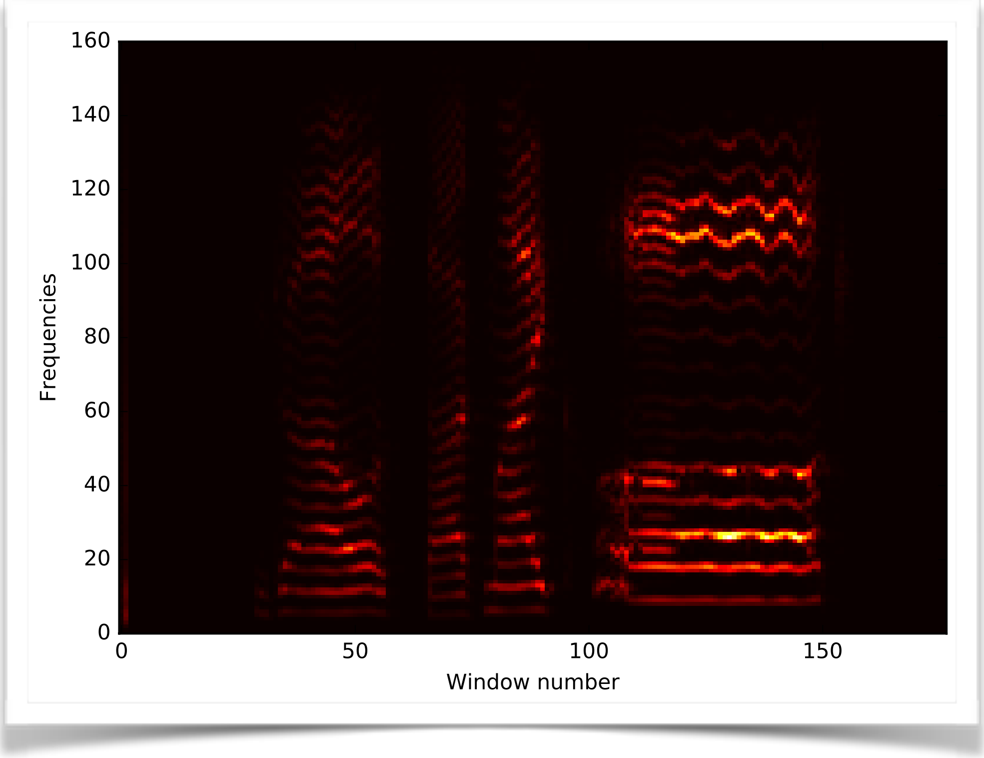
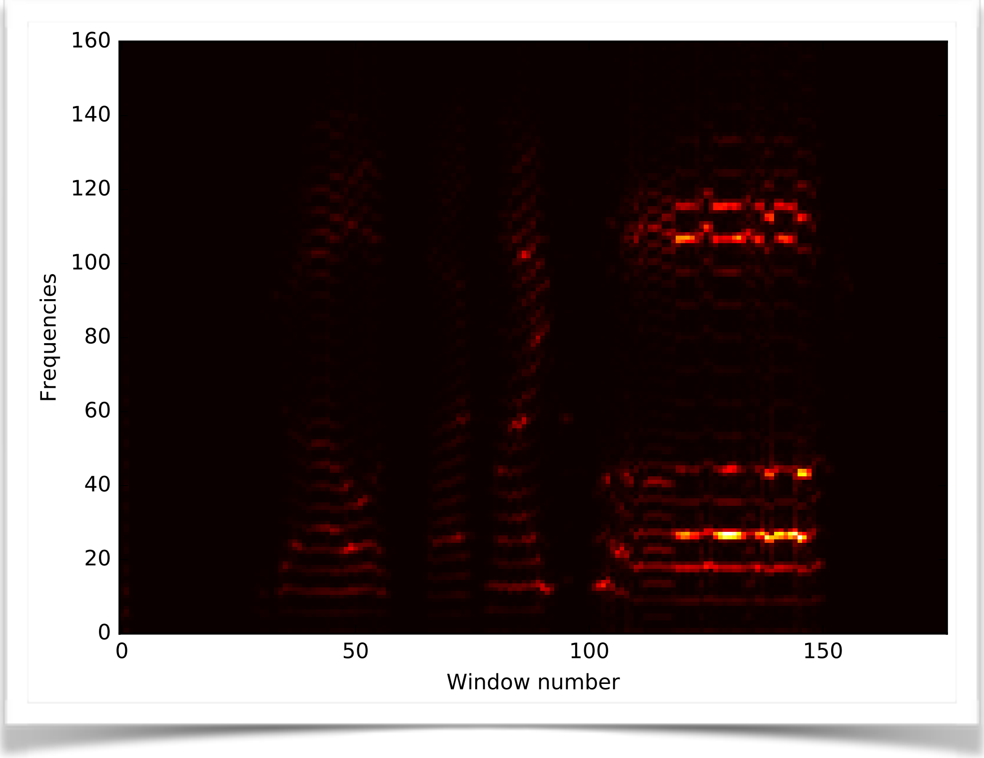
Speech Compression

Benjamin Villalonga - Department of Physics
UIUC
Algorithms Interest Group - 02-21-2017





Types of Audio Compression

- Lossless
  - General (lossless) signal compression algorithms apply
  - Optimized for a type of signal (music, speech, …)
- Lossy
  - General (lossy) signal compression algorithms apply
  - Optimized for a type of signal (music, speech, …)
  - **More important:** Optimized for human perception!

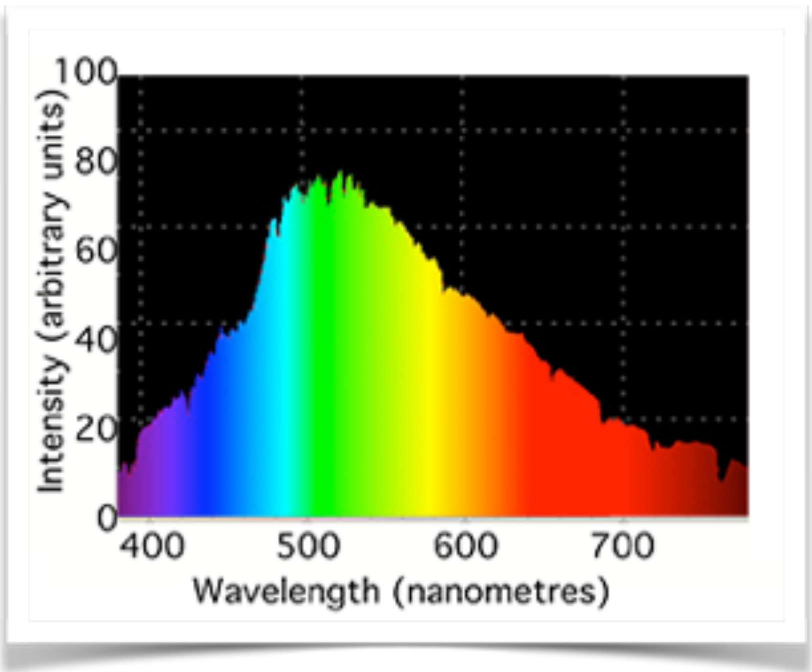
Principle:
Drop everything that will not be heard by a human

Types of Audio Compression

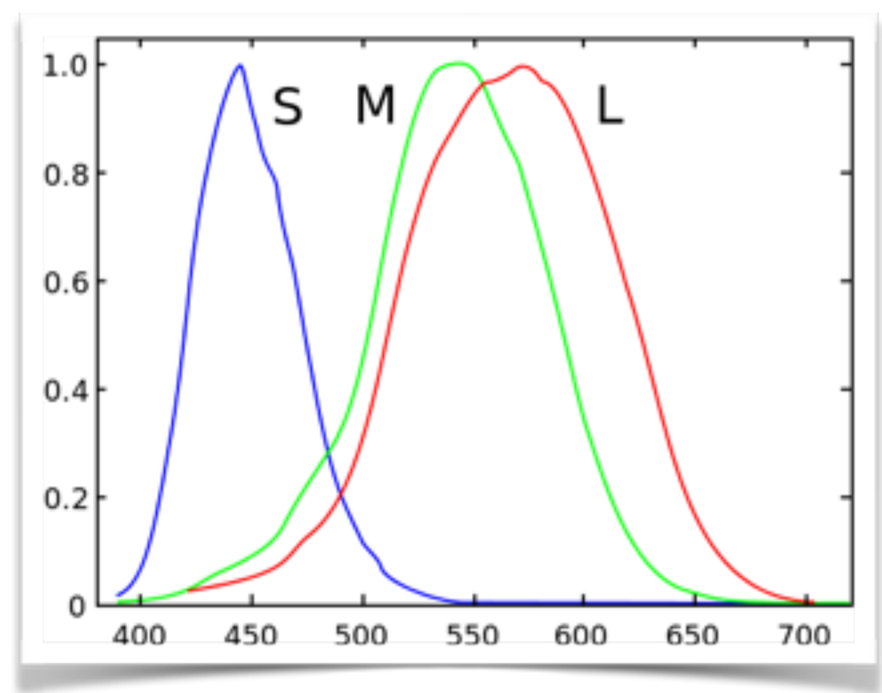
- Lossless
  - General (lossless) signal compression algorithms apply
  - Optimized for a type of signal (music, speech, …)
- Lossy
  - General (lossy) signal compression algorithms apply
  - Optimized for a type of signal (music, speech, …)
  - **More important:** Optimized for human perception!

