Parametric vocoding (for speech synthesis) (and coding)

Phil Garner

Idiap Research Institute

Contents

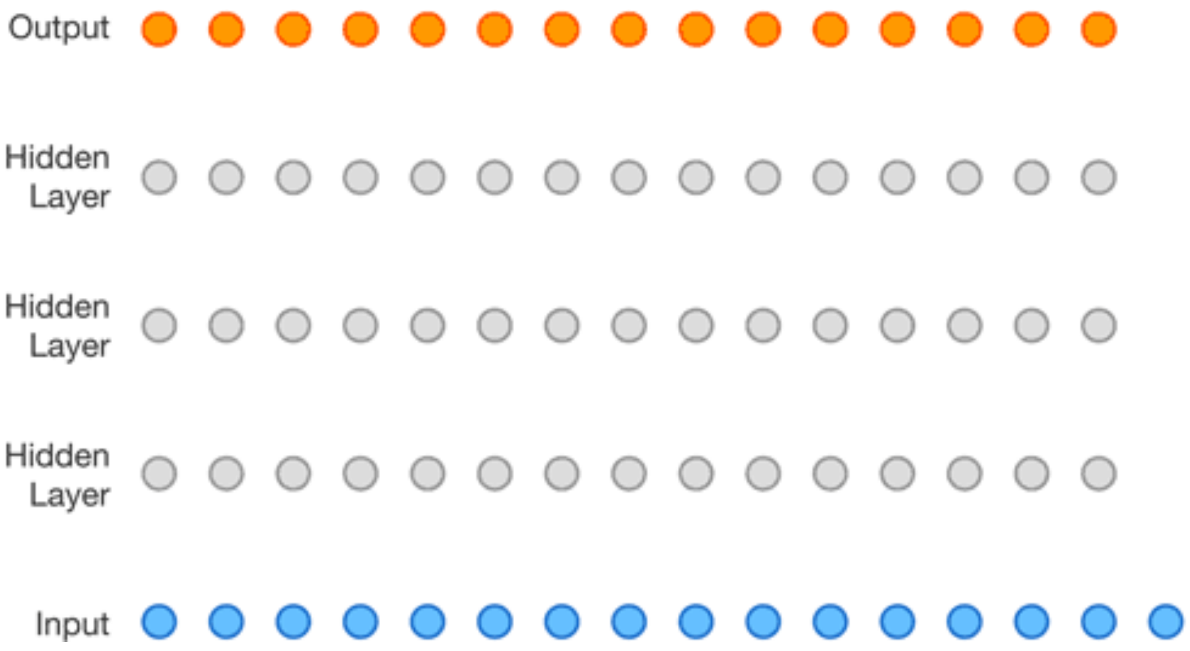
- Part 1: Background applications
  - Some justification for all this
- Part 2: Approaches to vocoding
  - A few insights into what's involved
  - Probably incomplete, probably a couple of years out of date
- Part 3: Shortcomings
  - Some things I tried to address

Isn’t this a hearing conference?

During the evolution of human speech, the articulatory motor system has presumably structured its output to match those rhythms the auditory system can best apprehend. Similarly, the auditory system has likely become tuned to the complex acoustic signal produced by combined jaw and articulator rhythmic movements. Both auditory and motor systems must, furthermore, build on the existing biophysical constraints provided by the neuronal infrastructure.

– Giraud and Poeppel (2012)
attributed to Heimbauer et al. (2011) and Liberman and Mattingley (1985)

Is it all moot?

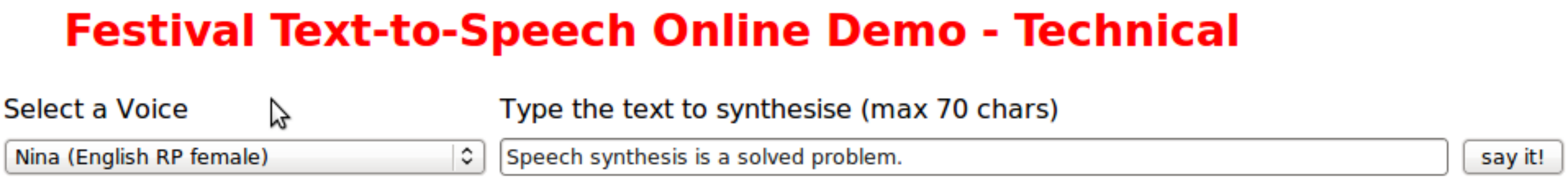


- There is now wavenet (left)
  - Synthesises waveform directly
- Also
  - Lyrebird https://lyrebird.ai
  - Char2Wav http://josesotelo.com/speechsynthesis/
  - Tacotron (not strictly waveform synthesis)
- Do we really need to understand the production mechanism?

Part 1

Background applications

Text to speech synthesis



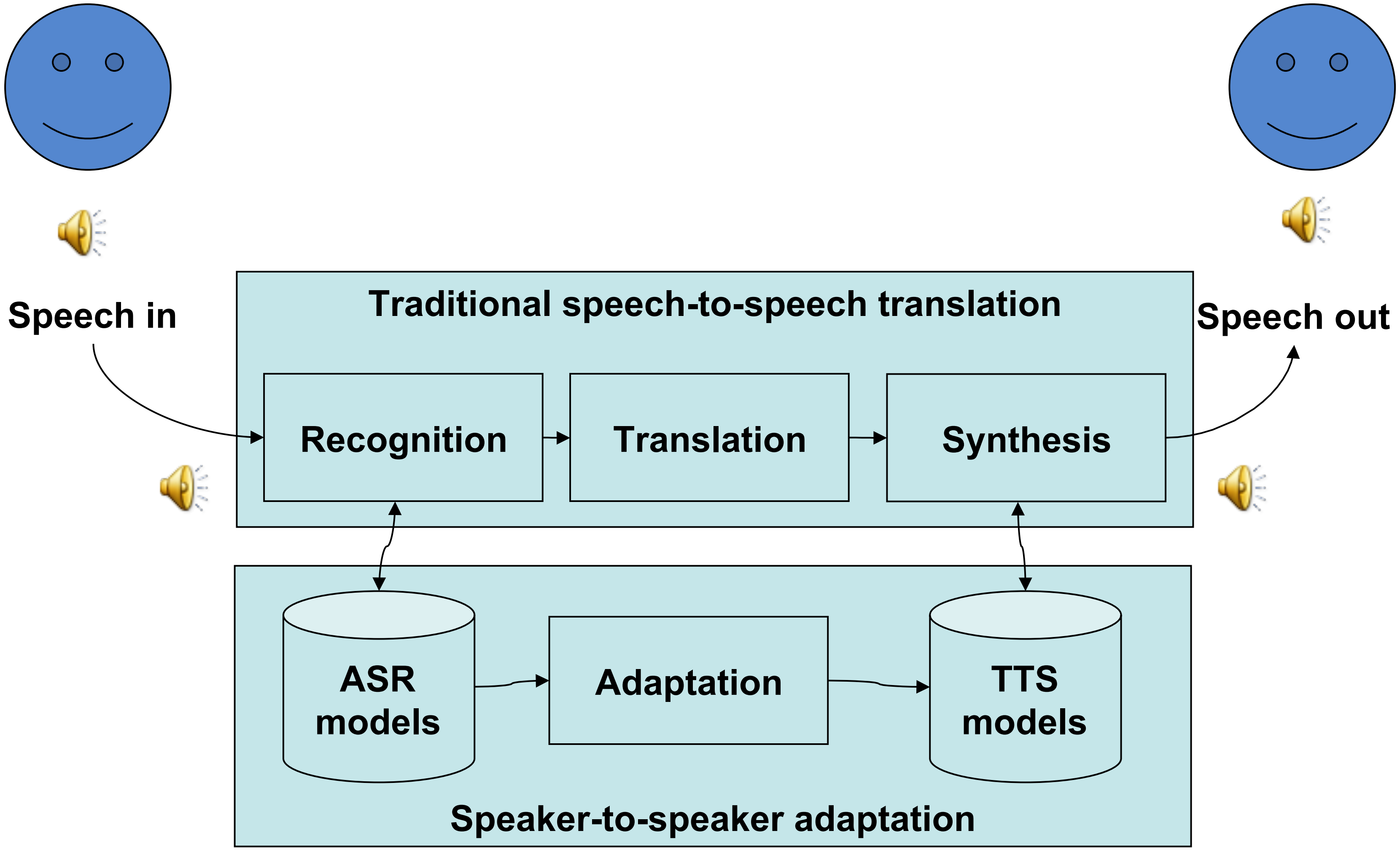
Festival Text-to-Speech Online Demo - Technical

Select a Voice: Nina (English RP female)

Type the text to synthesise (max 70 chars): Speech synthesis is a solved problem.

- It works!
  - Basic TTS *in its raw form* is a solved problem
  - http://www.cstr.ed.ac.uk/projects/festival/

