Orientation

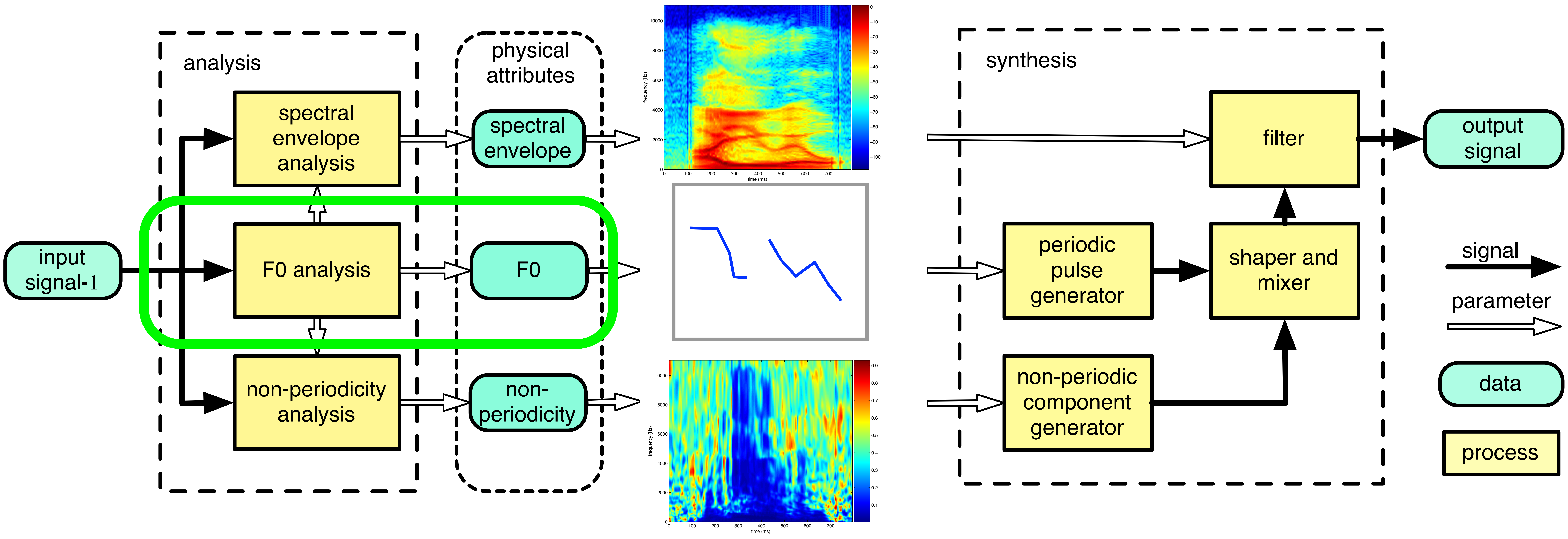
- Modules 1 to 5
  - Unit selection speech synthesis
  - The database
  - Evaluation
- Module 6
- Assignment

Orientation

- Module 6 (today's class)
  - Parameterising speech
  - Features that we want to model
  - A representation that can be modelled

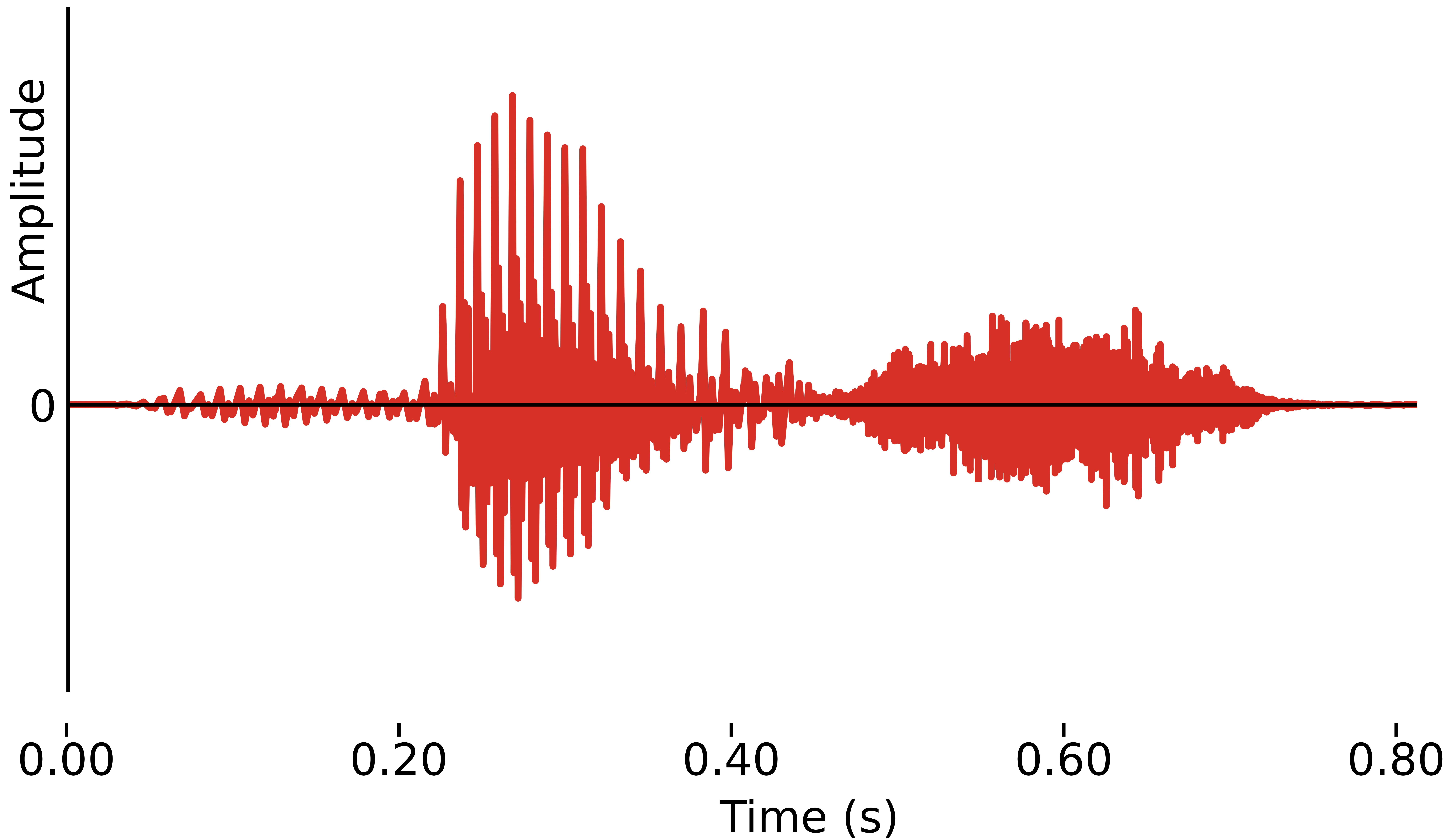
- A 'deep dive' into F0 estimation
  - F0 is a key feature we want to extract
  - RAPT is a classical example of a signal processing algorithm