Principle:
Drop everything that will not be heard by a human

Example with colors



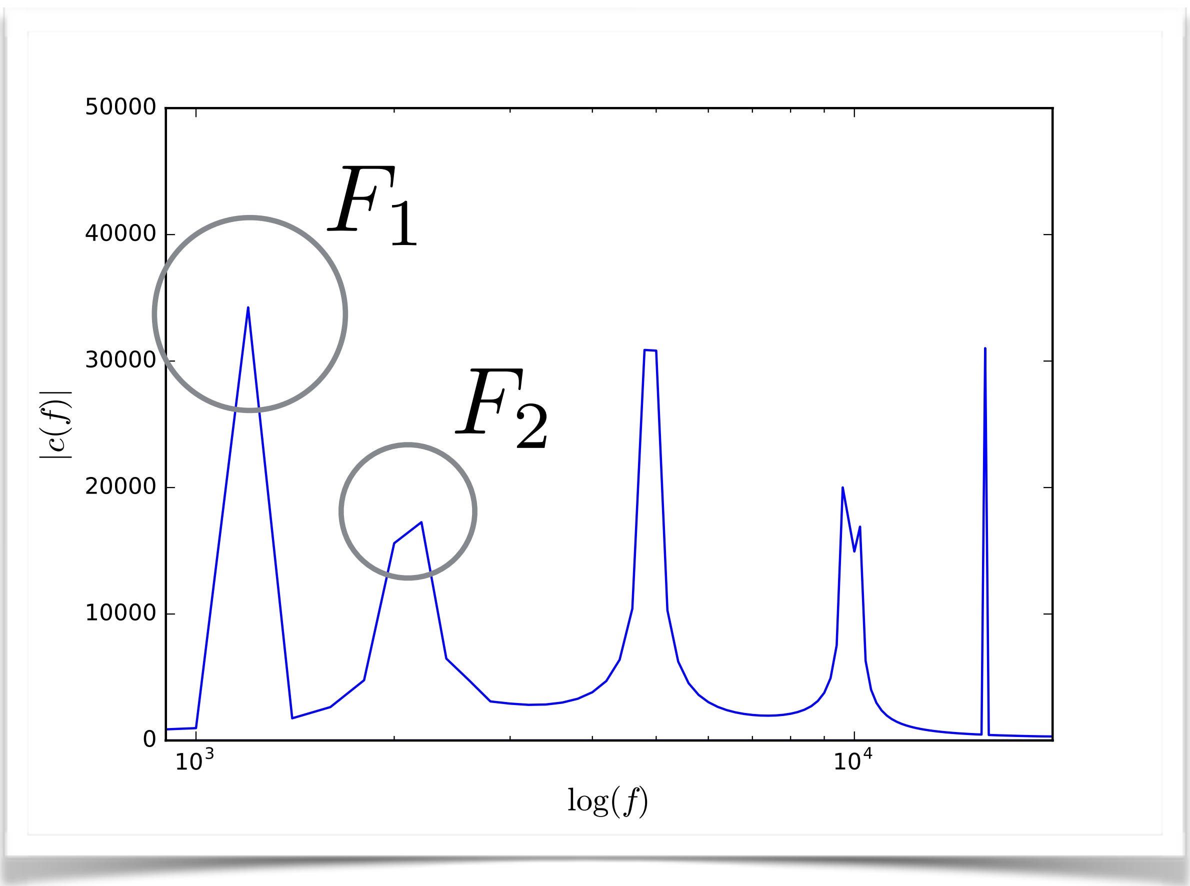
Visible radiation is an infinite dimensional quantity.



However, RGB seems to reproduce all colors (for humans).

Huge compression of information!

Example with speech

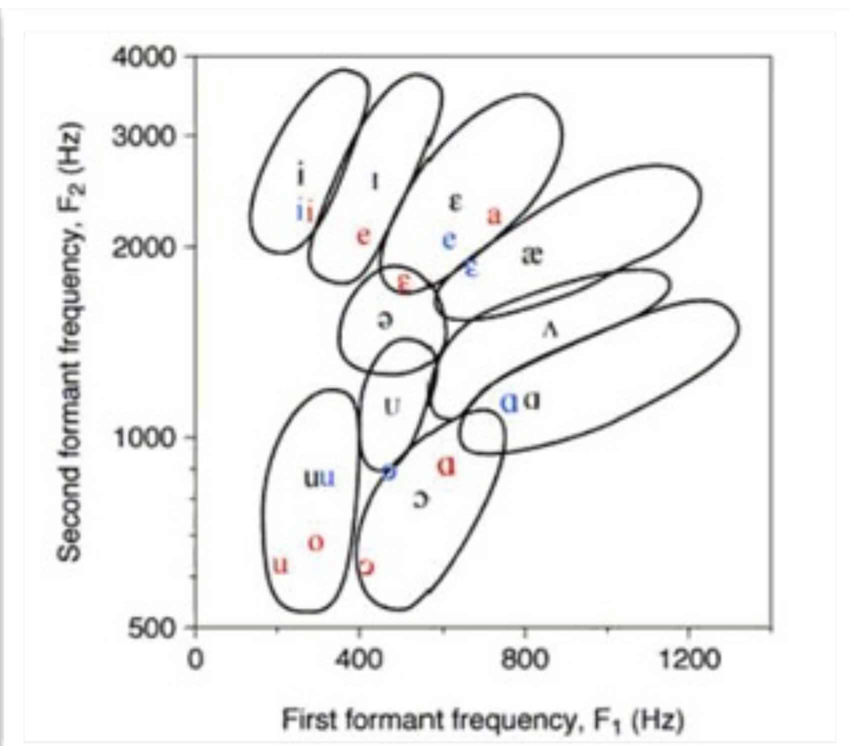


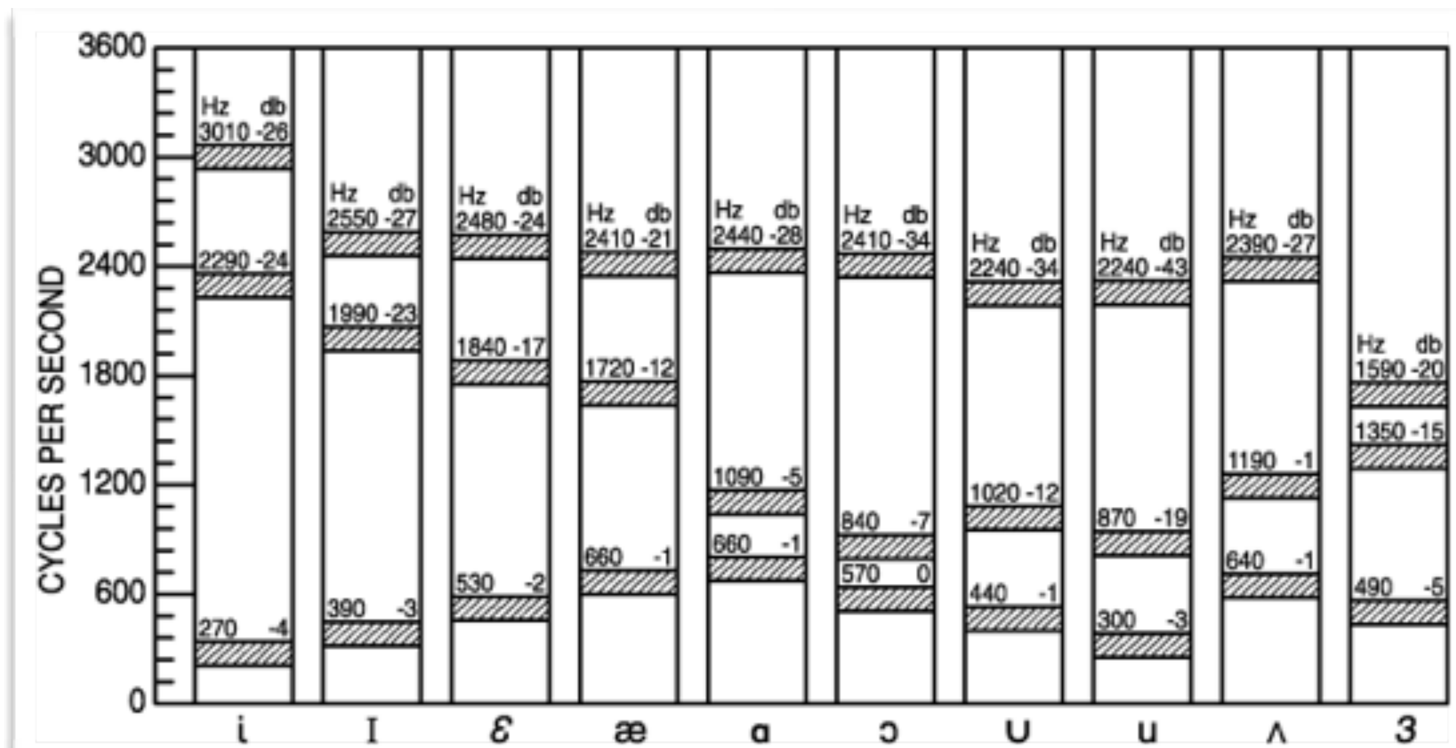
~~Each person has a footprint.~~

Each (person + vowel) has a footprint.

