of incorrect **text** normalisation

Figure 3.1 Inaccurate expansion of 24/7

Extraneous information, detracts from the point you're making

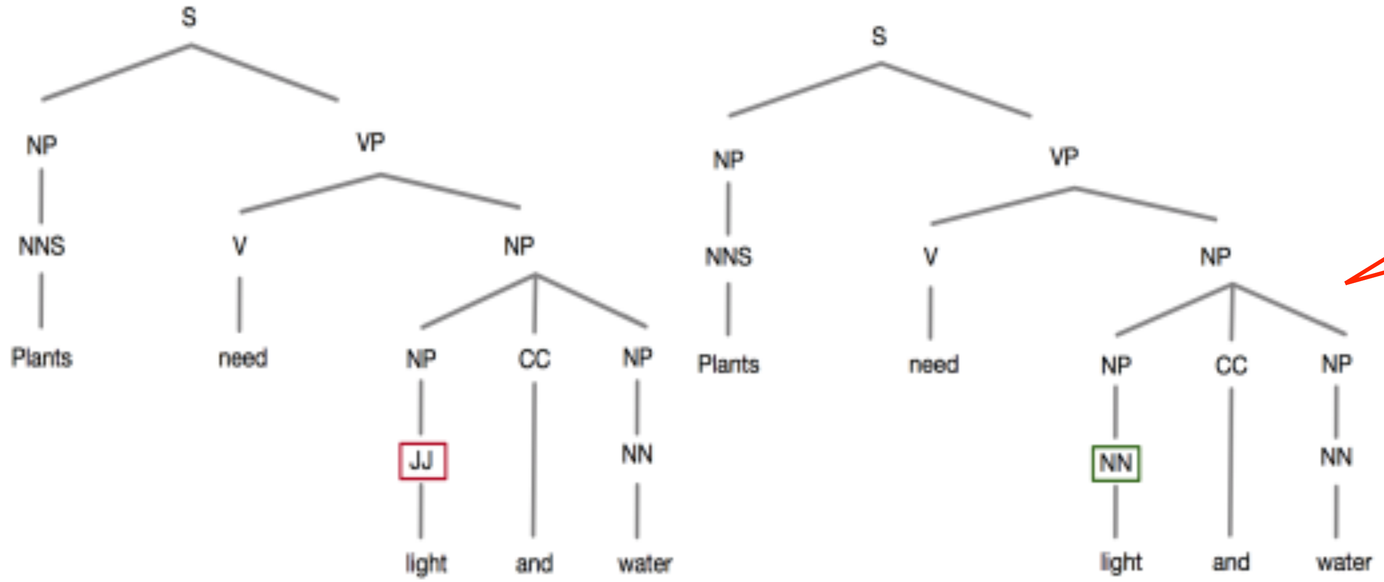


Figure is about POS tagging, so why does it also show a parse tree?

Figure 2 POS tagging for the example sentence "Plants need light and water". The result on the left-hand side comes from Festival; the figure on the right-hand side is the expected interpretation by humans.

Too much information?