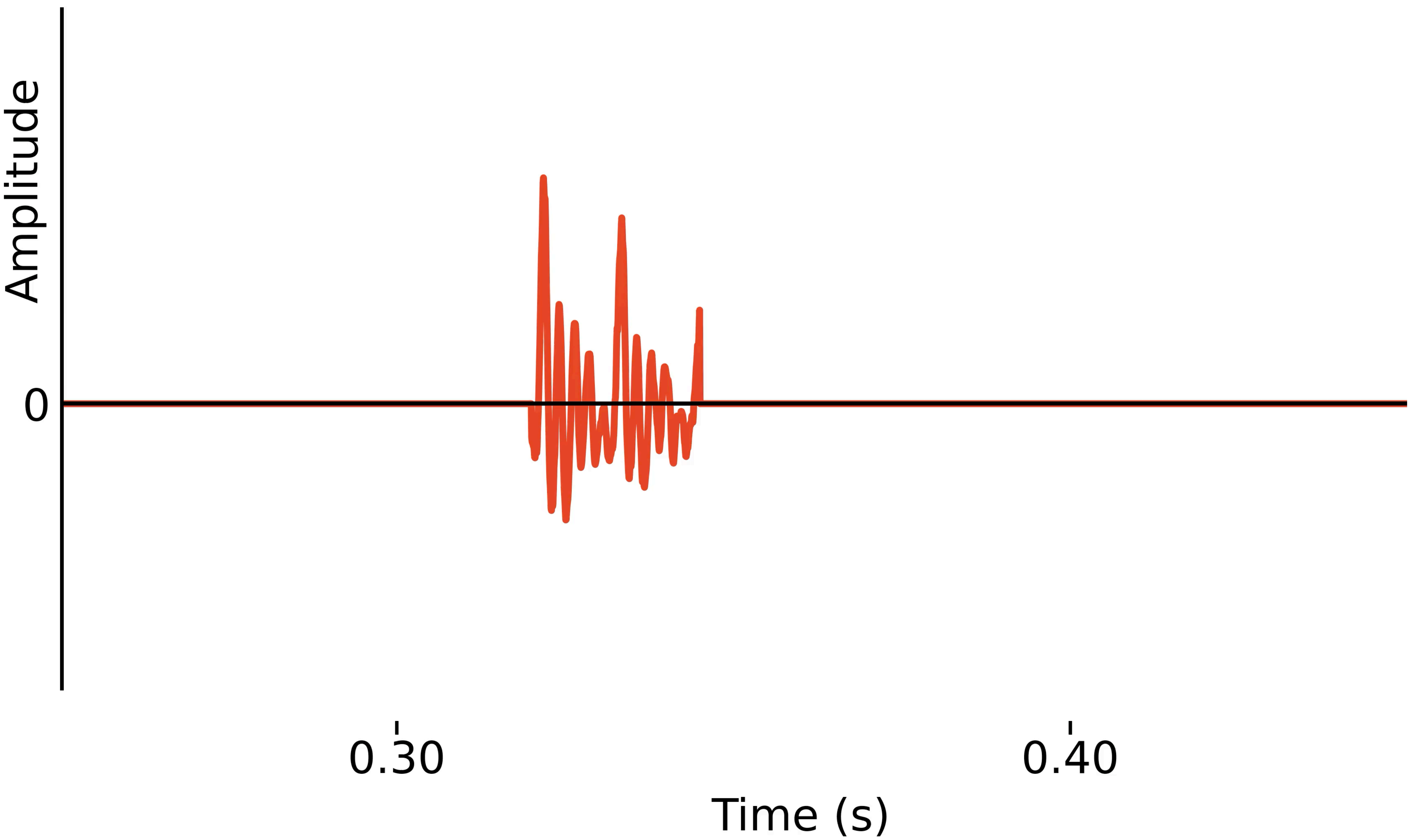
Orientation



Warm-up

- check your units !
  - time
  - frequency
  - sampling rate
  - sampling interval
  - samples
  - frame
- convert between time and samples
- describe a frame of samples from a longer waveform


Amplitude
0
0.00
0.20
0.40
0.60
0.80
Time (s)


Amplitude
0
0.30
0.40
Time (s)

F0 estimation (‘pitch tracking’)

- Discussion points

What's the relationship between samples and frames in Equation 2.1 ?