The EMIME scenario



Speech in

Traditional speech-to-speech translation

Speech out

Recognition

Translation

Synthesis

ASR models

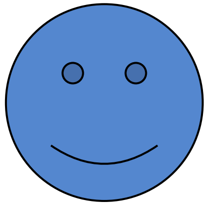
Adaptation

TTS models

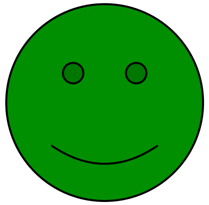
Speaker-to-speaker adaptation

Example of adaptation

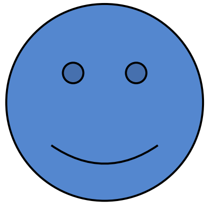
Natural speech



Average voice

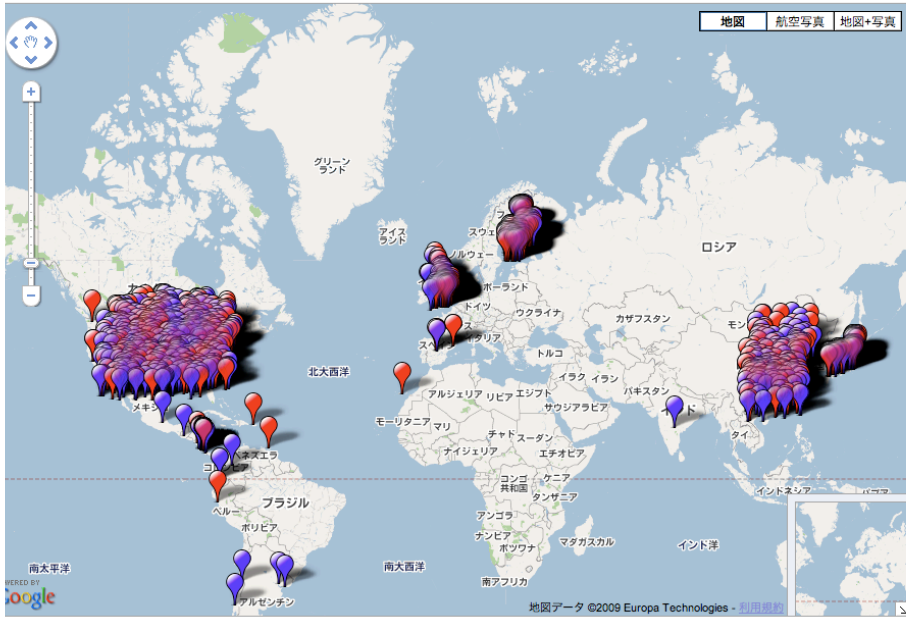


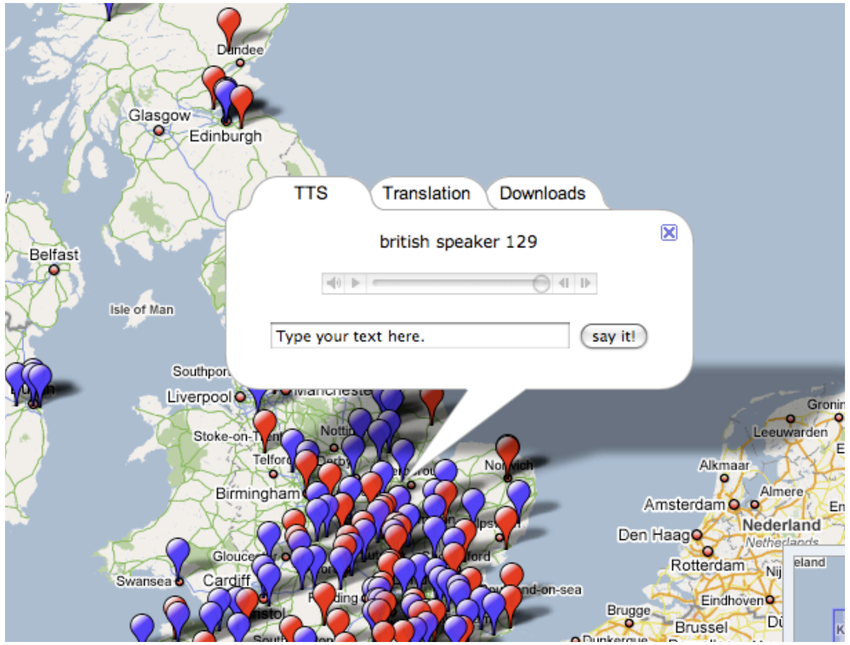
Adapted voice



- Note:
  - 16 kHz speech (not such high quality for synthesis)
  - Wall St. Journal samples (an ASR database, more males)
  - VTLN adaptation (quite simple, gender characteristics)

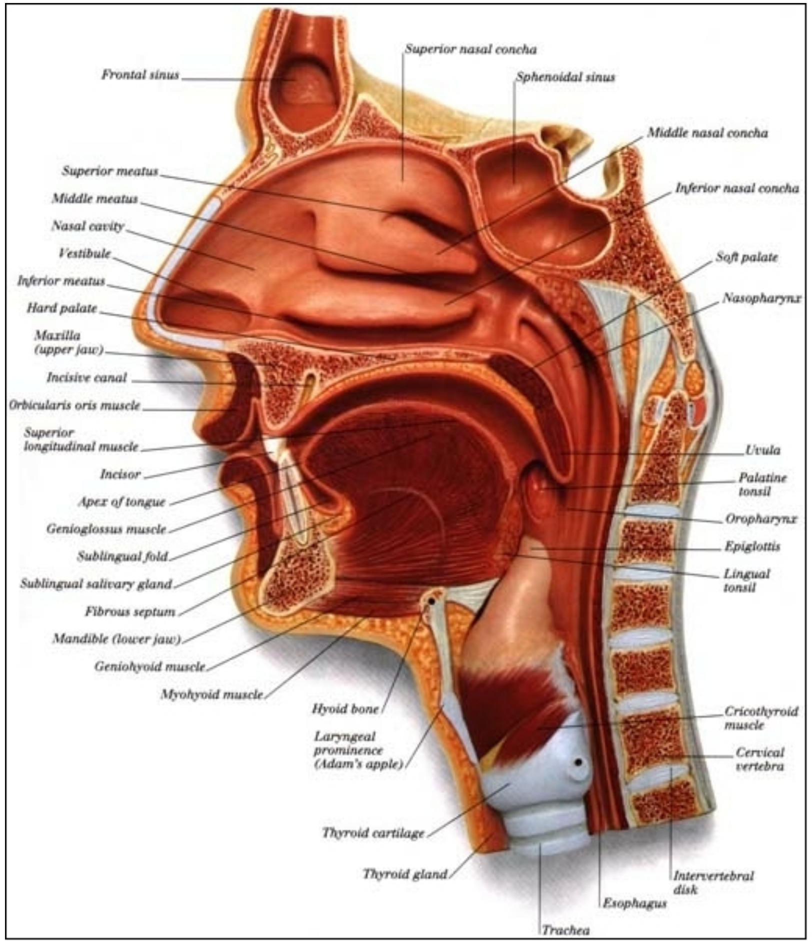
What can be done with adaptation





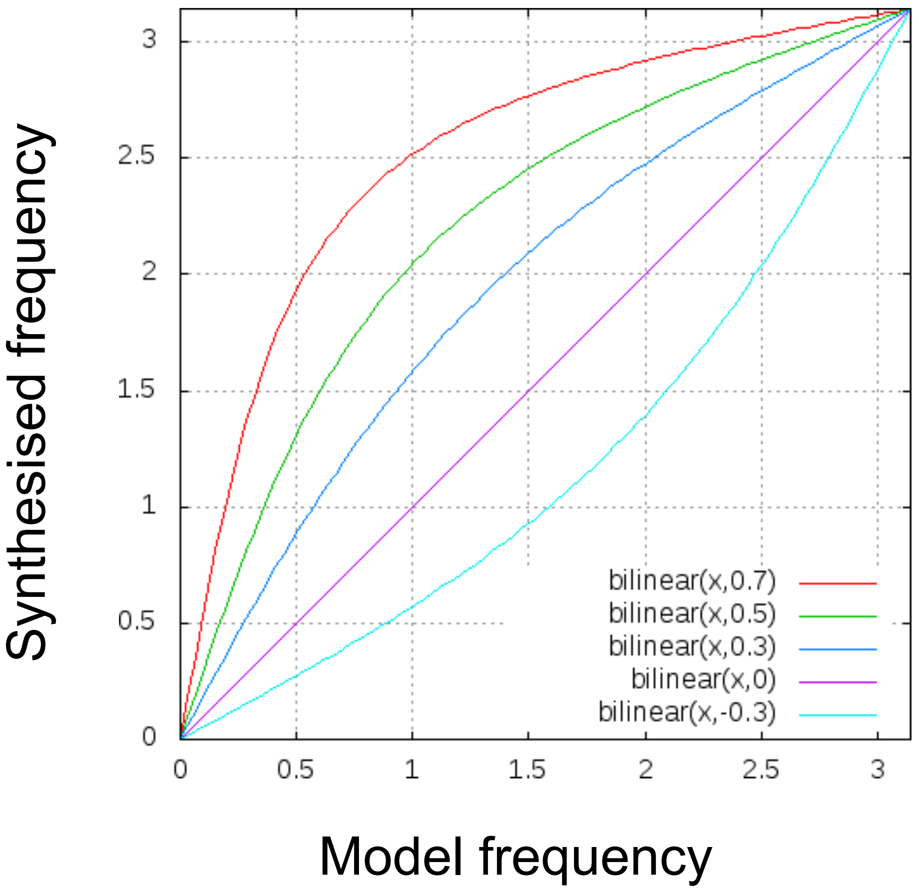
http://homepages.inf.ed.ac.uk/jyamagis/Demo-html/map-new.html

Vocal tract length normalisation



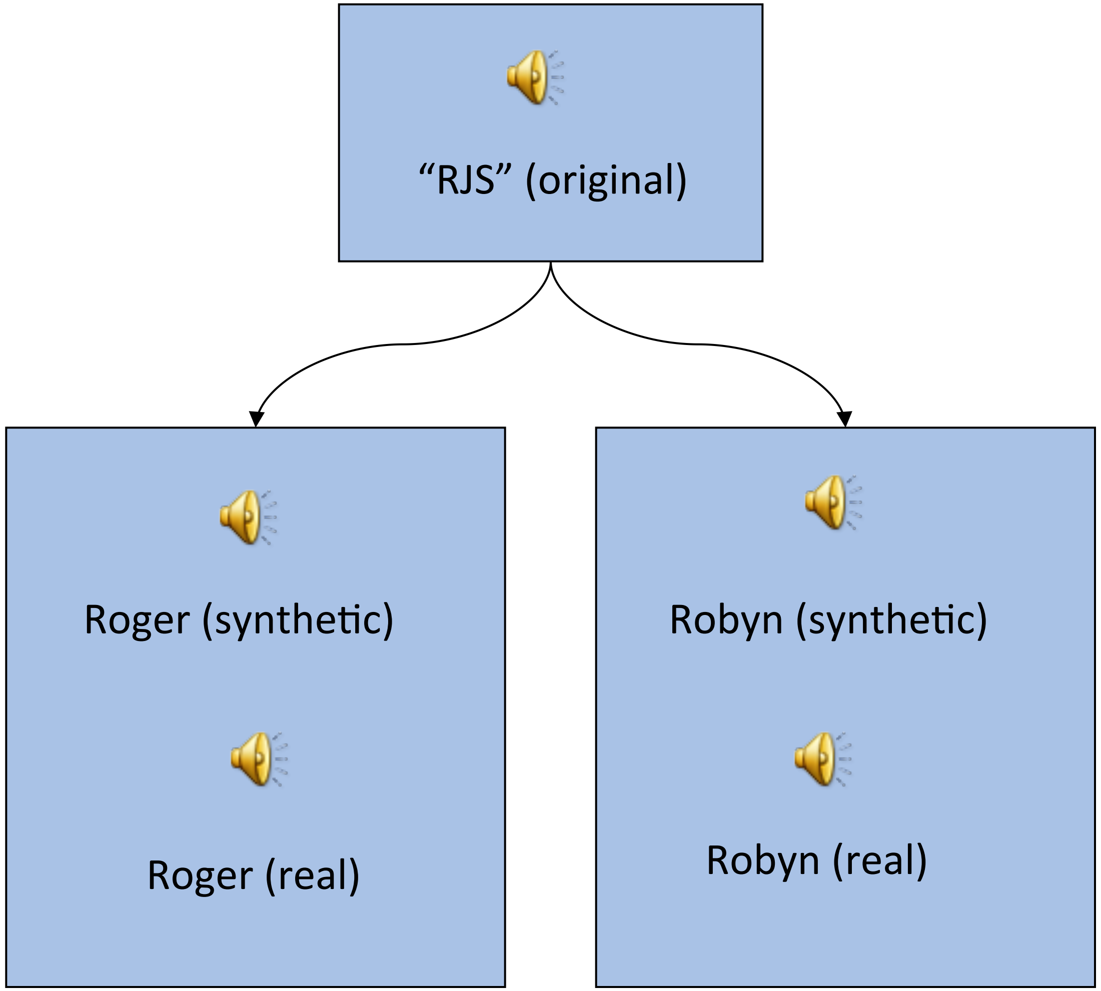
- Distance between larynx and lips
- Longer in males
  - "Lower voice"
- Shorter in females
  - "Higher voice"
- Adaptation does more than alter vocal tract length
  - But VTL is easy to visualise

Bilinear warping



- Bilinear warping
  - can stretch or shrink a vocal tract
  - Enough to distinguish male / female
  - It can also be represented as a linear transform
- Very quick to learn
  - Only one speech example!
  - Only one parameter to learn
  - Can be prior for more sophisticated techniques

Fast, difficult adaptation



“RJS” (original)

Roger (synthetic)