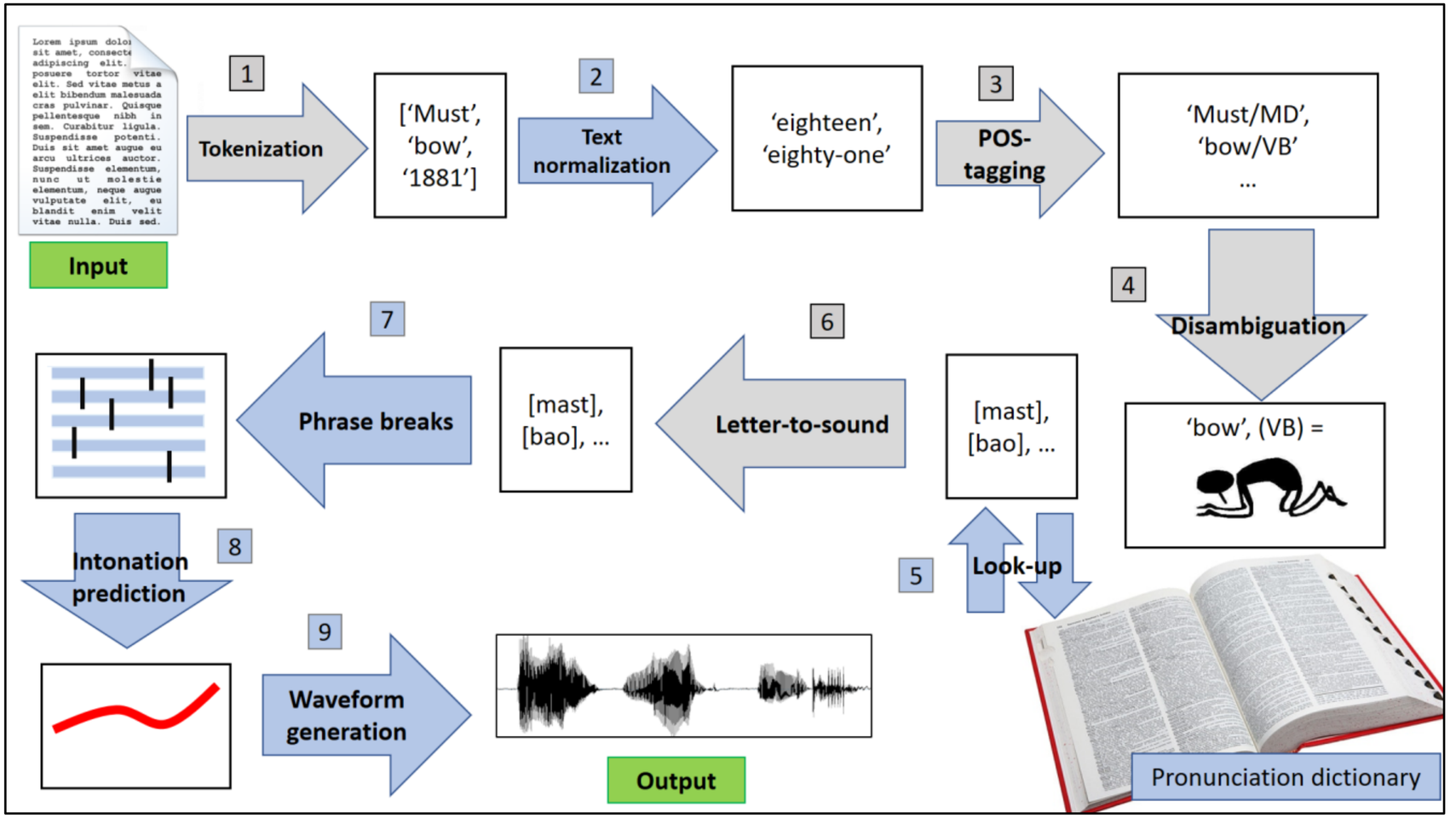
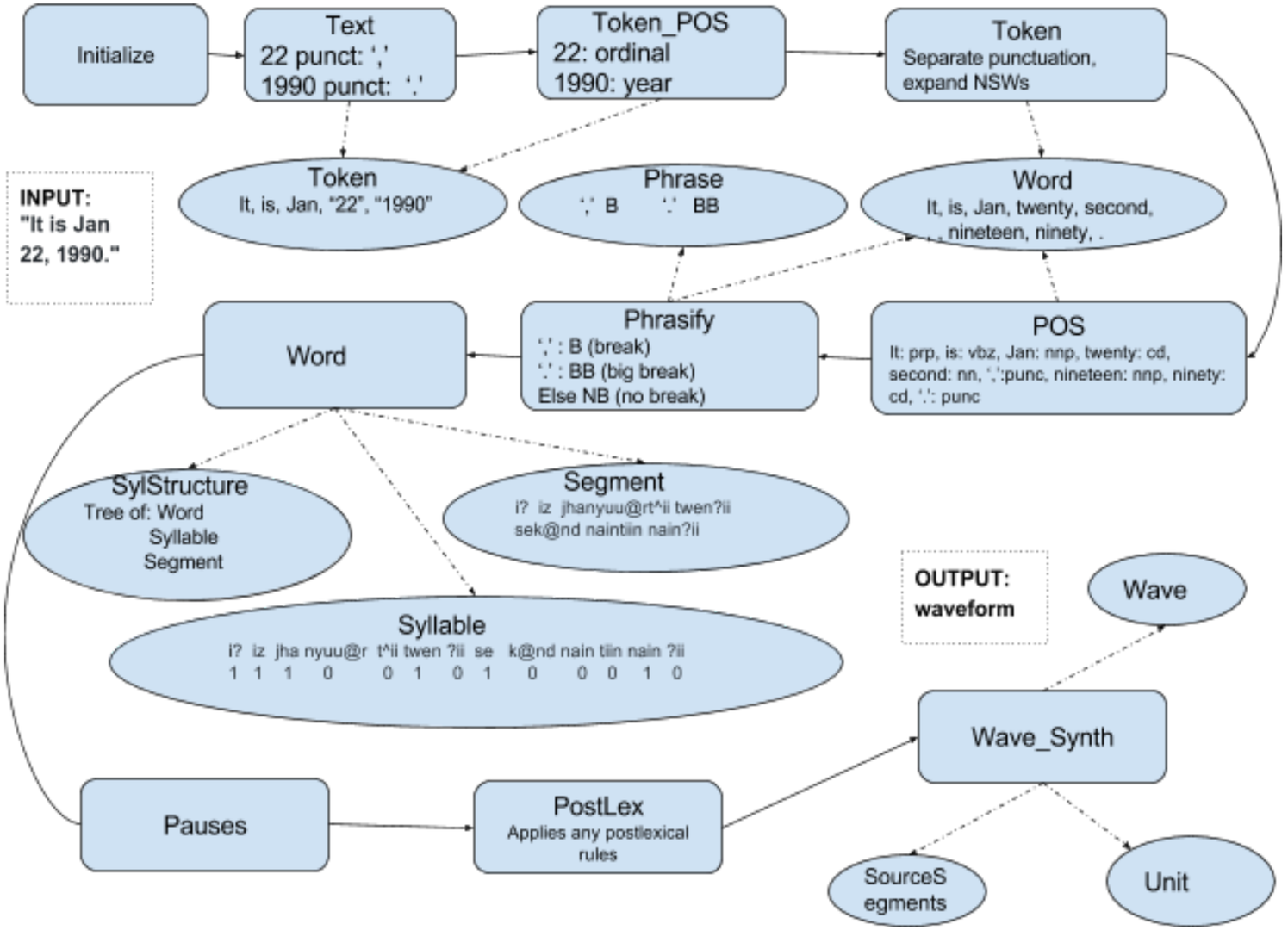