2.2.2. Autocorrelation

The autocorrelation function (ACF) of the speech signal, or of a pre-processed version of it, is a traditional source of period candidates [31]. Given s_p, $p = 0, 1, 2, \ldots$, a sampled speech signal with sampling interval $T = 1/F_s$, analysis frame interval t, and analysis window size w, at each frame we advance $z = t/T$ samples with $n = w/T$ samples in the autocorrelation window. w is chosen to be at least twice the longest expected glottal period; s is assumed to be zero outside the window. t is sized to sample adequately the time course of changes in F0. The ACF of K samples length, $K < n$, may then be defined as

$$R_{i,k} = \sum_{j=m}^{m+n-k-1} s_j s_{j+k}, \quad k = 0, K-1; \; m = iz; \; i = 0, M-1, \qquad (2.1)$$

where i is the frame index for M frames, and k is the *lag index* or *lag*. As outlined in

These equations are the *almost* same, except for notation

$$R_{i,k} = \sum_{j=m}^{m+n-k-1} s_j s_{j+k}, \quad k = 0, K-1; \; m = iz; \; i = 0, M-1, \qquad (2.1)$$

$$r_t(\tau) = \sum_{j=t+1}^{t+W} x_j x_{j+\tau},$$


0
198.60
198.70
198.80
198.90
200
100
0
198.6
198.65
198.7
198.75
198.8
198.85
198.9
198.95


0
198.60
198.70
198.80
198.90
200
100
0
198.6
198.65
198.7
198.75
198.8
198.85
198.9
198.95

Discuss the relative importance of each point, and how RAPT deals with it

- F0 changes with time, often with each glottal period.
- Sub-harmonics of F0 often appear that are sub-multiples of the "true" F0.
- In many cases when strong sub-harmonics are present, the most reasonable objective F0 estimate is clearly at odds with the auditory percept.
- Vocal-tract resonances and transmission-channel filtering can emphasize harmonics other than the first, causing F0 estimates that are multiples of the true F0.
- Occasionally F0 actually does jump up or down by an octave!
- Voicing is often very irregular at voice onset and offset leading to minimal waveshape similarity in adjacent periods.
- Panels of expert humans do not agree completely on the locations of voice onset and offset.
- Narrow-band filtering of unvoiced excitation by certain vocal-tract configurations can lead to signals with significant apparent periodicity.
- The amplitude of voiced speech has a wide dynamic range from low in voiced stop consonant closures to high in open vowels.
- It is difficult to distinguish periodic background noise from breathy voiced speech.

- Some voiced speech intervals are only a few glottal cycles in extent.

Draw a diagram that shows candidate generation

- Hint : start with Figure 2 (the correllogram)

Annotate **N_CANDS** on your diagram

Constant	Meaning	Value
$F0_{min}$	minimum F0 to search for (Hz)	50
$F0_{max}$	maximum F0 to search for (Hz)	500
t	analysis frame step size (sec)	.01
w	correlation window size (sec)	.0075
CAND_TR	minimum acceptable peak value in NCCF	.3
LAG_WT	linear lag taper factor for NCCF	.3
FREQ_WT	cost factor for F0 change	.02
VTRAN_C	fixed voicing-state transition cost	.005
VTR_A_C	delta amplitude modulated transition cost	.5
VTR_S_C	delta spectrum modulated transition cost	.5
VO_BIAS	bias to encourage voiced hypotheses	0.0
DOUBL_C	cost of exact F0 doubling or halving	.35
A_FACT	term to decrease ϕ of weak signals	10000
N_CANDS	max. number of hypotheses at each frame	20

Find a diagram in the slides on which you can annotate **CAND_TR**

Constant	Meaning	Value
$F0_{min}$	minimum F0 to search for (Hz)	50
$F0_{max}$	maximum F0 to search for (Hz)	500
t	analysis frame step size (sec)	.01
w	correlation window size (sec)	.0075
CAND_TR	minimum acceptable peak value in NCCF	.3
LAG_WT	linear lag taper factor for NCCF	.3
FREQ_WT	cost factor for F0 change	.02
VTRAN_C	fixed voicing-state transition cost	.005
VTR_A_C	delta amplitude modulated transition cost	.5
VTR_S_C	delta spectrum modulated transition cost	.5
VO_BIAS	bias to encourage voiced hypotheses	0.0
DOUBL_C	cost of exact F0 doubling or halving	.35
A_FACT	term to decrease ϕ of weak signals	10000
N_CANDS	max. number of hypotheses at each frame	20

Draw a diagram describing the dynamic programming

- What are the states?
  - and how many are there?
- What are the transitions?
- What is the local cost?
  - Hint: it's different for voiced vs unvoiced candidates
- What is the transition cost?
  - Hint: it depends on voicing status