Roger (real)

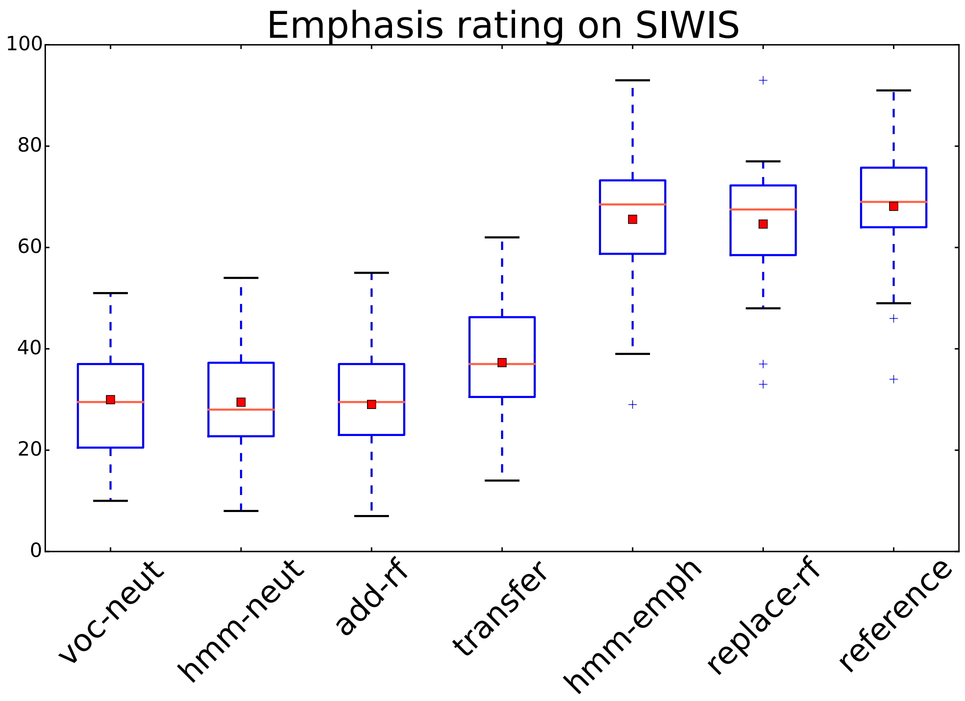
Robyn (synthetic)

Robyn (real)

- Note:

- 48 kHz samples (quite good!)
- RJS is a HTS2010 / Festival / Blizzard challenge voice
- VTLN + SMAPLR adaptation
- Roger: 1 sentence
- Robyn: 4 sentence

Speech synthesis

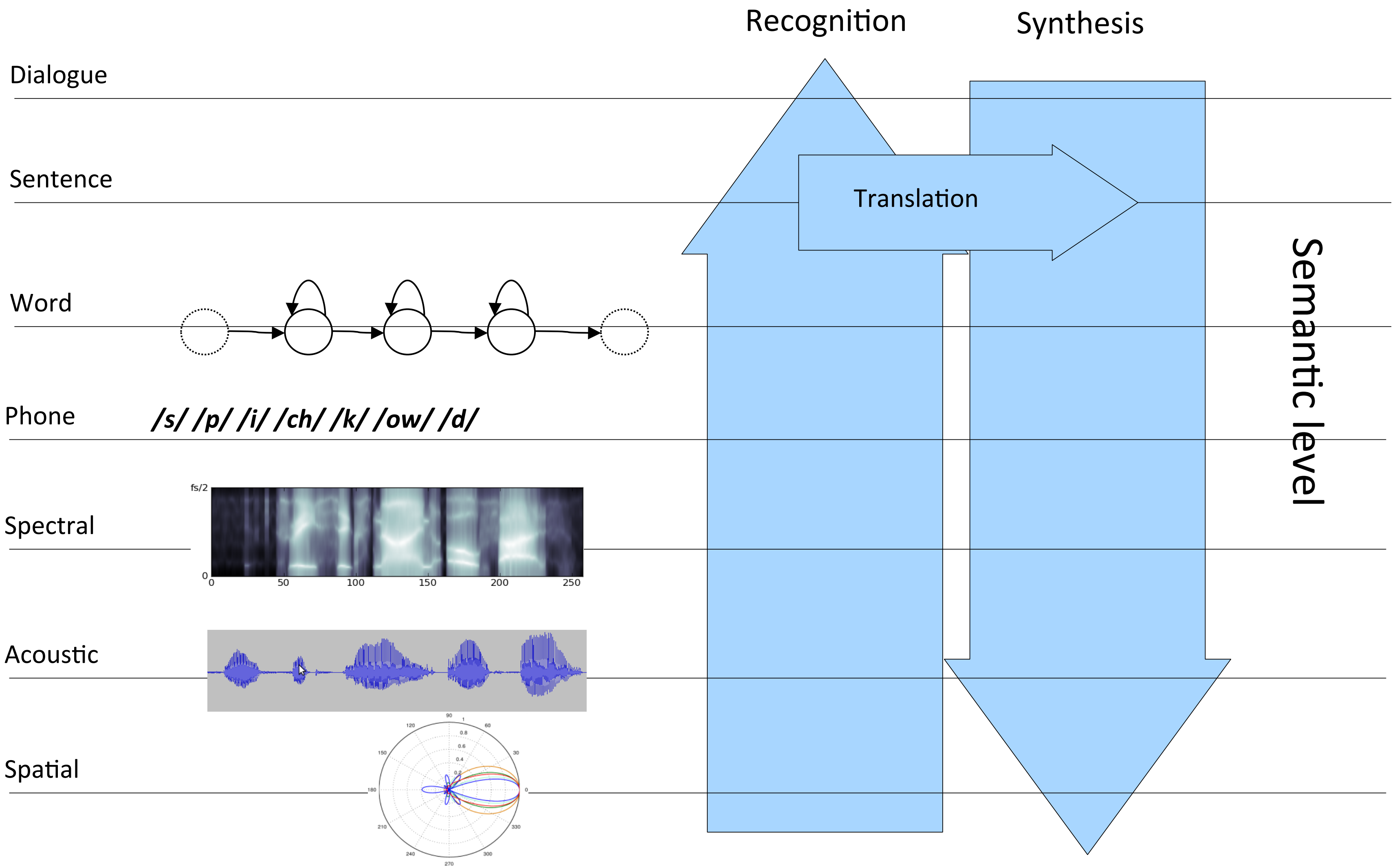


Random box plot

- TTS involves lots of listening tests e.g., Alex’s page
  - Which sounds most like the original?
  - Which is most natural?
  - None of them!
    They all sound unnatural!
- There is a characteristic “buzzy” quality
  - The effect of the vocoder is stronger than that of the thing we’re trying to modify

Part 2

Approaches to vocoding



Recognition

Synthesis

Translation

Semantic level

Dialogue

Sentence

Word

Phone ***/s/ /p/ /i/ /ch/ /k/ /ow/ /d/***

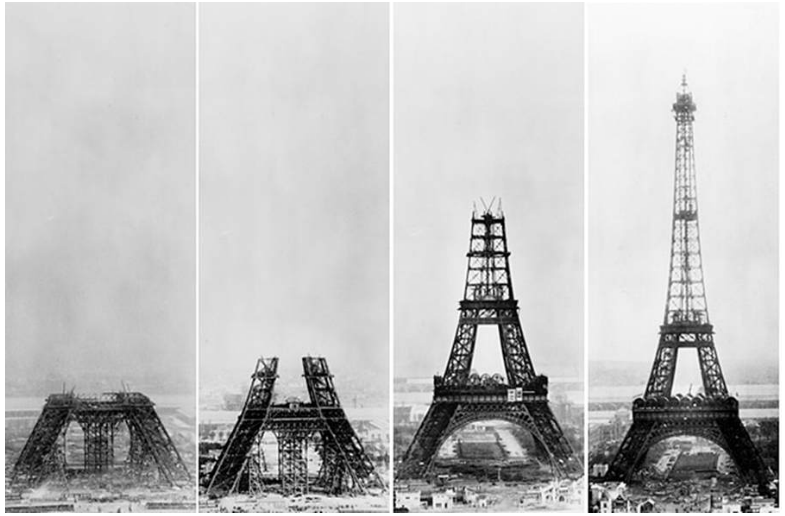
Spectral

Acoustic

Spatial

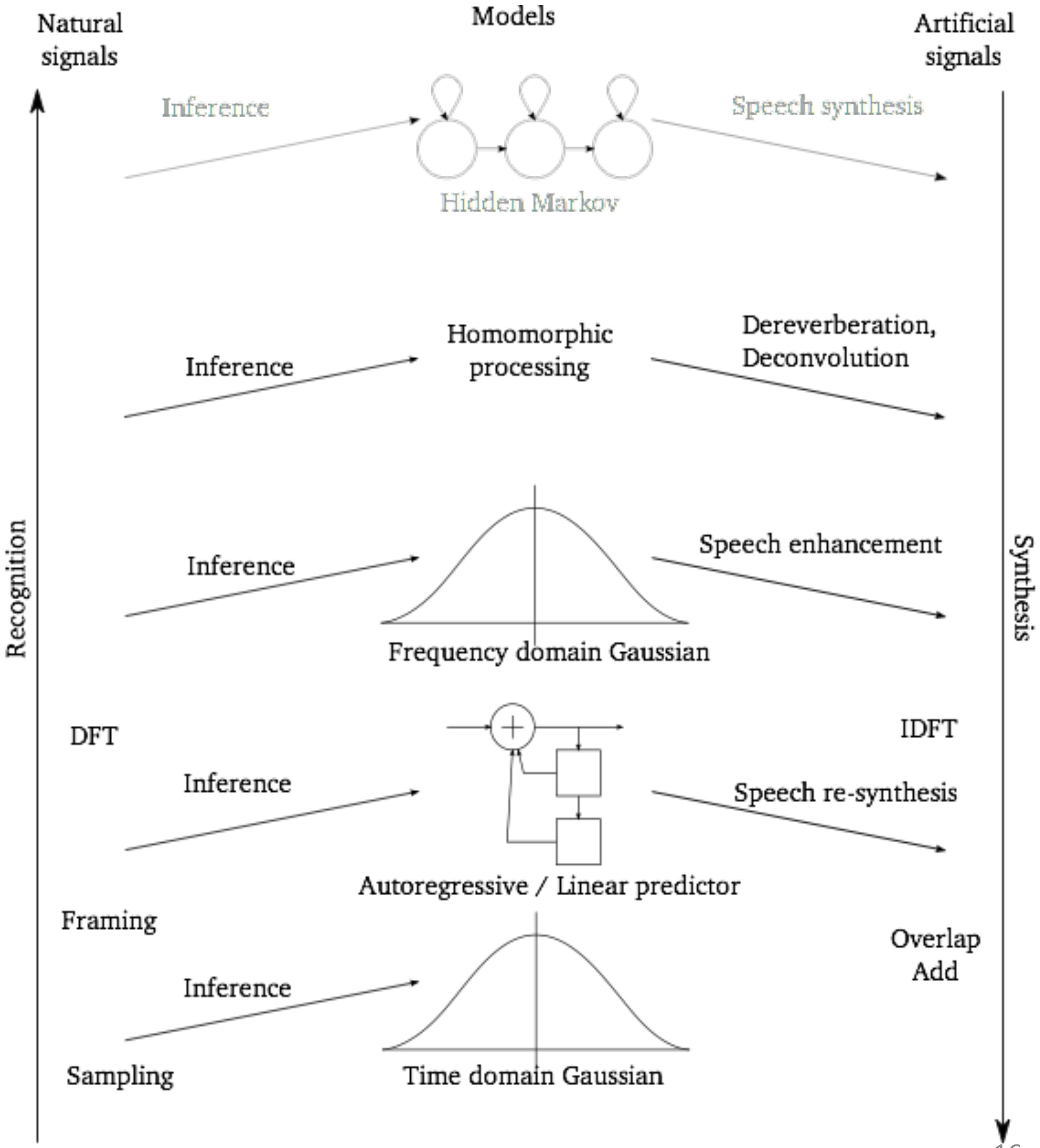
"One cannot understand speech recognition without understanding speech synthesis.
One cannot understand either without understanding speech coding."
- (in the spirit of) James Baker, ICASSP 2012, Kyoto.