Formants

Frist 2 formants tell us a lot …





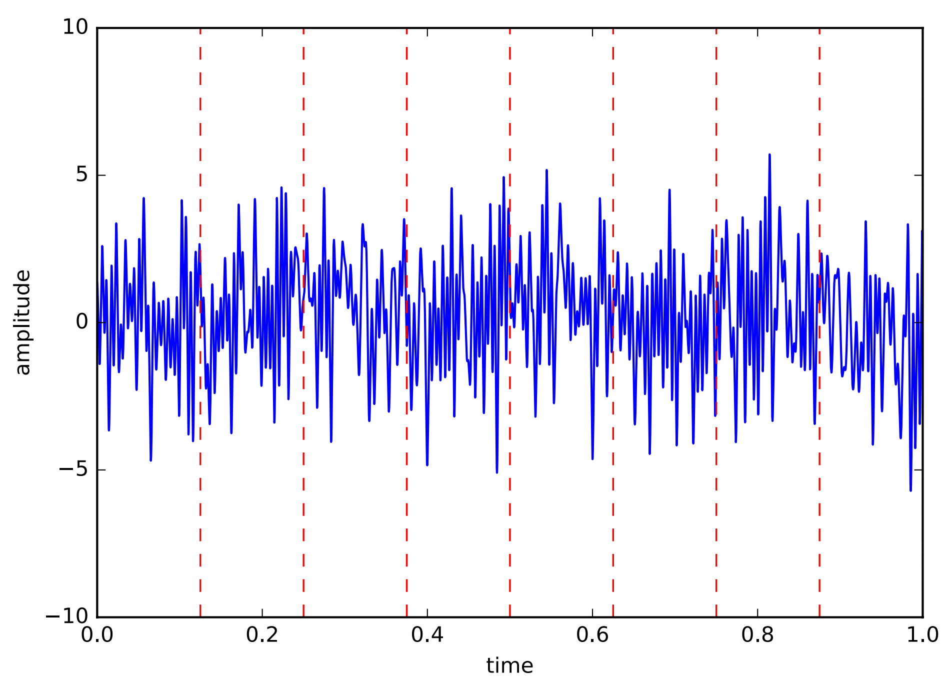
Whatever we do to compress, keep formants' information!

Linear Predictive Coding (LPC)

$$\tilde{x}_i = \sum_{k=1}^{p} a_k x_{i-k}$$

Model of order **p**

Assumption: Only storing the **p a**'s and the first **p** values of x_i is sufficient to reconstruct an entire signal.

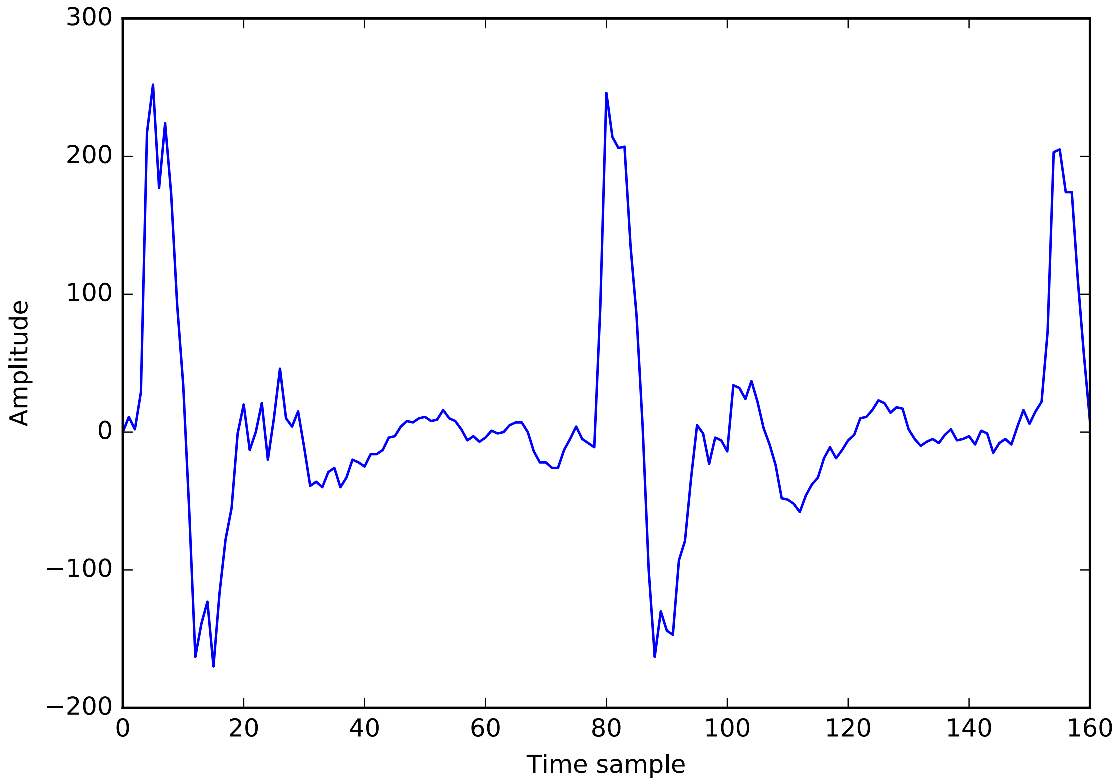


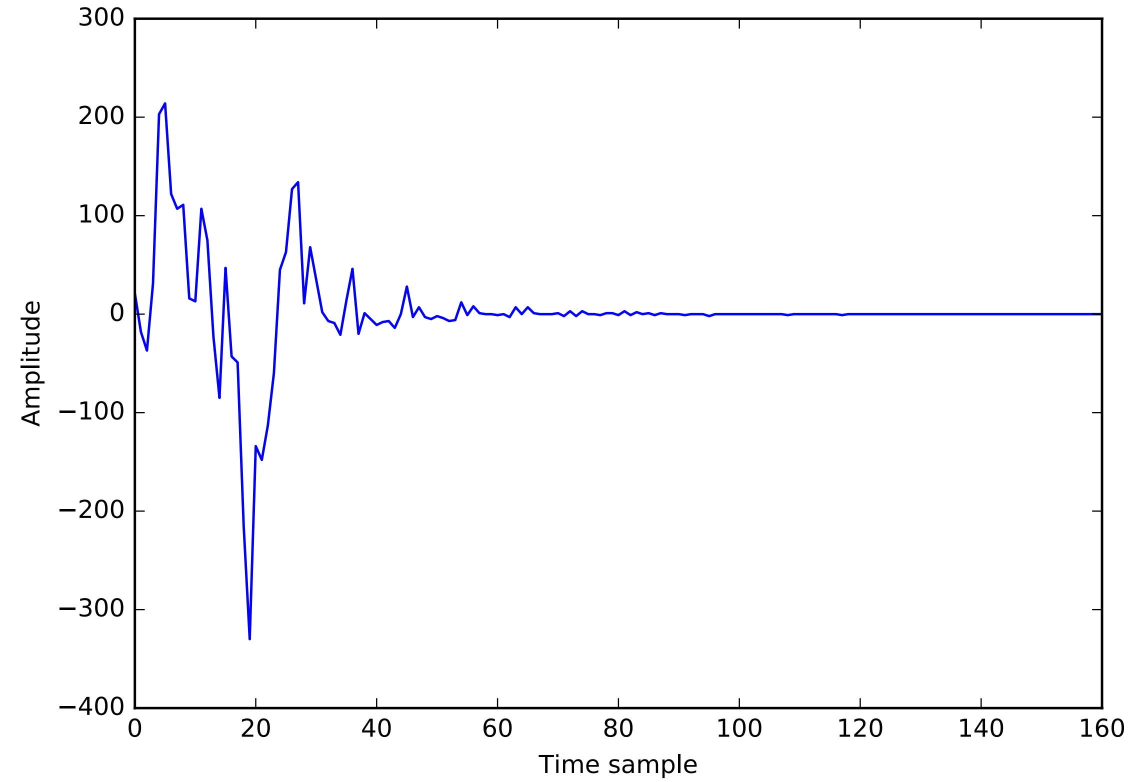
Apply this assumption to small *windows* ($\sim 30ms$).