Annotate your diagram describing the dynamic programming with

Constant	Meaning	Value
$F0_{min}$	minimum F0 to search for (Hz)	50
$F0_{max}$	maximum F0 to search for (Hz)	500
t	analysis frame step size (sec)	.01
w	correlation window size (sec)	.0075
CAND_TR	minimum acceptable peak value in NCCF	.3
LAG_WT	linear lag taper factor for NCCF	.3
FREQ_WT	cost factor for F0 change	.02
VTRAN_C	fixed voicing-state transition cost	.005
VTR_A_C	delta amplitude modulated transition cost	.5
VTR_S_C	delta spectrum modulated transition cost	.5
VO_BIAS	bias to encourage voiced hypotheses	0.0
DOUBL_C	cost of exact F0 doubling or halving	.35
A_FACT	term to decrease ϕ of weak signals	10000
N_CANDS	max. number of hypotheses at each frame	20

Orientation

- Assignment

Module 4 – the database

The quality of unit selection depends on good quality recorded speech, with accurate labels

Log in

Start | Videos | Readings | Quiz | Examples | Class | Finish

Download the slides for the module 4 videos

Total video to watch in this module: 58 minutes

Key concepts

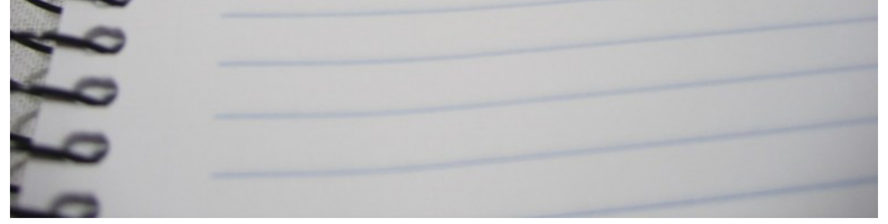
The base units (e.g., diphones) can occur in many different contexts. This makes it difficult to record a database that covers all possible units-in-context.

Script design

We can design the recording script in a way that should be better than randomly-selected text, in terms of coverage and other desirable properties.

Annotating the database

There are several reasons for avoiding manual annotation of the database. Instead, we will borrow methods from Automatic Speech Recognition.

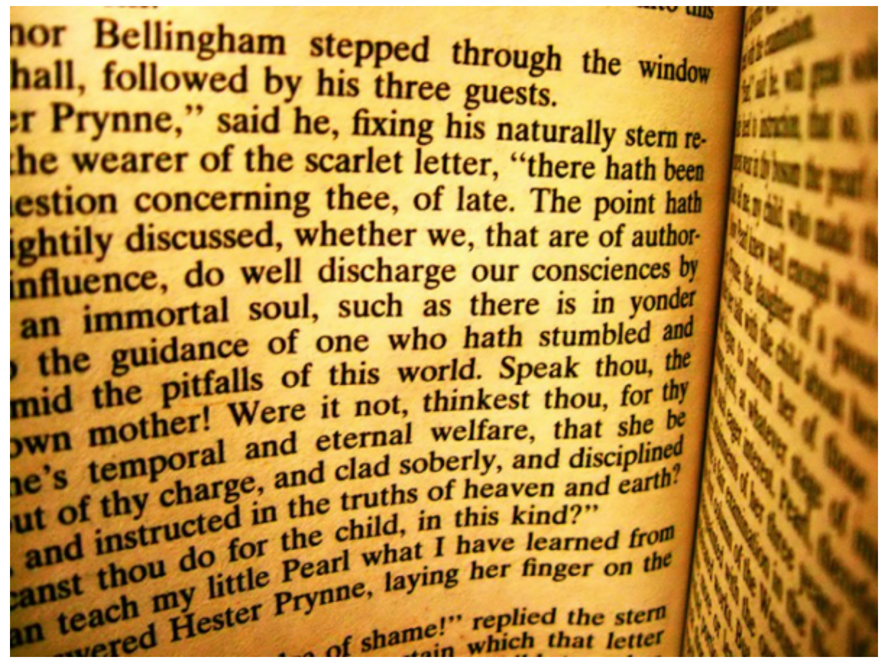


Prepare your workspace

We're going to be generating quite a lot of differ
workspace in which to keep them.

Milestones

To keep on track, check your progress against t
if you can.



The recording script

Because unit selection relies so heavily on the
carefully about exactly what speech we should

Make the recordings

With our carefully chosen script, we now need t
voice talent to record it. Consistency is the key
over multiple sessions.

Module 4 – the database

The quality of unit selection depends on good quality recorded speech, with accurate labels

Log in

Start | Videos | Readings | Quiz | Examples | Class | Finish

Download the slides for the module 4 videos

Total video to watch in this module: 58 minutes

Key concepts

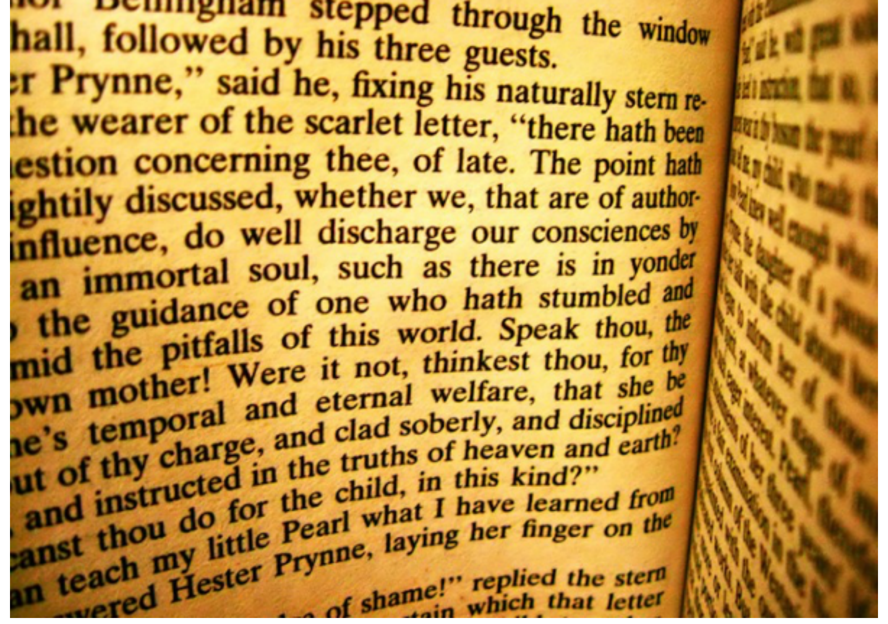
The base units (e.g., diphones) can occur in many different contexts. This makes it difficult to record a database that covers all possible units-in-context.