Similar views…



…pinched from Andrea

There is no hope at the high levels if the lower ones don't work!

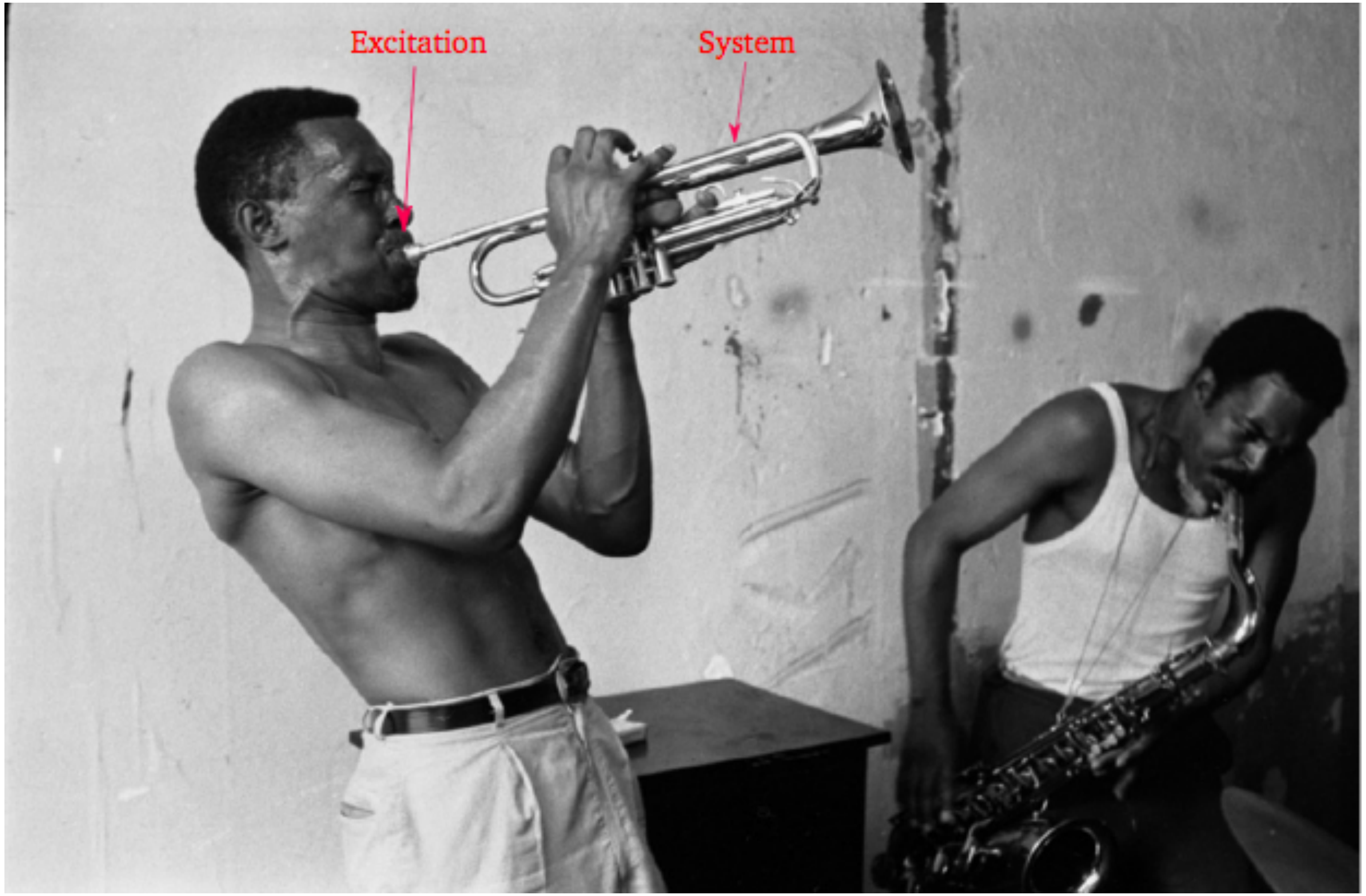


Non-parametric vs. parametric

- MPEG-like codecs do not care about content
  - What comes out is what went in
  - May use LP, may not
  - mp3 uses LP
- Generally structure doesn't matter

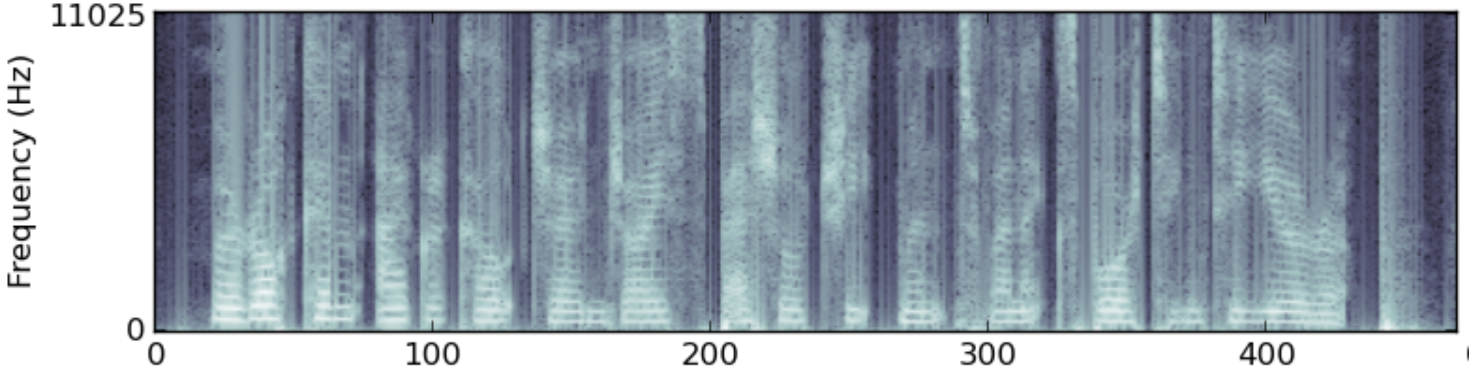
- In TTS, we want to be able to control
  - Voice character
  - Pitch
  - Emotion
- Structure matters!
- Examples:
  - STRAIGHT, YANG, Vocaine
  - WORLD: http://ml.cs.yamanashi.ac.jp/world/english/

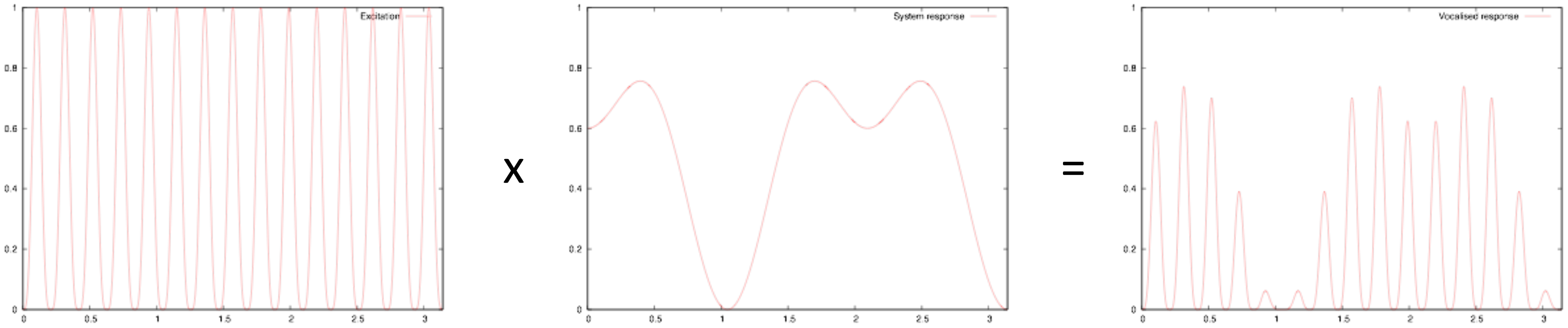
The source-filter model



Donald Ayler, 1942–2007
(with Albert Ayler, 1936–1970)
© The Wire Magazine

Source-filter for speech



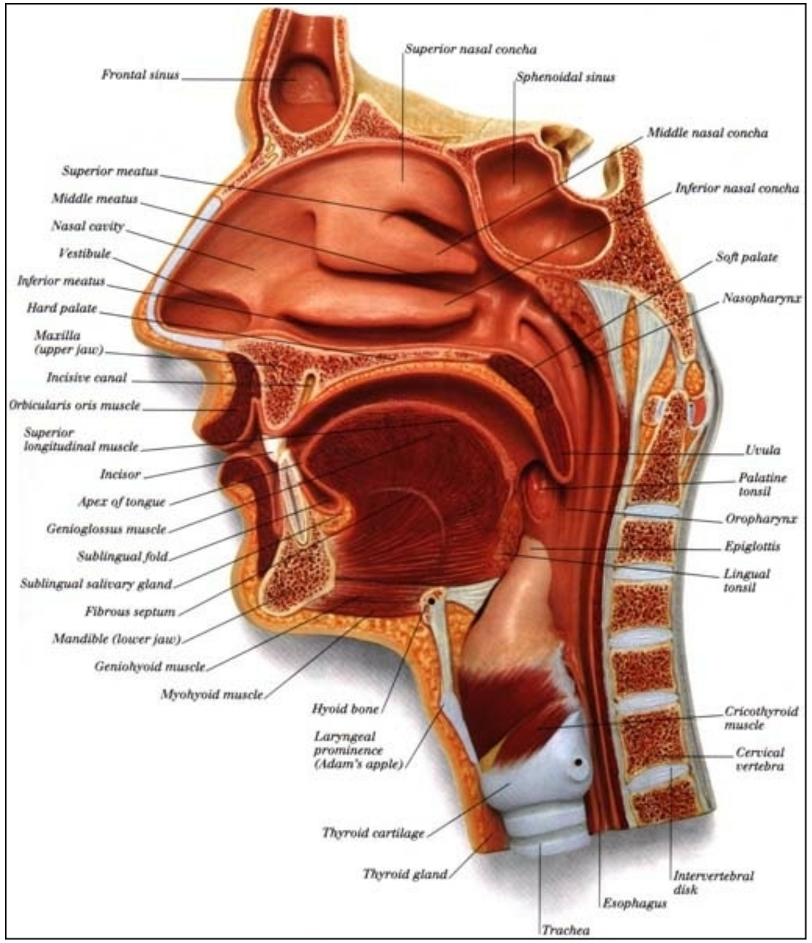


Excitation

System

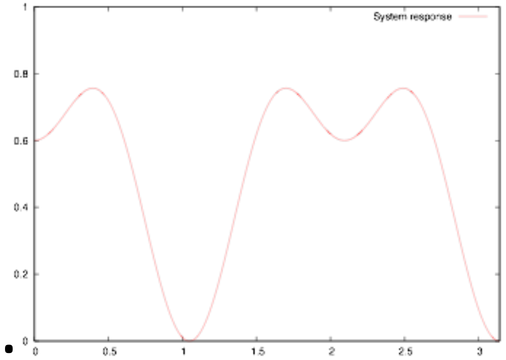
(Simplistic frequency-domain view)