Optimize **a**'s for each window.

(Appendix A)

By experience...





Signal decays for optimal **a**'s.

$\Downarrow$

Excite periodically.

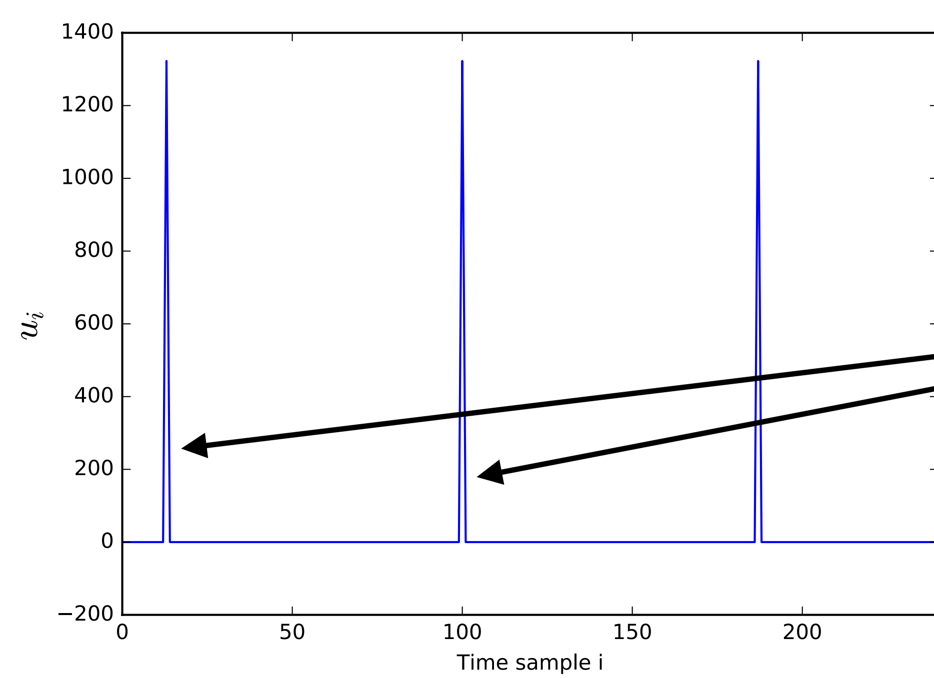
Add excitation

$$\tilde{x}_i = \sum_{k=1}^{p} a_k x_{i-k} + \boxed{G u_i}$$

Background excitation.

G - magnitude of excitation (gain) (Appendix B)

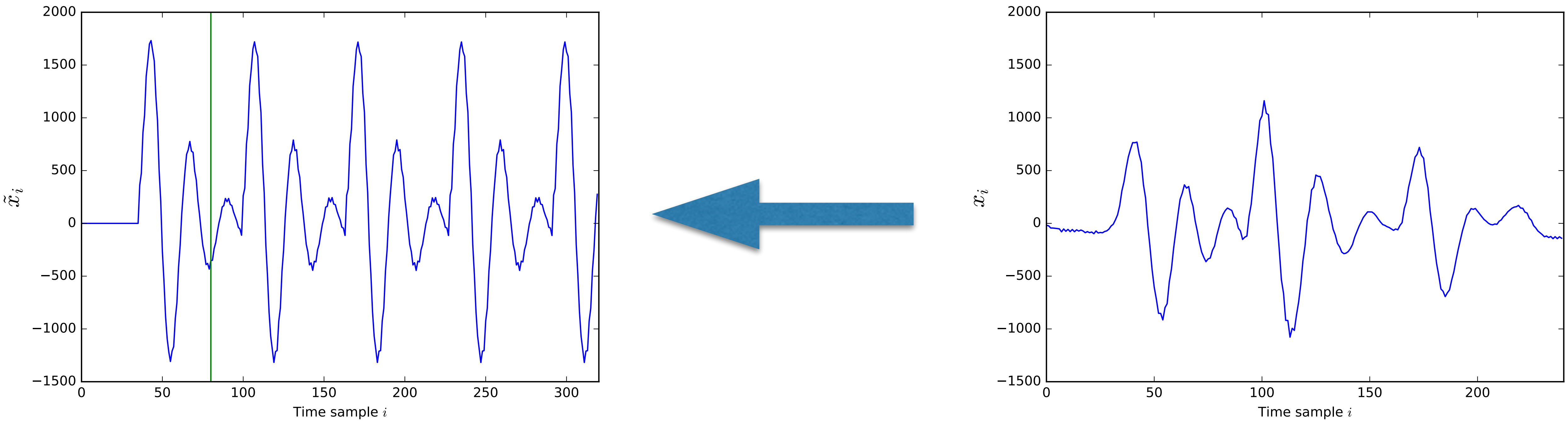
u - profile of excitation



Fundamental period

(Appendix C)

Exciting the signal



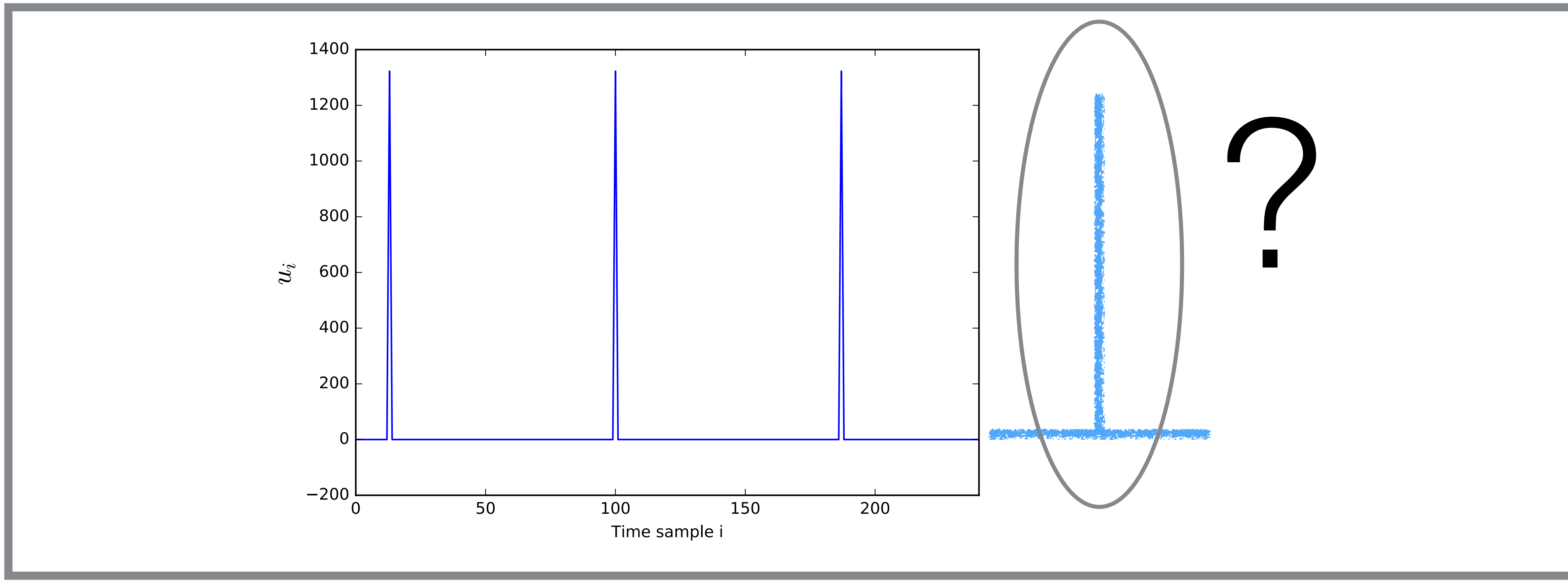
In this case **G** = 363 & **p** = 10.