Script design

We can design the recording script in a way that should be better than randomly-selected text, in terms of coverage and other desirable properties.

Annotating the database

There are several reasons for avoiding manual annotation of the database. Instead, we will borrow methods from Automatic Speech Recognition.



The recording script

Because unit selection relies so heavily on the c
carefully about exactly what speech we should r

Make the recordings

With our carefully chosen script, we now need t
voice talent to record it. Consistency is the key
over multiple sessions.

Prepare the recordings

Move your recordings into the workspace, conv
do some sanity checking.

Label the speech

The labels are obtained from the text using the
we then need to align them to the recorded spe
automatic speech recognition.

Pitchmark the speech

Module 6 – speech signal analysis & modelling

Epoch detection, F0 estimation and the spectral envelope. Representing them for modelling. We also consider aperiodic energy. Then, we can analyse and reconstruct speech: this is called vocoding.

Log in

Start | Videos | Readings | Class | Quiz | Finish

Download the slides for the module 6 videos

Total video to watch in this module: 81 minutes

Key concepts

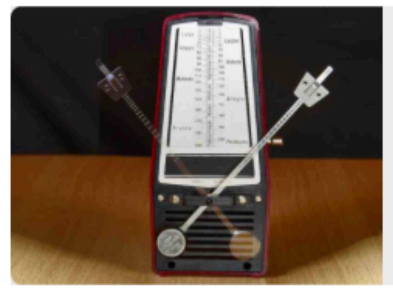
Extracting certain features from the speech signal is an essential first step before further processing, such as concatenating candidate unit waveforms, modifying the prosody of a speech signal, or statistical modelling.

Epoch detection

Epochs are moments in time, often defined as the Glottal Closure Instants, in voiced speech. Locating them consistently is necessary for some types of signal processing, such as Pitch Synchronous Overlap Add (PSOLA) methods.

F0 estimation (part 1)

Whilst epochs are moments in time, the fundamental frequency (F0) is the rate of vibration of the vocal folds expressed in Hertz. It is generally estimated over a frame spanning several pitch periods, containing multiple epochs.



F0 estimation (part 2)

Some F0 estimation algorithms apply pre-processing to the speech waveform, and all use post-processing to select from multiple candidate values for F0. Most algorithms have several parameters that need to be carefully chosen.

Spectral envelope estimation

Until now, we have conflated the vocal tract frequency response with the spectral envelope. We now take a strictly signal-based view of speech, and define the spectral envelope more carefully.

Speech signal modelling

After we parameterise a speech signal, we need to decide how best to represent those parameters for use in statistical modelling, and eventually how to reconstruct the waveform from them.

Label the speech

The labels are obtained from the text using the
we then need to align them to the recorded spe
automatic speech recognition.

Pitchmark the speech

The signal processing used for waveform conca
requires the speech database to have the indivi

Build the voice

The final stages of building the voice involve cre
and join costs, plus the representation of the sp

Run the voice

We're done! Time to find out what it sounds like

Module 6 – speech signal analysis & modelling

Epoch detection, F0 estimation and the spectral envelope. Representing them for modelling. We also consider aperiodic energy. Then, we can analyse and reconstruct speech: this is called vocoding.

Log in

Start | Videos | Readings | Class | Quiz | Finish

Download the slides for the module 6 videos

Total video to watch in this module: 81 minutes

Key concepts

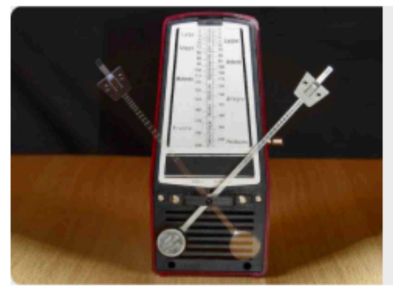
Extracting certain features from the speech signal is an essential first step before further processing, such as concatenating candidate unit waveforms, modifying the prosody of a speech signal, or statistical modelling.

Epoch detection

Epochs are moments in time, often defined as the Glottal Closure Instants, in voiced speech. Locating them consistently is necessary for some types of signal processing, such as Pitch Synchronous Overlap Add (PSOLA) methods.

F0 estimation (part 1)

Whilst epochs are moments in time, the fundamental frequency (F0) is the rate of vibration of the vocal folds expressed in Hertz. It is generally estimated over a frame spanning several pitch periods, containing multiple epochs.