Fig. 1: The TTS pipeline

Too much information?



Annotation

what are the arrows pointing at?

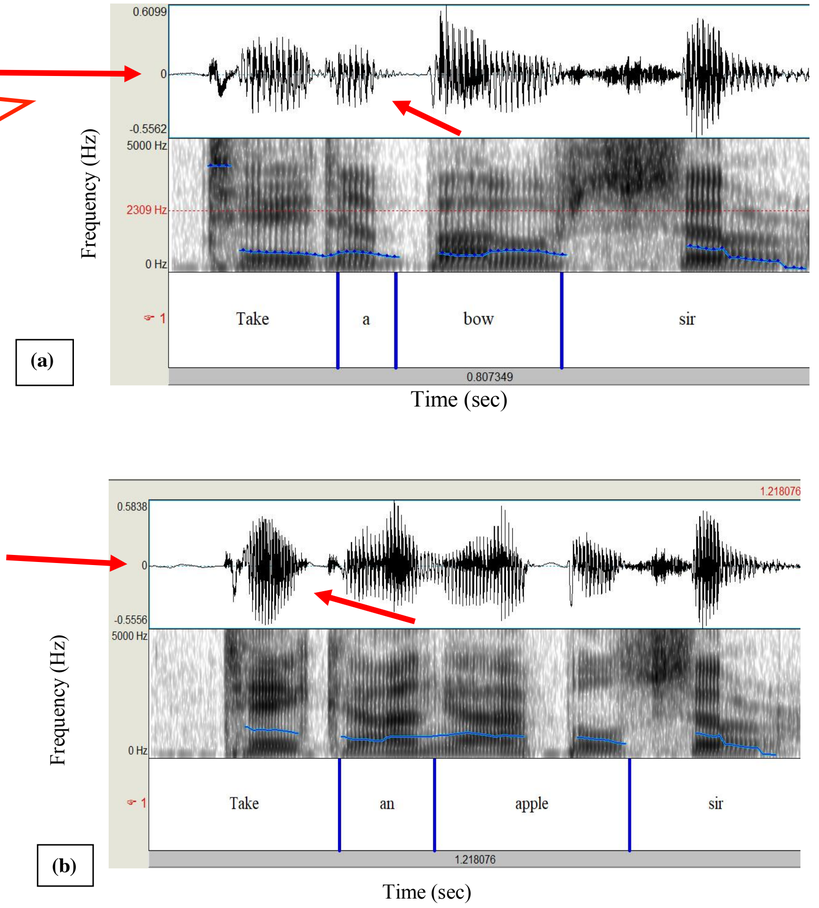


Figure 4. (a) *Waveform and spectrogram of "Take a bow sir" where "take" seems to be slightly anomalous in its form.*

(b) *Waveform and spectrogram of "Take an apple sir" where "take" sounds natural.*

Annotation