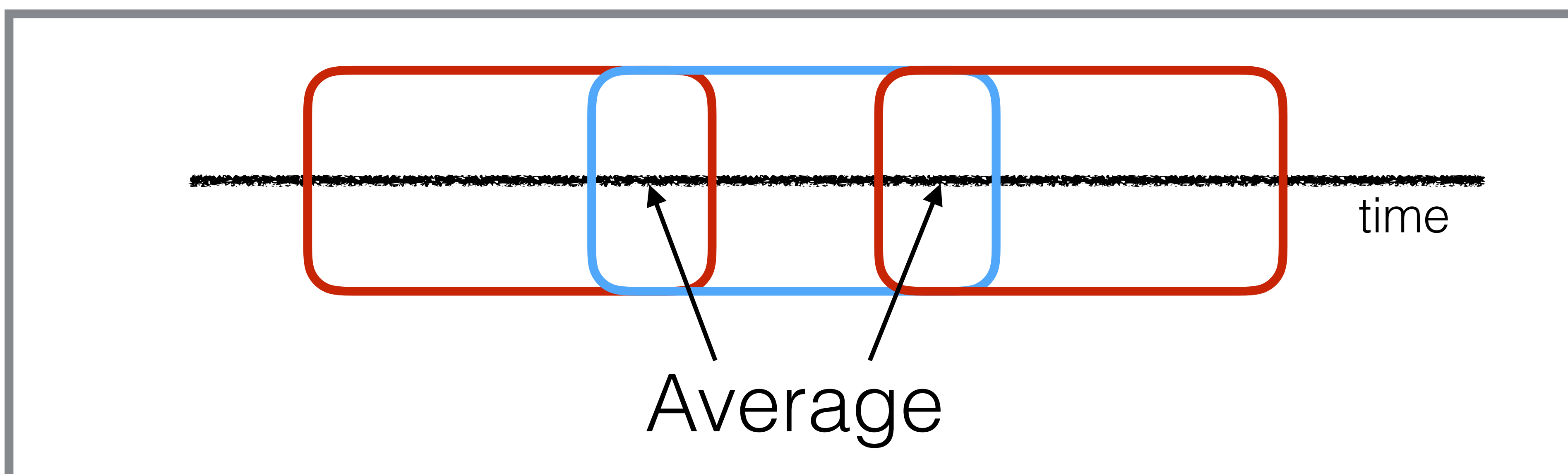
Good news:

No need to store **x**'s!

- This is for *voiced* window
- For *unvoiced* (p, t, g, …): white noise

Two more things





Storage

For each window only need to store:

- **p** coefficients
- one gain **G**
- one fundamental period **T**

As opposed to **n** samples per window.

Appendix A: Optimal **a**'s

$$\tilde{x}_i = \sum_{k=1}^{p} a_k x_{i-k}$$

Want to minimize error: $E = \sum_i e_i^2 \qquad e_i = x_i - \tilde{x}_i$

$$\boxed{e_i \text{ orthogonal to } x_i}$$

$$R_k = \sum_{i=0}^{N-k} x_i x_{i+k} \qquad k = 0, ..., p$$

$$R_{matrix,kl} = R_{|k-l|} \qquad \boxed{R_{matrix,kl} \cdot a_l = R_k}$$

Appendix B: Getting the gain **G**

$$\boxed{x_i \text{ same energy as } \tilde{x}_i}$$

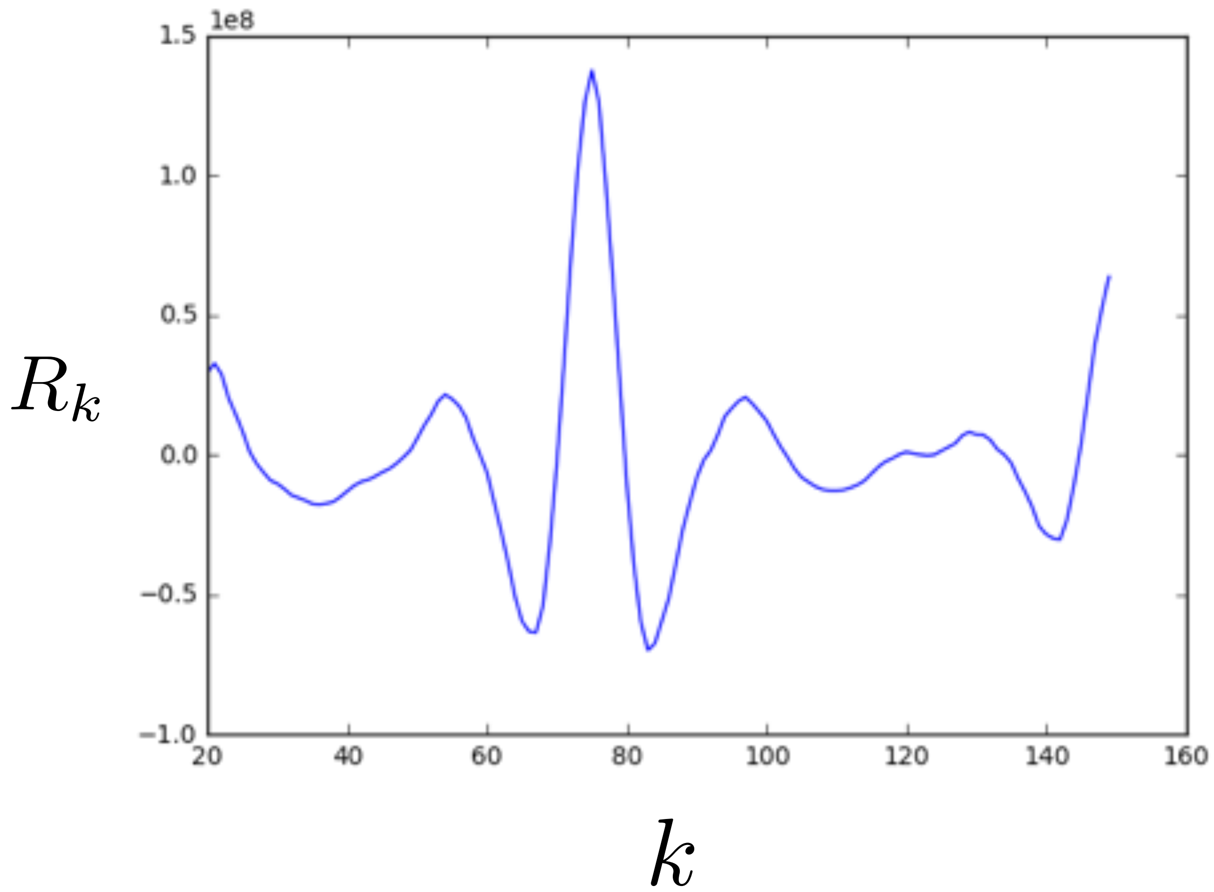
$$\sum_{i=0}^{N} x_i^2 = \sum_{i=0}^{N} \tilde{x}_i^2$$

$$\Downarrow$$

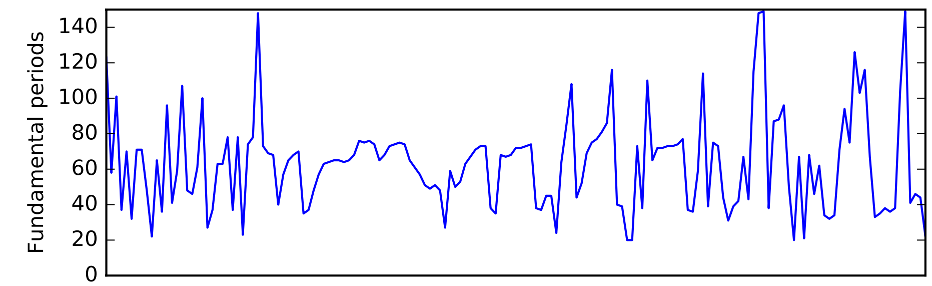
$$\boxed{G^2 = R_0 - \sum_{k=1}^{p} R_k}$$

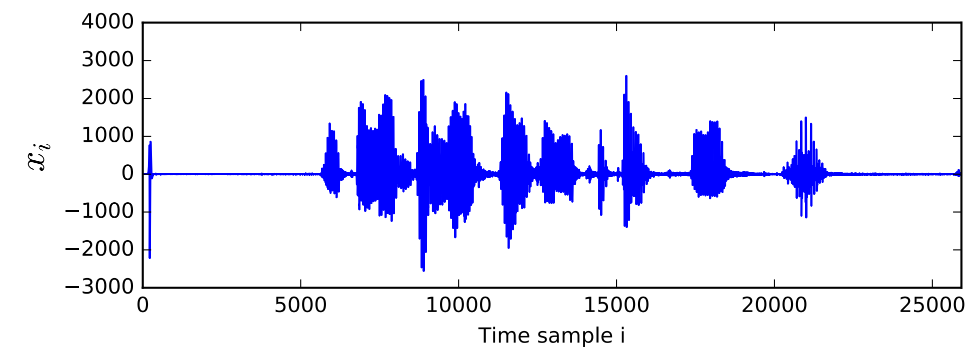
Appendix C: Fundamental Period

Highest peak in periods:



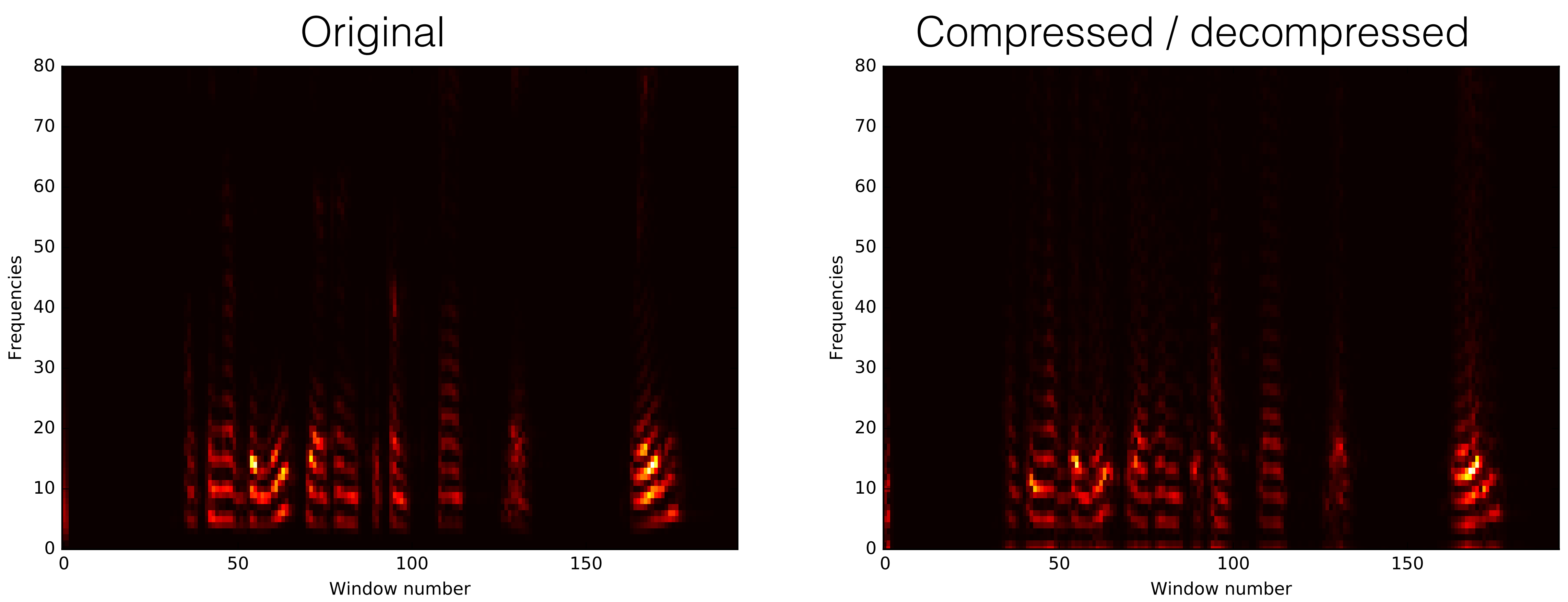
Allow fundamentals down to 50 Hz





Results

- Speech is reproduces amazingly well
- Compression rates of up to 80

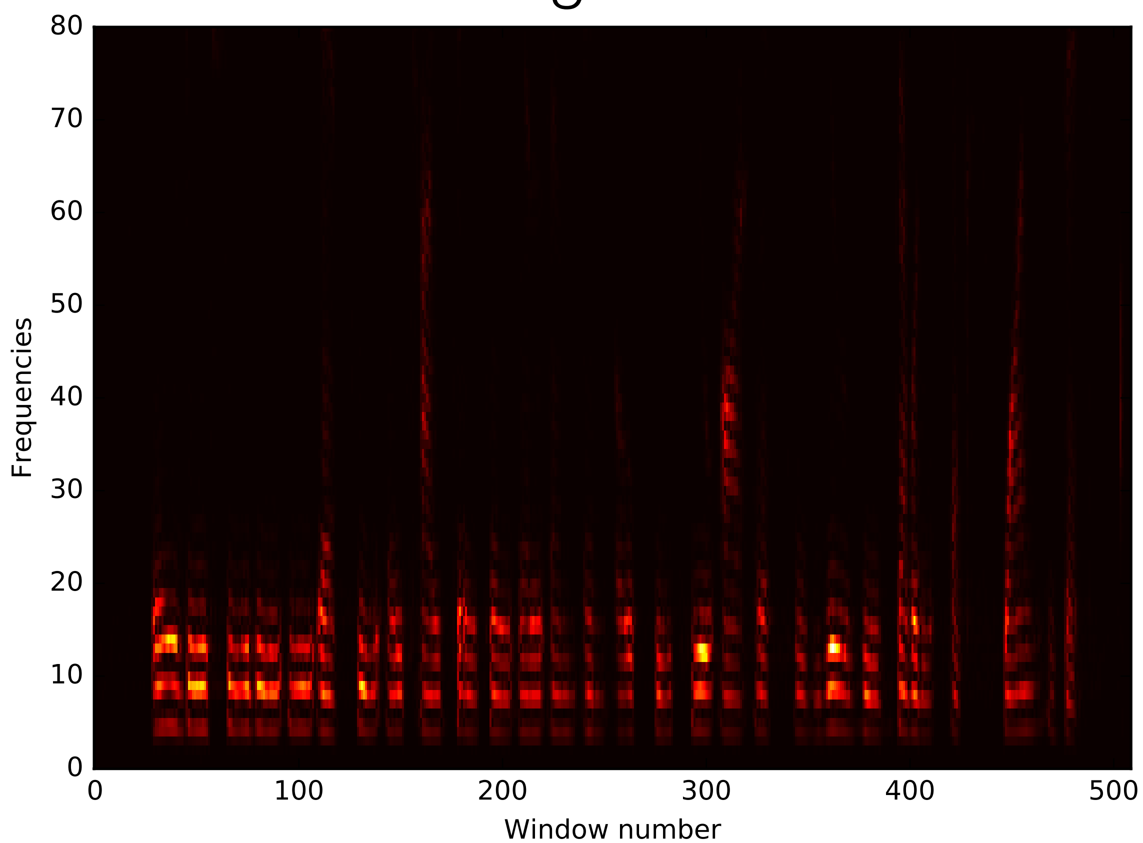


p = 5 vs. **n** = 160 $\Rightarrow$ **compression rate** = 32

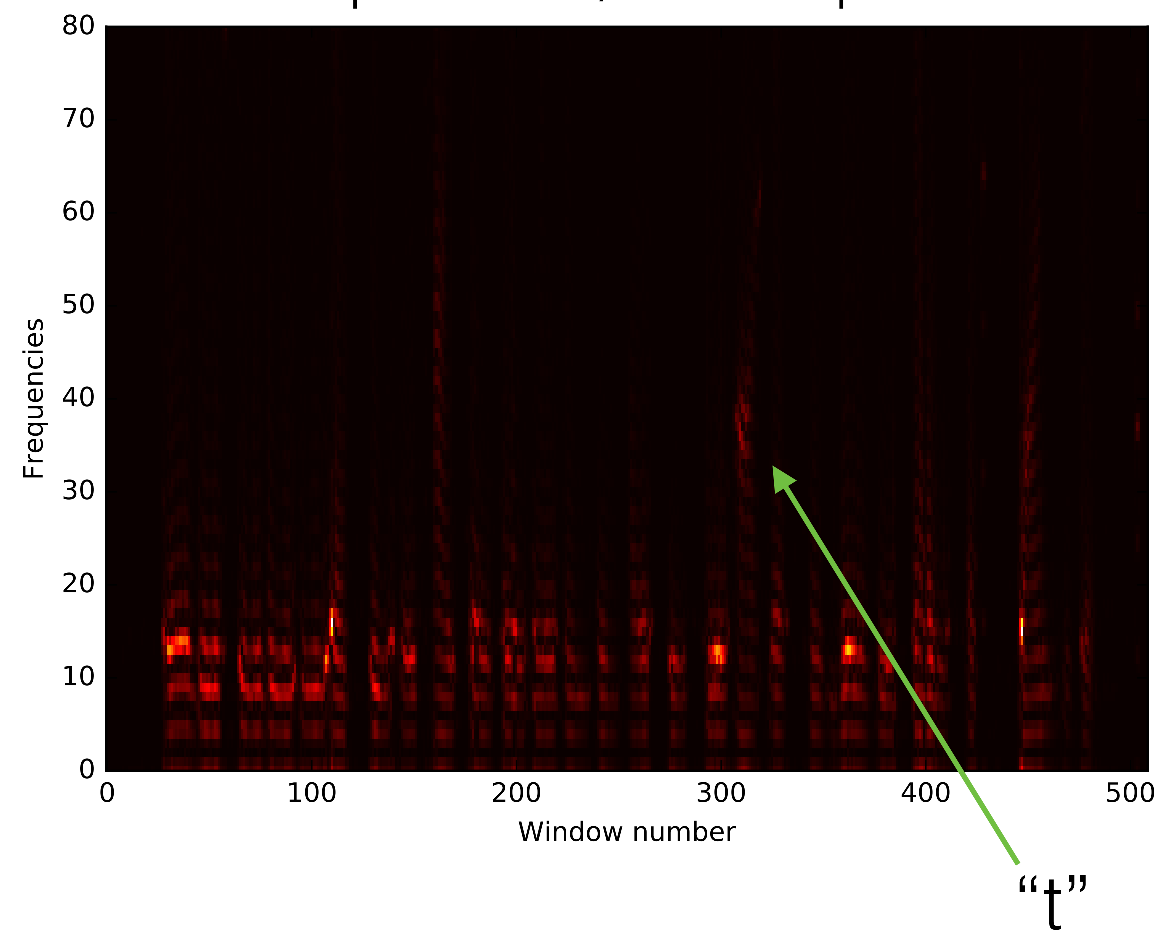
Results

- Consonants are hard (alphabet)