Parametric vocoding



- Parametric vocoding breaks the signal into physiologically motivated components
  - System
    - Spectral shape
  - Excitation
    - Pitch
      - May include voicing detection
    - Harmonic component (think vowels)
      - Depends on voicing
    - Noise component (think consonants)
    - Harmonic to noise ratio
      - Could be band-dependent

Spectral shape

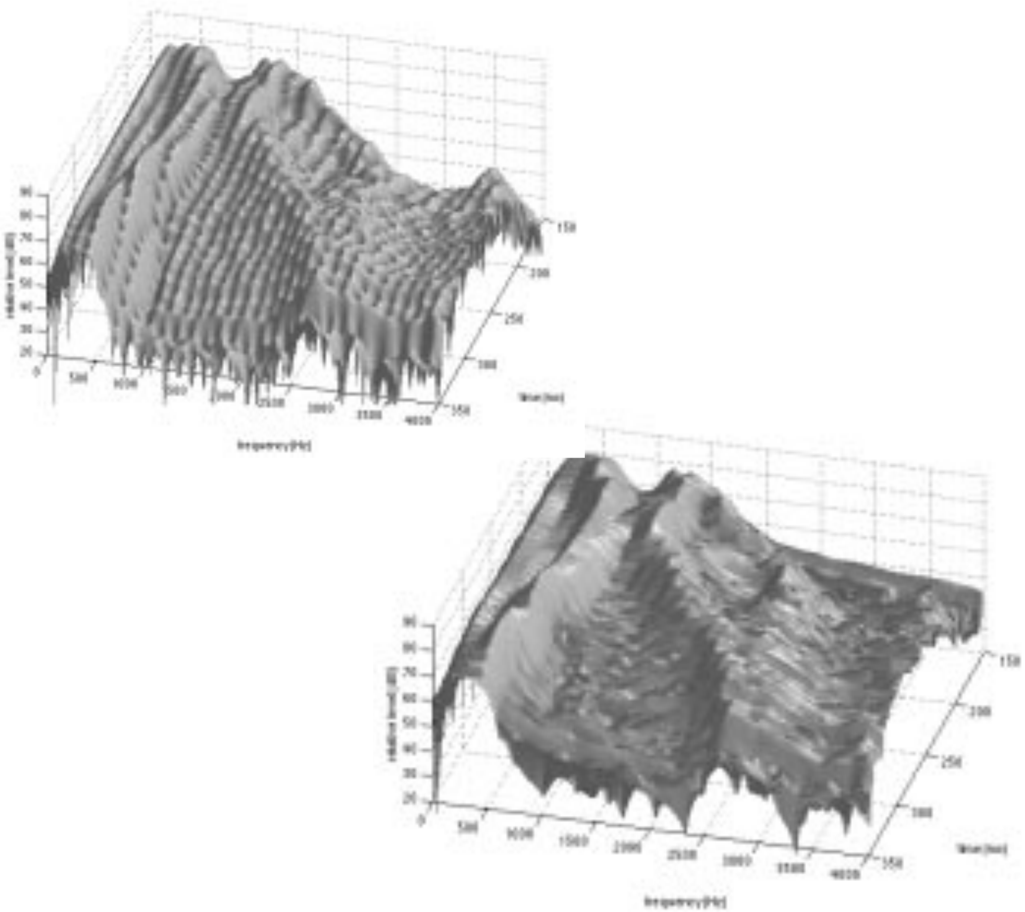


- Spectral shape is quite well understood! Two choices:

- DFT
  - STRAIGHT is a DFT method

- Linear prediction
  - Tends to be lower dimensional
  - Attractive as it has a physiological basis

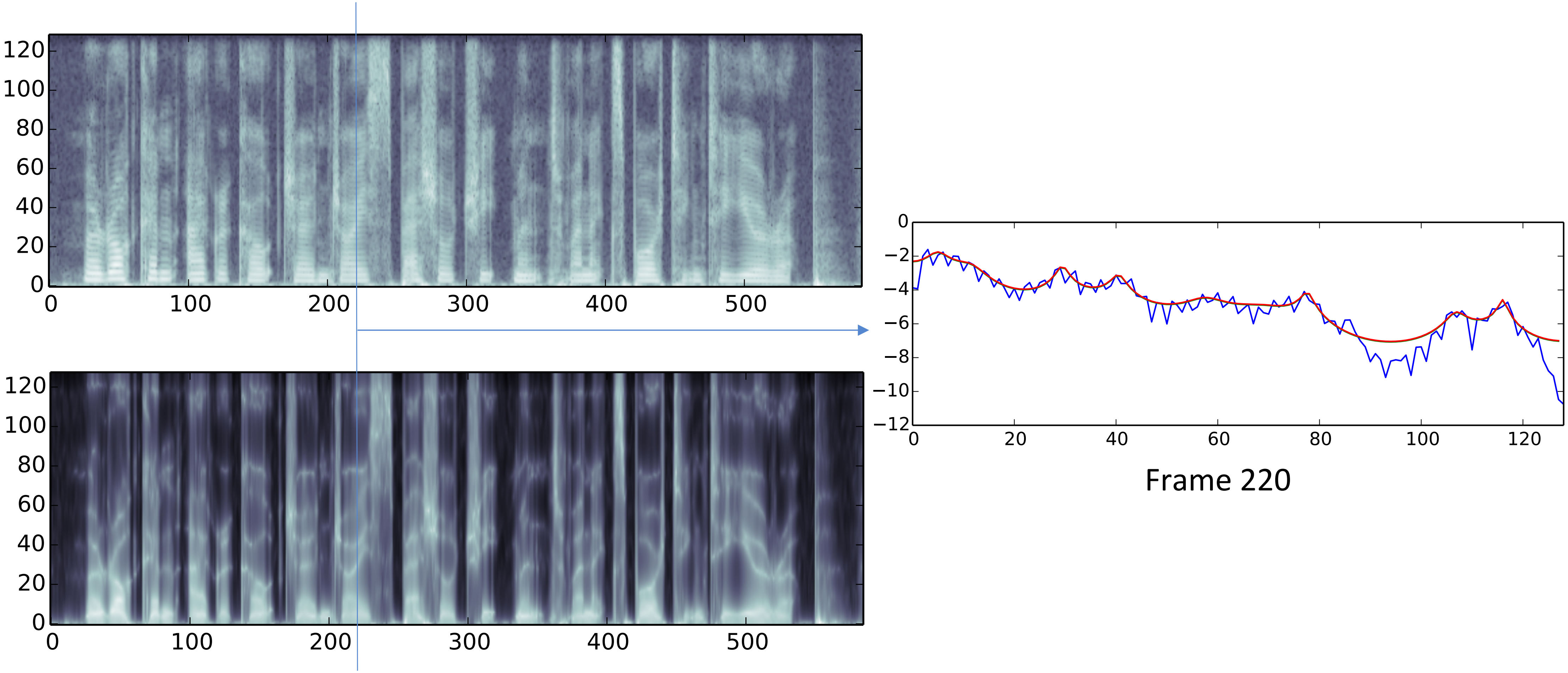
STRAIGHT &c



- STRAIGHT is essentially a pitch-synchronous DFT
  - Gets rid of pitch artefacts
- Later iteration:
  "Tandem STRAIGHT"
- STRAIGHT is known as a vocoder
  - Uses mixed excitation (see later)
  - See also: CheapTrick (Morise, 2015)