F0 estimation (part 2)

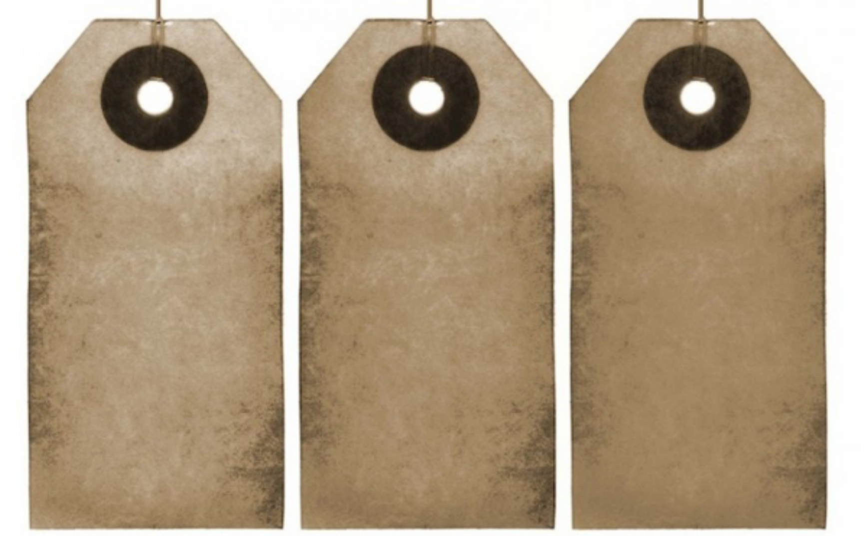
Some F0 estimation algorithms apply pre-processing to the speech waveform, and all use post-processing to select from multiple candidate values for F0. Most algorithms have several parameters that need to be carefully chosen.

Spectral envelope estimation

Until now, we have conflated the vocal tract frequency response with the spectral envelope. We now take a strictly signal-based view of speech, and define the spectral envelope more carefully.

Speech signal modelling

After we parameterise a speech signal, we need to decide how best to represent those parameters for use in statistical modelling, and eventually how to reconstruct the waveform from them.



The labels are obtained from the text using the
we then need to align them to the recorded spe
automatic speech recognition.

Pitchmark the speech

The signal processing used for waveform conca
requires the speech database to have the indivi

Build the voice

The final stages of building the voice involve cre
and join costs, plus the representation of the sp

Run the voice

We're done! Time to find out what it sounds like

Module 3 – unit selection target cost functions

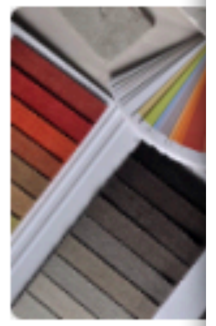
The target cost is critical to choosing an appropriate unit sequence. Several different forms are possible, using linguistic features, or acoustic properties, or a combination of both.

Log in

Start | Videos | Readings | Quiz | Class | Finish

Down

Total vide



Module 4 – the database

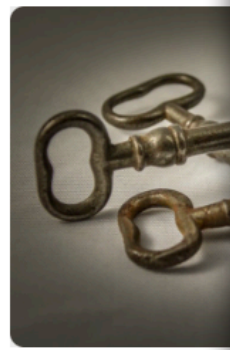
The quality of unit selection depends on good quality recorded speech, with accurate labels

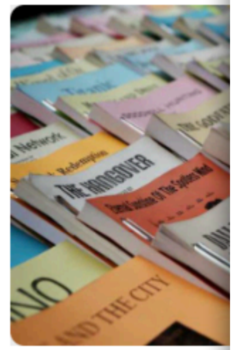
Log in

Start | Videos | Readings | Quiz | Examples | Class | Finish

Dow

Total video





Module 6 – speech signal analysis & modelling

Epoch detection, F0 estimation and the spectral envelope. Representing them for modelling. We also consider aperiodic energy. Then, we can analyse and reconstruct speech: this is called vocoding.

Log in

Start | Videos | Readings | Class | Quiz | Finish

Download the slides for the module 6 videos

Total video to watch in this module: 81 minutes

Key concepts

Extracting certain features from the speech signal is an essential first step before further processing, such as concatenating candidate unit waveforms, modifying the prosody of a speech signal, or statistical modelling.

Epoch detection

Epochs are moments in time, often defined as the Glottal Closure Instants, in voiced speech. Locating them consistently is necessary for some types of signal processing, such as Pitch Synchronous Overlap Add (PSOLA) methods.

Run the voice

We're done! Time to find out what it sounds like

Improvements and variations

It would take too long to tune every aspect of th
problems and see how to fix them. It's also easy
discover the effect on the synthetic speech.

Evaluation

The main form of evaluation should be a listenir
there are other ways to evaluate, and potentially

Writing up

Because you kept such great notes in your logb
and painless.

Module 5 – evaluation

How do we know how good our synthesiser is? Can we use formal evaluation to decide how to improve it?

Log in

Start | Videos | Readings | Quiz | Class | Finish

Download the slides for the module 5 videos

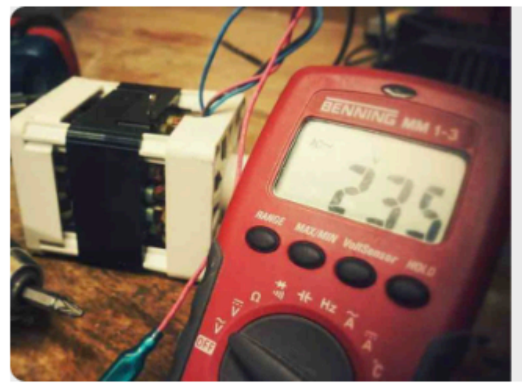
Total video to watch in this module: 65 minutes

Why? When? Which aspects?