Original



Compressed / decompressed

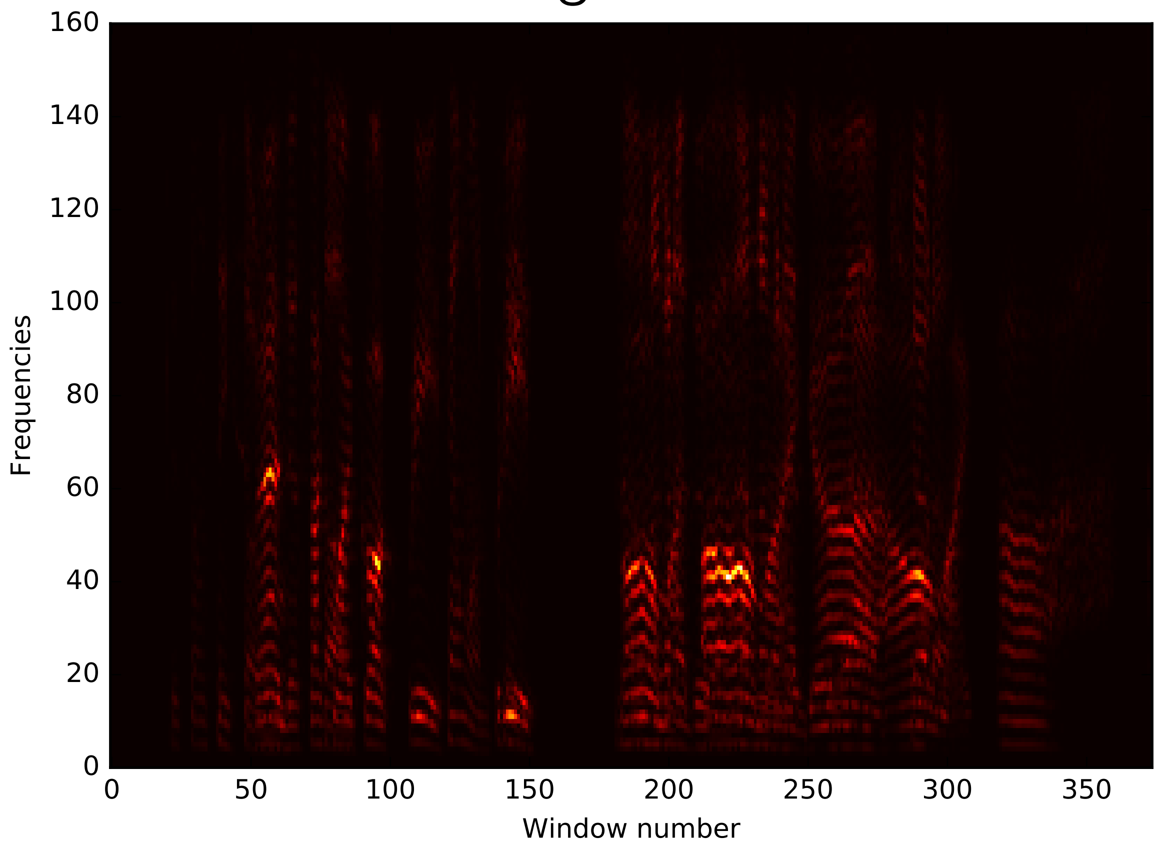


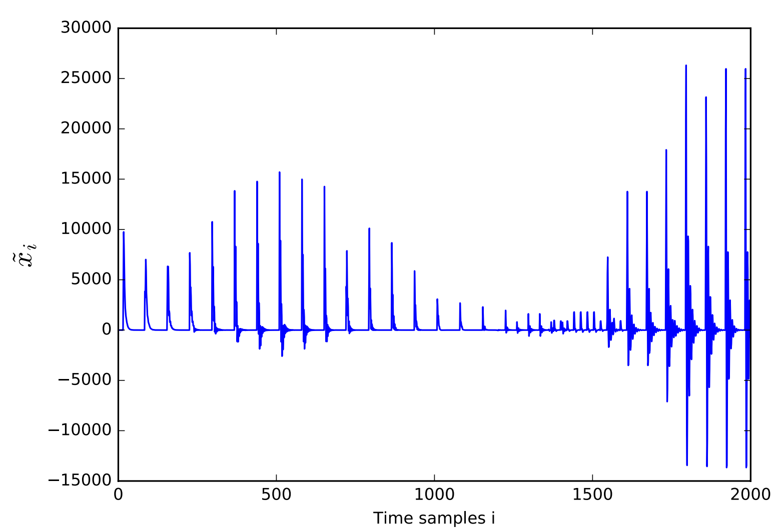
p = 5 vs. **n** = 160 $\Rightarrow$ **compression rate** = 32

Results

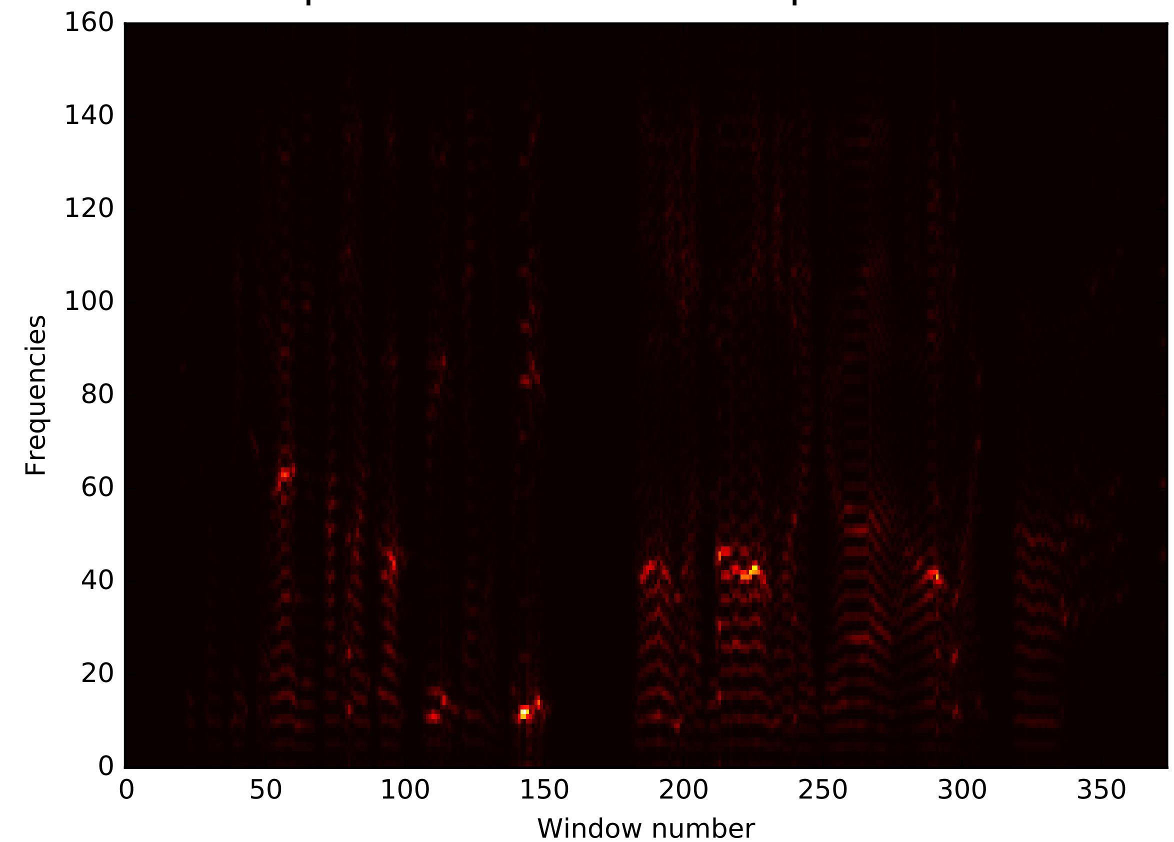
- Two simultaneous voices break the algorithm