Kawahara, Masuda-Katsuse & de Cheveigné, 1998

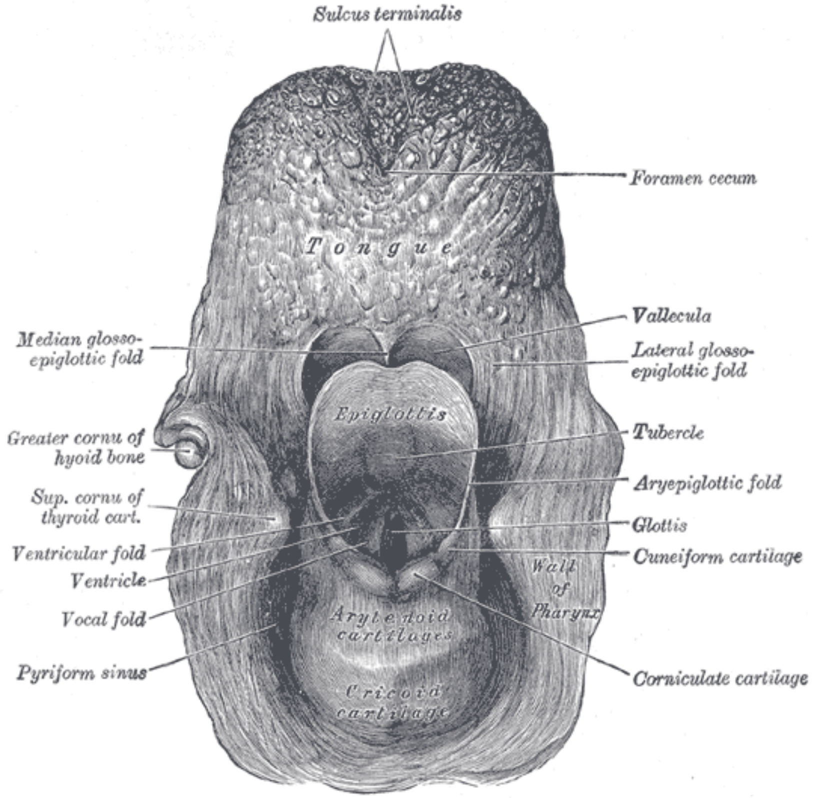
Linear prediction

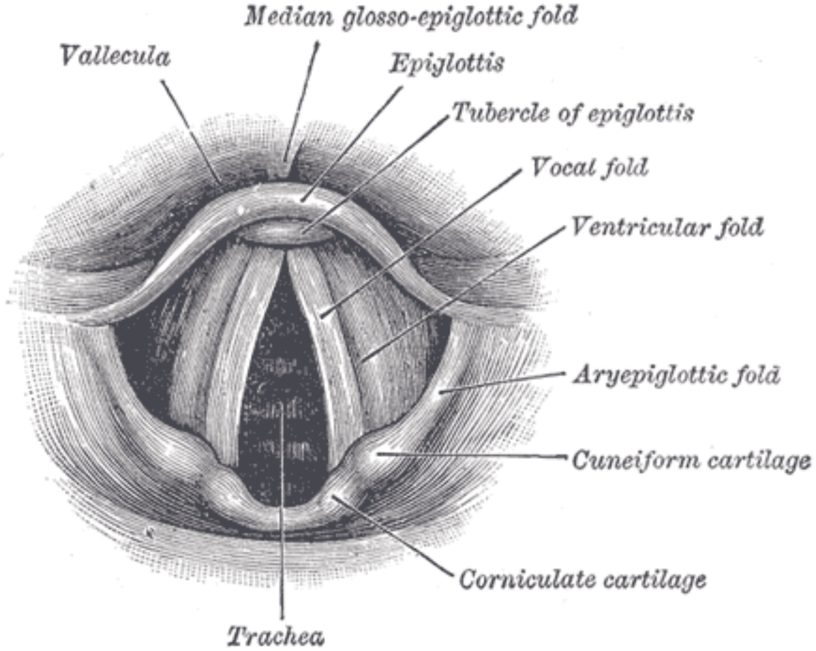


Frame 220

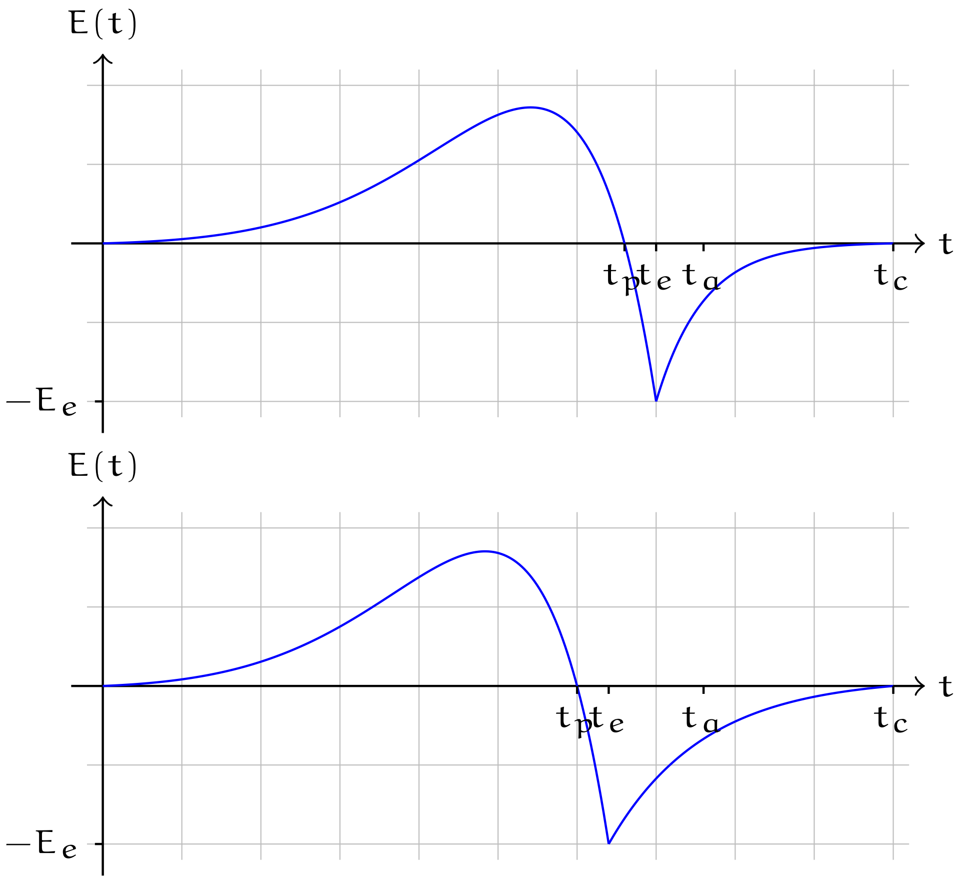
Excitation

Vocal folds do not produce impulses!





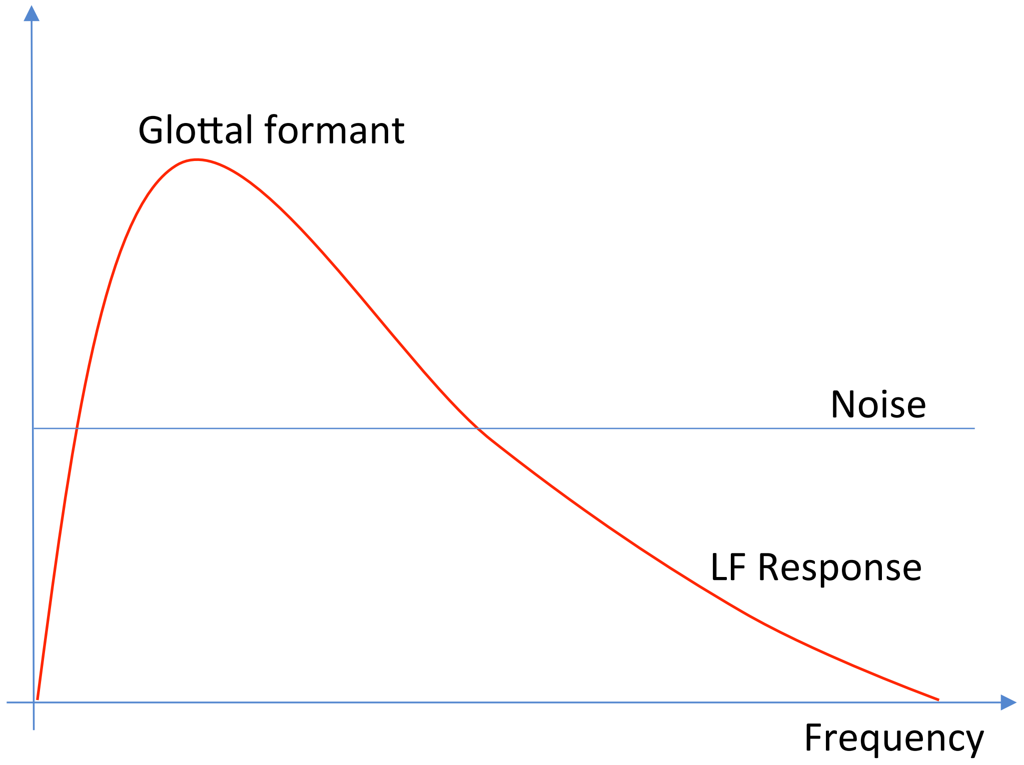
Excitation



- Our best guess is the LF model
  - Opening phase
  - Closing phase
  - Discontinuity when vocal folds come into contact
- This is the glottal flow derivative

Gunnar Fant, Johan Liljencrants, and Qi-Guang Lin.
“A Four-Parameter Model of Glottal Flow”. In: *STL-QPSR* 26.4 (1985). Paper presented at the French-Swedish Symposium, Grenoble, April 22–24, 1985, pp. 001–013.
URL: http://www.speech.kth.se/qpsr

LF in frequency domain



- LF shapes vocalisation, not noise
  - HF is just noise
  - There is cut-off
- The whole thing is subject to system response

Separation of harmonics from noise

- Mixed excitation (Yoshimura et al. 2001)

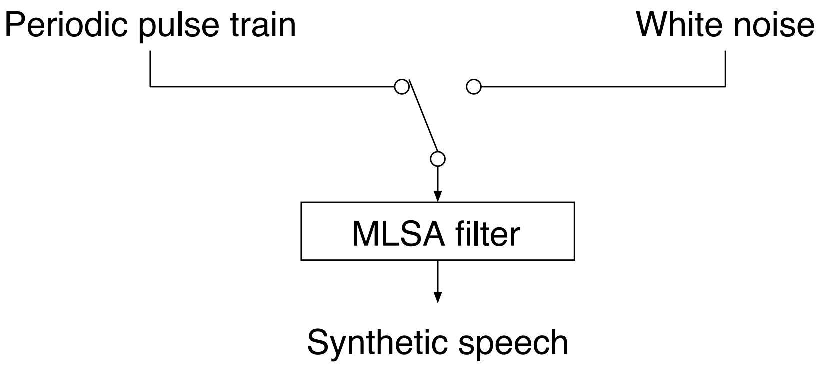


Figure 2: Traditional excitation model.

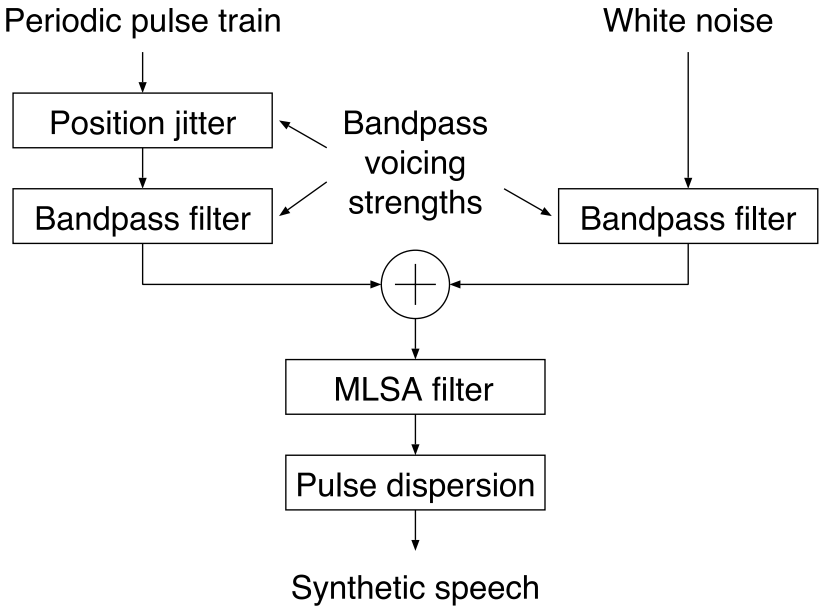
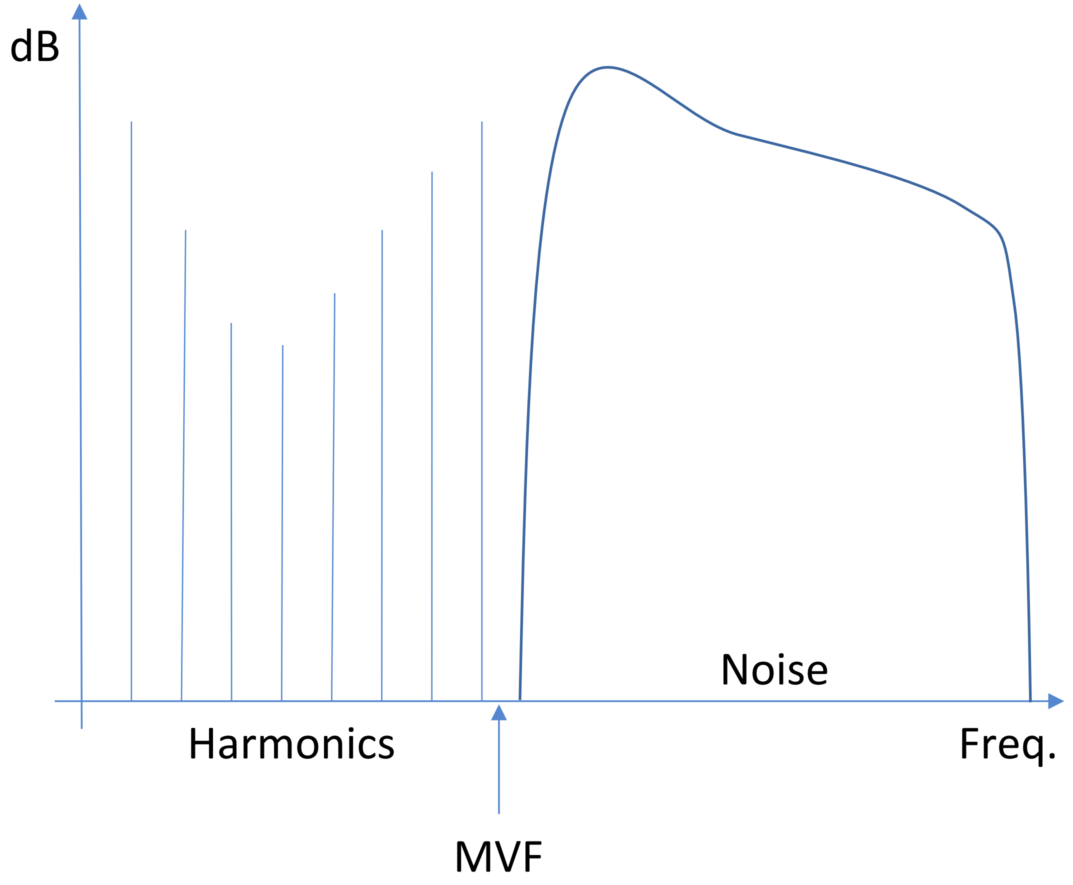


Figure 3: Mixed excitation model.