What are our goals when evaluating synthetic speech?

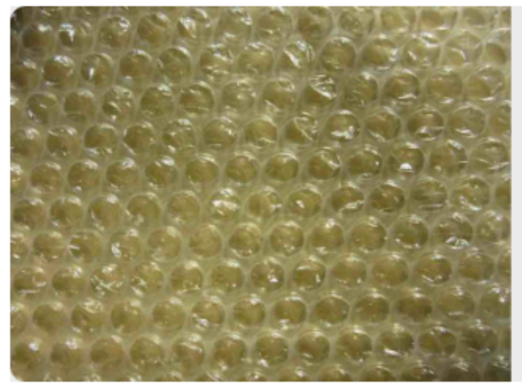
Subjective evaluation

The most commonly used, and the most reliable, method of evaluation is to ask people to listen to synthetic speech and provide a response. Often that is simply a preference, or an opinion score.



Objective evaluation

It would be convenient to avoid using listeners and instead use an objective, or algorithmic, measure to evaluate synthetic speech. This is possible, but only to a rather limited extent. Use with caution!



Wrap up

We'll conclude with a reminder of what to do with the outcome of an evaluation, and some recommendations for what types of test are suitable for various purposes.

We're done! Time to find out what it sounds like

Improvements and variations

It would take too long to tune every aspect of th problems and see how to fix them. It's also easy discover the effect on the synthetic speech.

Evaluation

The main form of evaluation should be a listenir there are other ways to evaluate, and potentially

Writing up

Because you kept such great notes in your logb and painless.

Speech Synthesis assignment marking scheme 2023-24

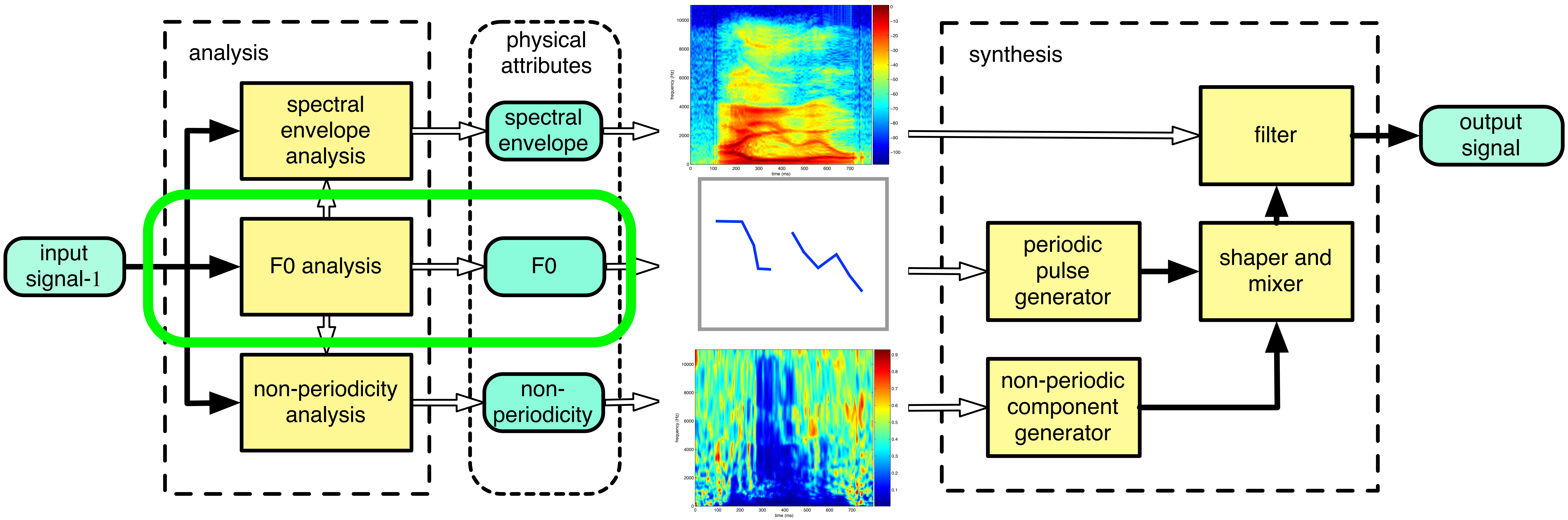
Category		Points available
Understanding (theory) **20 points**	Title, abstract	5
	Explaining unit selection	5
	Theoretical connections to current methods	10
Critical thinking (putting theory into practice) **20 points**	Data: script, dictionary, recording, alignment	5
	Signal processing: pitchmarking, F0, etc	5
	Practical implications for current methods	10
Evaluation **20 points**	Experimental design	10
	Execution of a basic listening test	5
	Conclusions	5
Scientific writing **20 points**	Conform with the journal style guide *and* anonymous submission, correct filename, exam number, state wordcount, page numbers	5
	Clarity, coherence, structure, presentation, figures & captions, bibliography	15
Additional (for a higher mark) **20 points**	*Any/all of these and/or going beyond the basic expectations in other ways:* • better script design (manual or automatic) • recording additional data • a more sophisticated listening test • forms of evaluation other than a listening test • using your knowledge of phonetics • …and so on	20
TOTAL		**100**

What next?