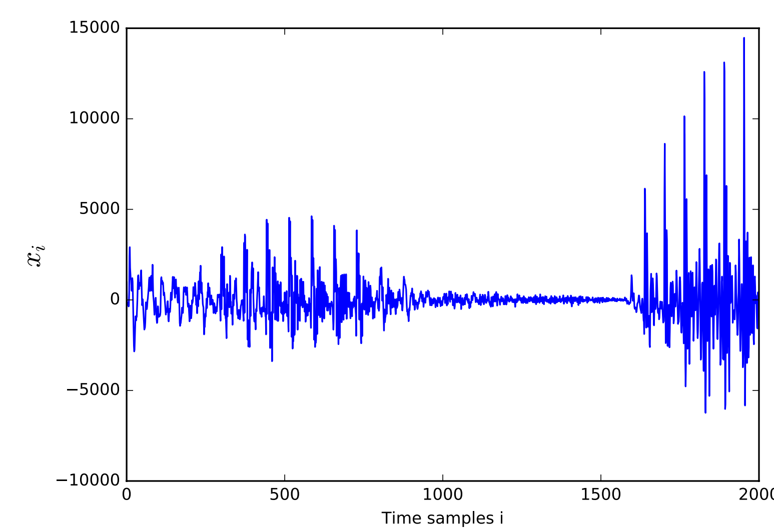
Original





Compressed / decompressed

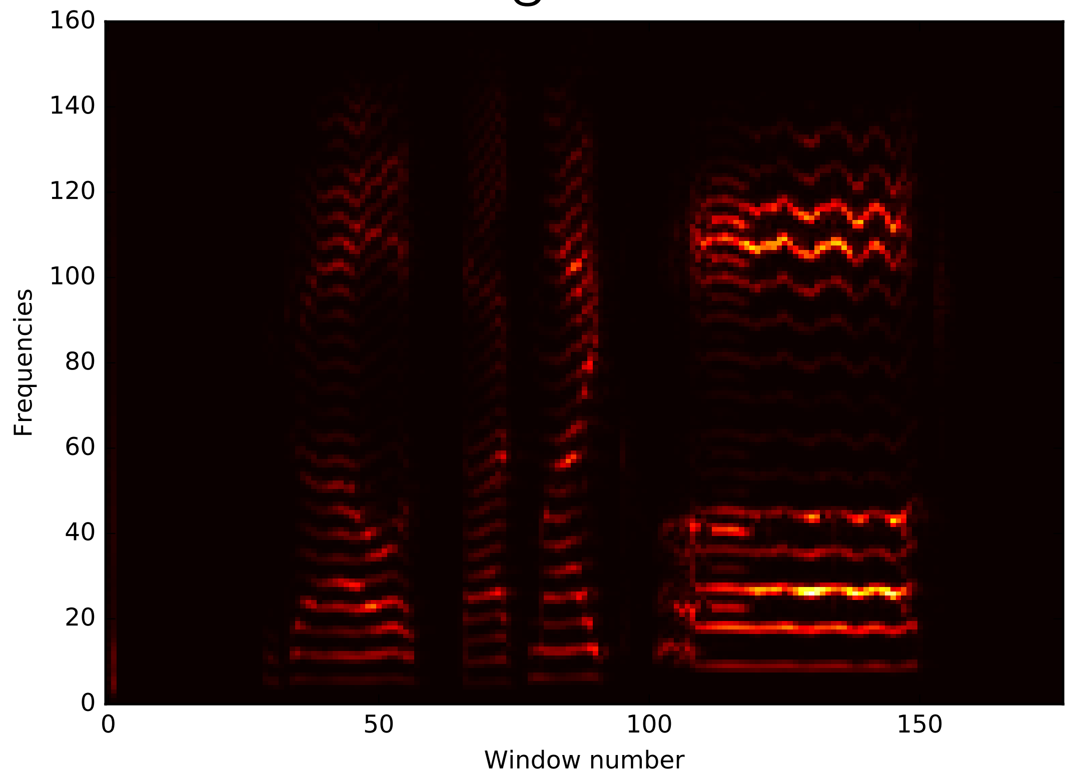




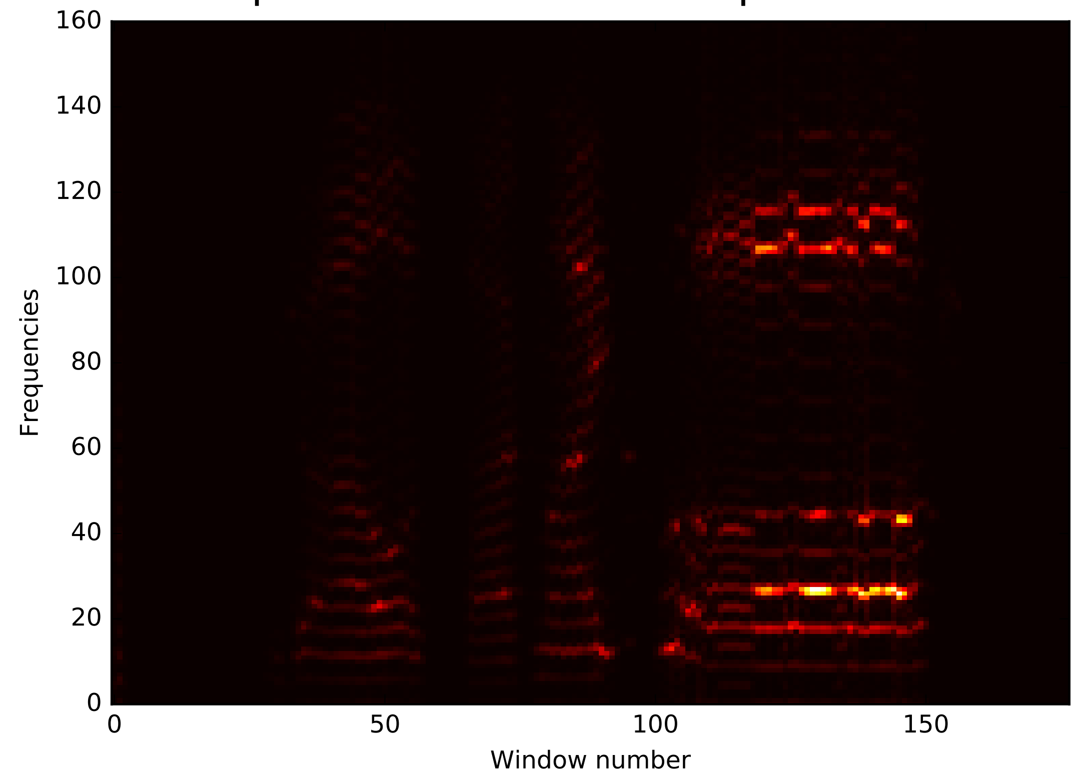
Results

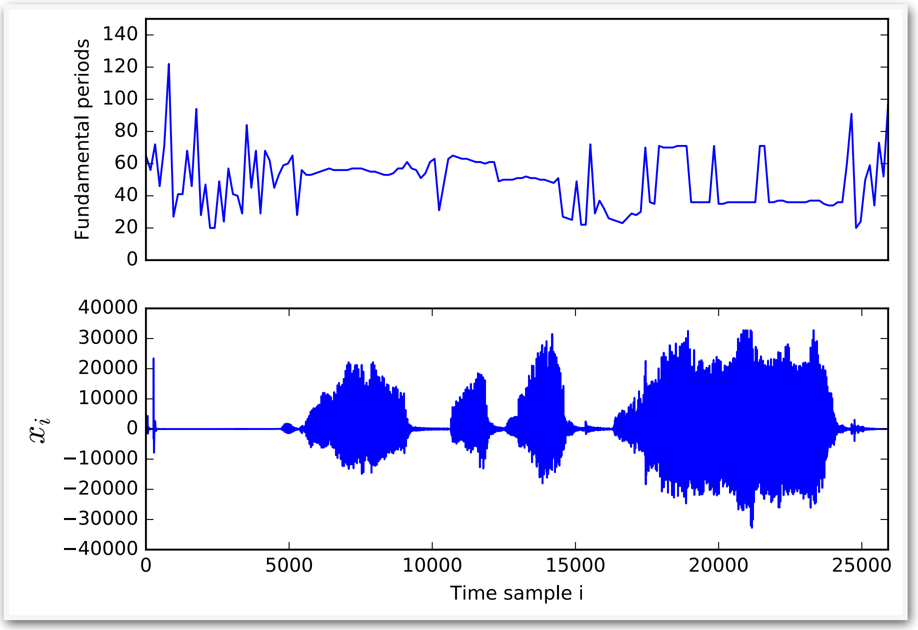
- Singing is hard to catch

Original



Compressed / decompressed





Thanks!