(MLSA: Mel Log Spectrum Approximation)

Separation of harmonics from noise



- HNM: Harmonic plus noise model (Y. Stylianou, PhD, 1996)
  - Determine a maximum voiced frequency
  - Model everything below that as harmonics
  - Model everything above that as noise
- These models are pretty good!

Part 3

Shortcomings

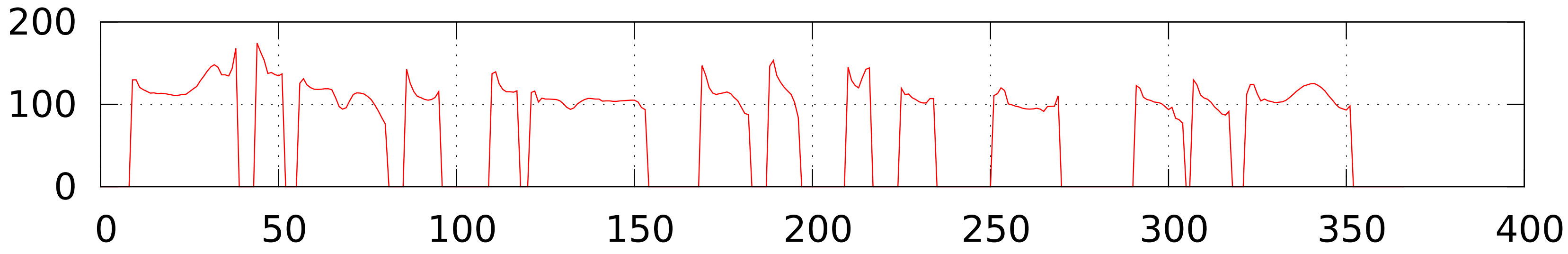
Degradation in vocoding

- The cause(s) of unnaturalness is (are) *not* well understood. There are several plausible theories:

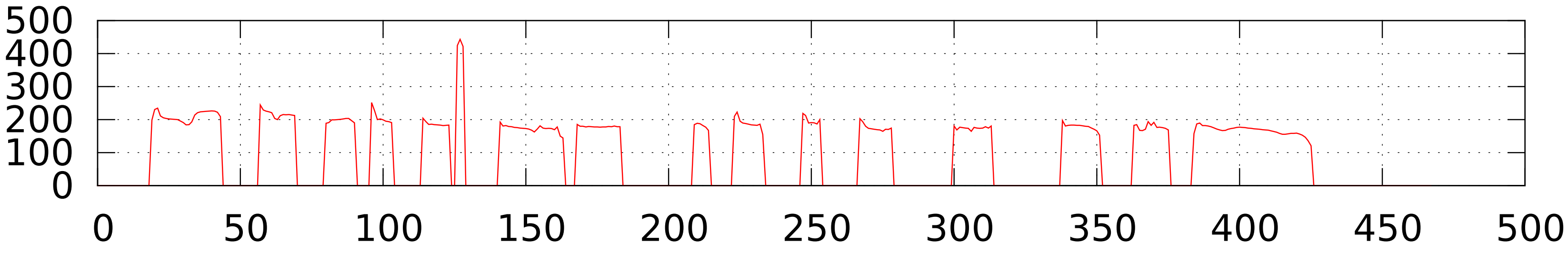
- Voicing detection
  - For sure, incorrect voicing will cause artefacts.
- Incorrect aperiodicity
  - Aperiodicity is the part that is not voicing.
  - Note: voiced fricatives have both periodic and aperiodic components.

- Framing
  - Mains hum is "buzzy"; the buzz is the frame rate.
- Plus:
  - It's worse in TTS than just vocoding.
  - It's worse if you mess with the signal.

Pitch Extractor



- Issues:
    - Not defined for unvoiced segments
    - Tendency towards doubling and halving errors
    - 1001 different algorithms



Is pitch extraction solved?

Junichi’s method:

“f0 is first extracted using a wide range over a whole database, then a range is determined for each speaker and f0 is extracted again using three methods. Finally a median value of the three methods is chosen.”

- So, something is broken!
- A hypothesis
  - Can we fix vocoding by fixing gross pitch errors?

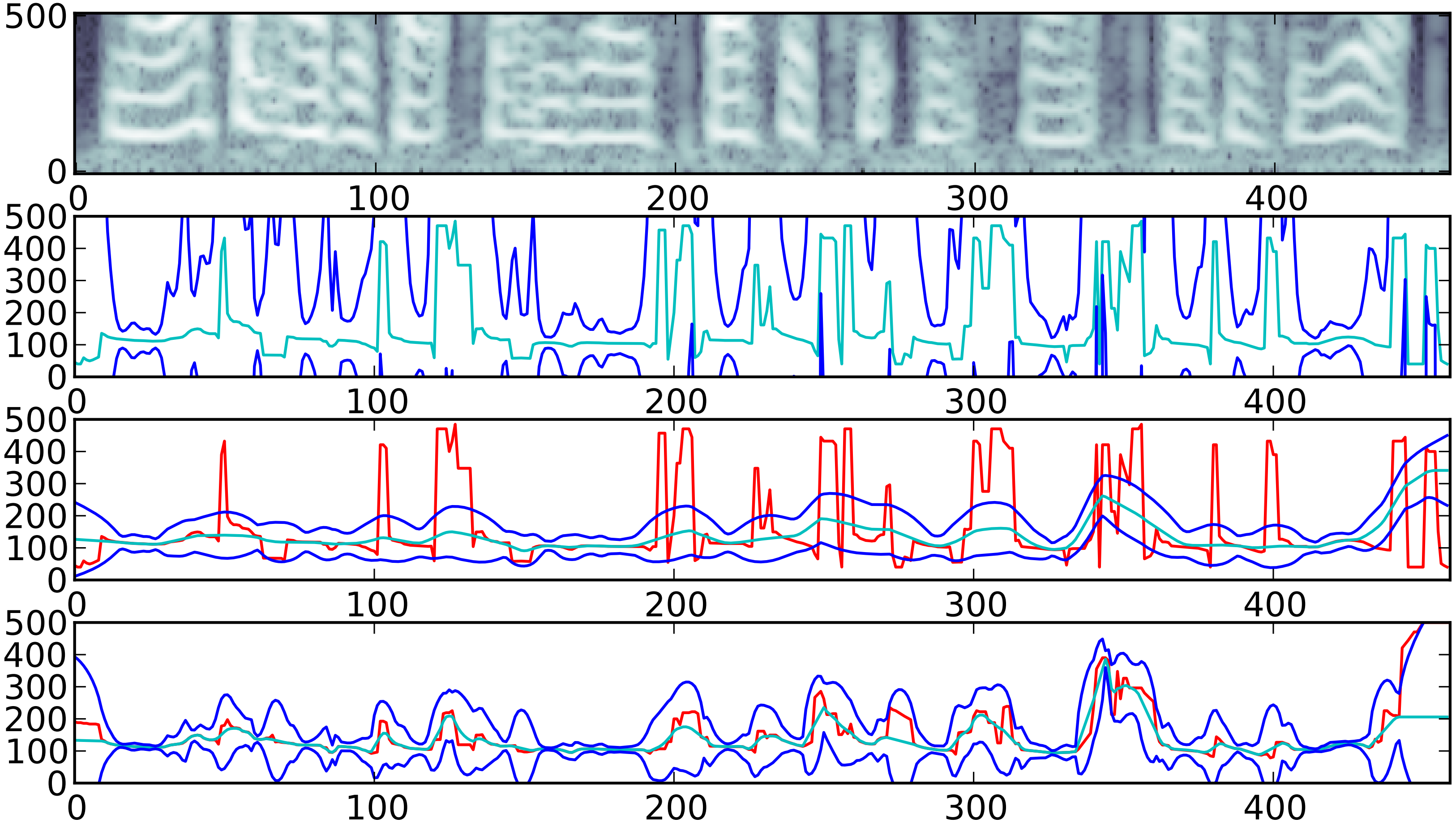
Basic pitch extraction

Paul Boersma. “ACCURATE SHORT-TERM ANALYSIS OF THE FUNDAMENTAL FRE- QUENCY AND THE HARMONICS-TO-NOISE RATIO OF A SAMPLED SOUND”. . In: *Proceedings of the Institute of Phonetic Sciences, Amsterdam*. 17. University of Amsterdam, 1993, pp. 97–110

- Boersma’s algorithm is in Praat
  - Pretty good
  - Essentially autocorrelation, but with lots of extra hacks

- Autocorrelation also gives a harmonic to noise ratio
  - You can convert HNR to a variance on the estimate
  - If so, you don’t need a voicing decision

Use HNR as precision in a Kalman filter



The SSP vocoder

- Given pitch and harmonic to noise ratio, you can build a vocoder! Ingredients
  - Continuous pitch extraction
  - Linear Prediction based spectral shape
  - “Assorted” excitation methods
- Implemented in python using numpy

- Excitation examples
  - Original
    (The natural recording at 22 kHz)
  - Whisper
    (Pure Gaussian noise excitation)
  - Robot
    (Pure impulse excitation, constant pitch)
  - Impulse
    (Impulse & noise excitation, continuous pitch)

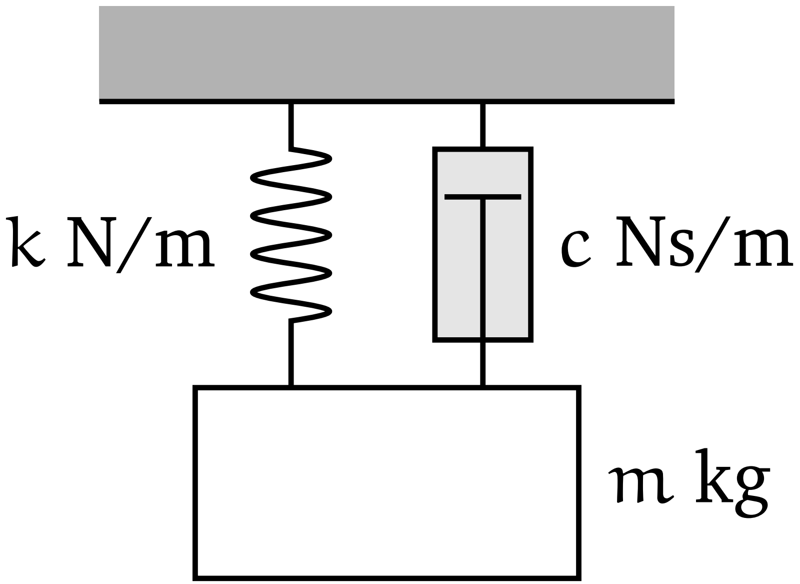
Tentative pitch conclusion

- LPC actually works very well
- Accurate pitch seems not to matter
  - At least, it's not the most important issue
- Rather, the synthetic speech is still "buzzy"
  - It's more likely to be excitation related
- Note: the Kaldi pitch tracker is a good choice too