- We have decomposed speech into
  - F0, plus a V/UV decision
  - smooth spectral envelope, parameterised as the Mel-cepstrum
  - band aperiodicity parameters
- We've seen how to reconstruct the waveform

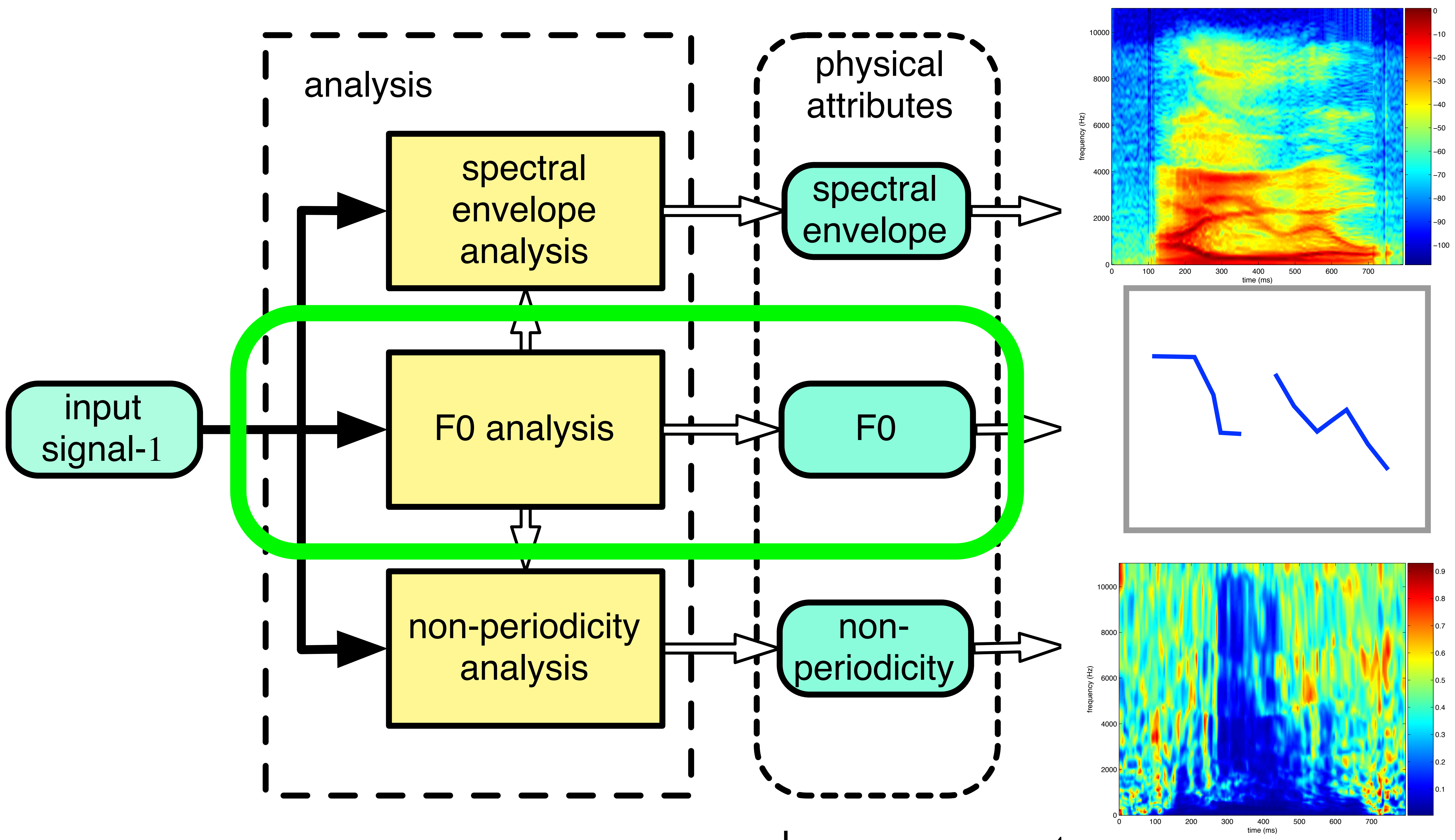
- Now we can insert a **statistical model** between the analysis and synthesis parts

What next?

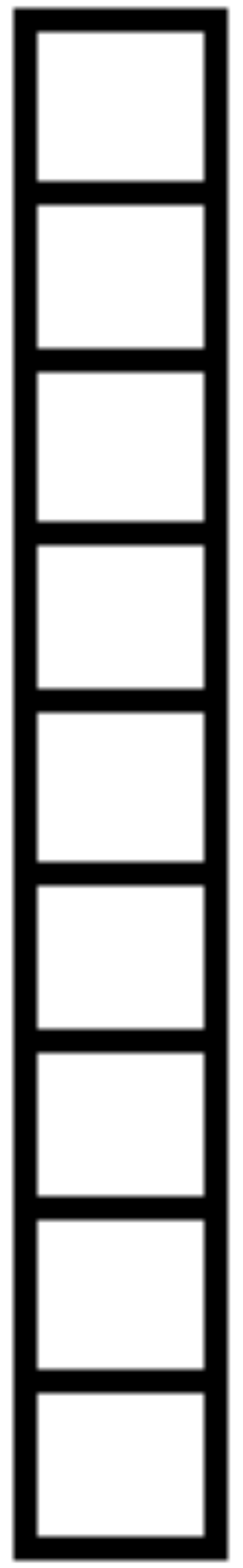


Figures: Hideki Kawahara

Speech parameters



speech parameters



feature vector

Speech parameters in more recent approaches