Philip N. Garner, Milos Cernak, and Petr Motlicek. "A Simple Continuous Pitch Estimation Algorithm". In: *IEEE Signal Processing Letters* 20.1 (Jan. 2013), pp. 102–105. DOI: 10.1109/LSP.2012.2231675

Pegah Ghahremani et al. "A Pitch Extraction Algorithm Tuned for Automatic Speech Recognition". In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. Florence, Italy, May 2014, pp. 2513–2517

LF automatic parameter extraction



- LF opening phase is an exponential sinusoid

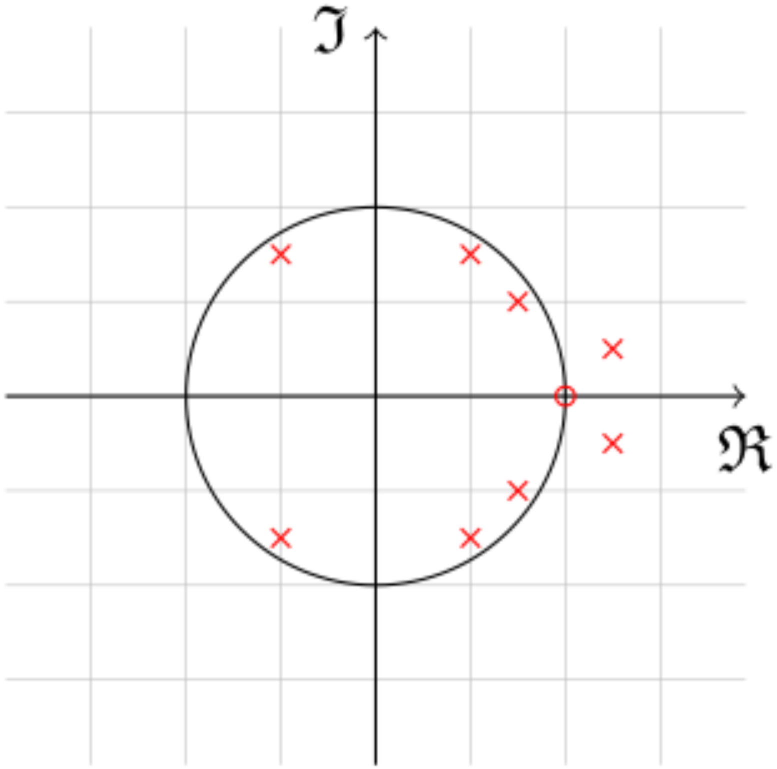
$$\frac{dU_g(t)}{dt} = E(t) = E_0 e^{\alpha t} \sin(\omega_g t)$$

- This is the same as a second order system

$$Y(s) = \frac{kX(s)}{ms^2 + cs + k}$$

$$Y(s) = \frac{\omega_0^2 X(s)}{\underbrace{s^2 + 2\zeta\omega_0 s + \omega_0^2}_{\text{Two poles}}}$$

Rough production model



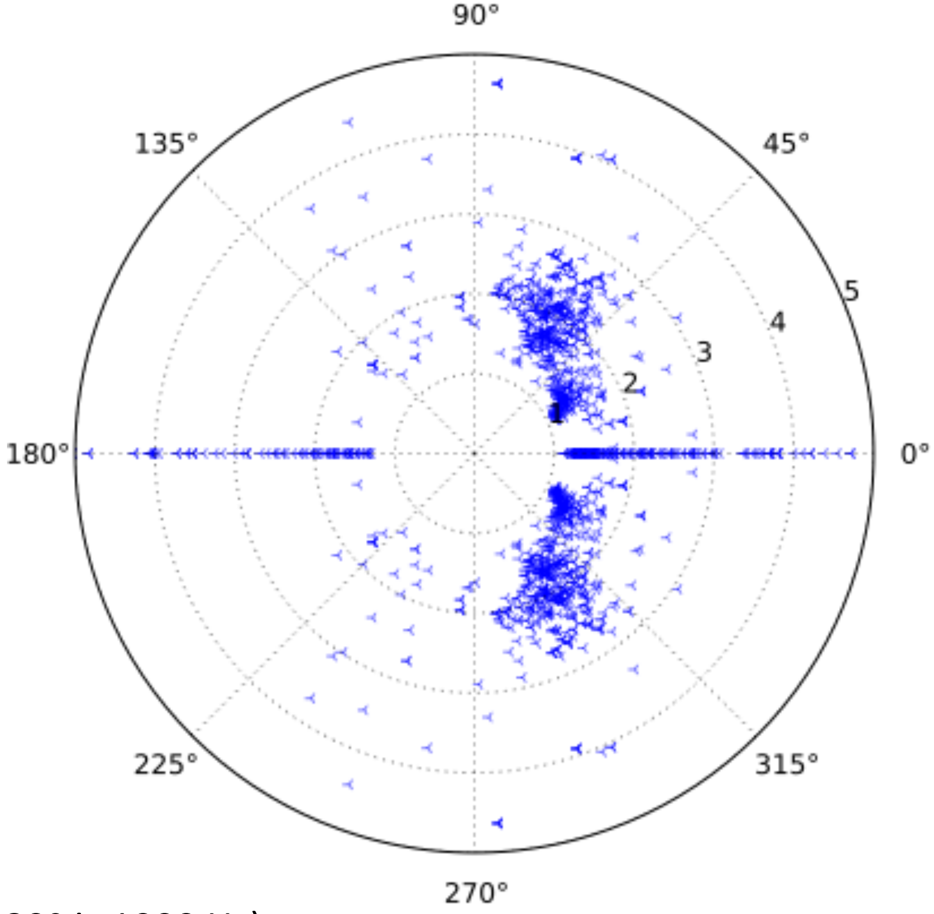
- There is a minimum / maximum phase distinction
  - Formants are inside the unit circle
  - Excitation is outside the circle
    The *glottal* formant
  - Usual (real) signal processing does not distinguish min/max phase
  - …but *complex cepstrum* does

Algorithm

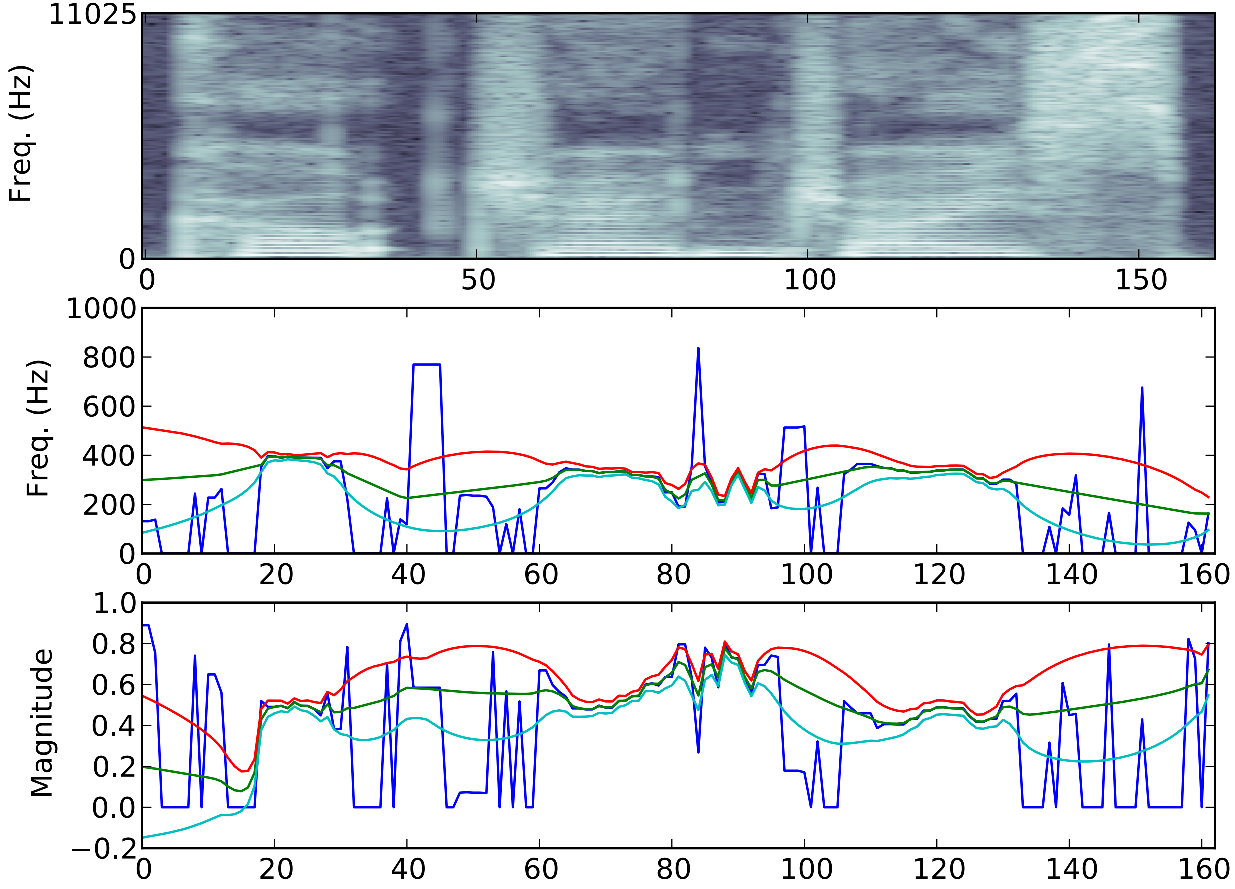
- Calculate complex cepstrum
- Replace min-phase portion with max-phase portion
- Invert (one sided) cepstrum, auto-correlation, order two LP

    - See also: R. Maia et al. / Speech Communication 55 (2013)

Glottal formant



(180º is 1000 Hz)



Examples

- The file that I was using during development
  - EMIME sample 🔊
- Milos got some rather good results with a LibriVox voice
  - Luke original 🔊
  - Luke STRAIGHT 🔊
  - Luke glottal 🔊

Excitation conclusion

- A better excitation model can get rid of buzzy character
  - Key seems to be in low pass filtering the impulsive part
    This LPF is represented by the LF model
  - Only makes sense when added to noise
  - Currently has a few issues
- Automatic extraction is tricky
  Gives all but one parameter
  - Introduces other distortions
  - Defined as differential glottal flow
    Actual glottal flow would be better

Closing

- Conclusions
  - Pitch and HNR lead to a Kalman smoothed continuous pitch estimate
  - Vocoding without a voicing decision is easier than with
  - Glottal excitation is available from the complex cepstrum
  - Glottal modelling leads naturally to a maximum voiced frequency
- Corollaries
  - It's all available open source in python
    https://github.com/idiap/ssp
  - Some in C++
    https://github.com/idiap/libssp