Rhythmic Demodulation for Zero-Resource Speech Recognition

Pascal Clark

24 July, 2012

also with

Les Atlas, Ivars Kirsteins, Greg Sell

Human Language Technology Center of Excellence, JHU

Motivating Question

What is the data rate of speech?

- Phonetic?
- Syllabic?
- Word?

The syllable in speech recognition:

- Phonological stability → "minimal recognition unit" (Fujimura, 1975)
- Demonstrated importance in human perception (Greenberg, 1997)
- Possible timing cues for theta waves in the brain (Ghitza and Greenberg, 2009)

Main Idea

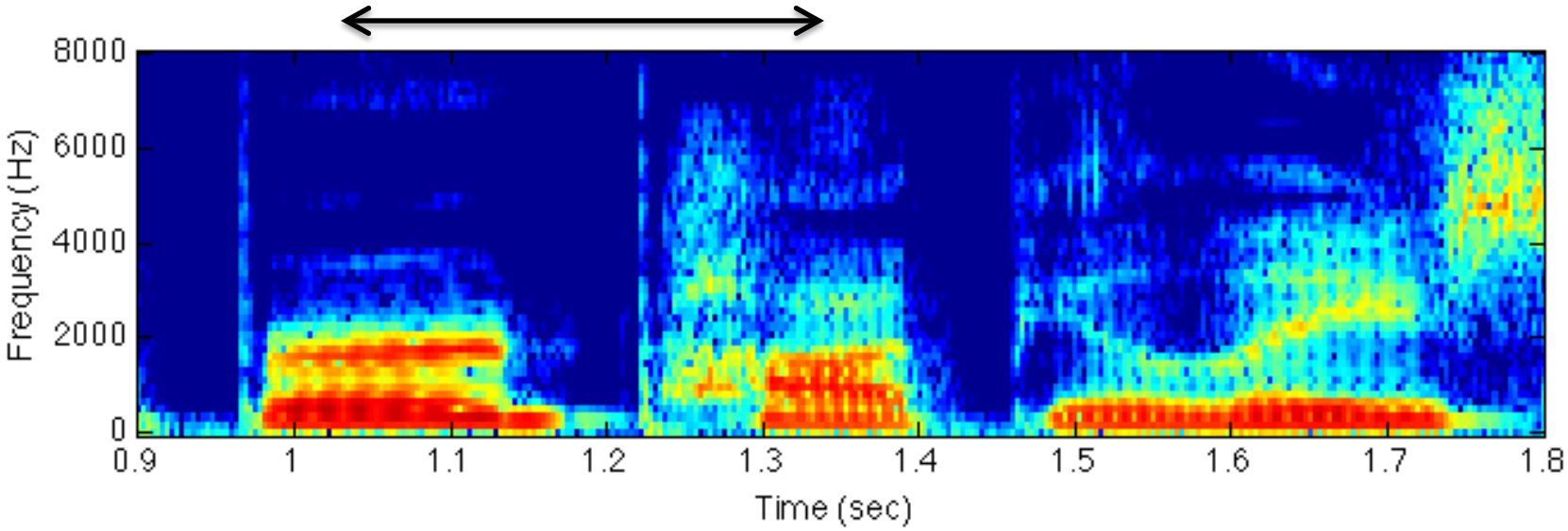
Sources of speaker-dependent temporal variations include:

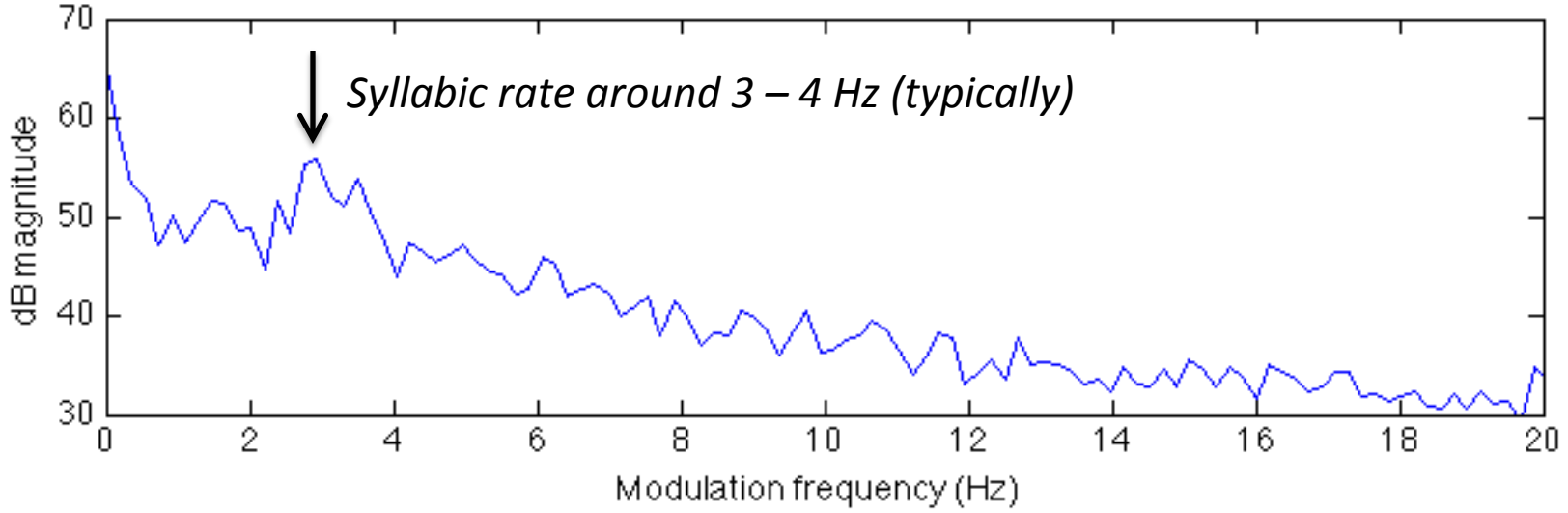
- Speaking rate mismatch
- Prosody, stress, pronunciation

Goal: An acoustic signal model for detecting and normalizing rhythmic variations between spoken terms.

Results: Two feature vector streams for Aren's Same/Different keyword detection evaluation.

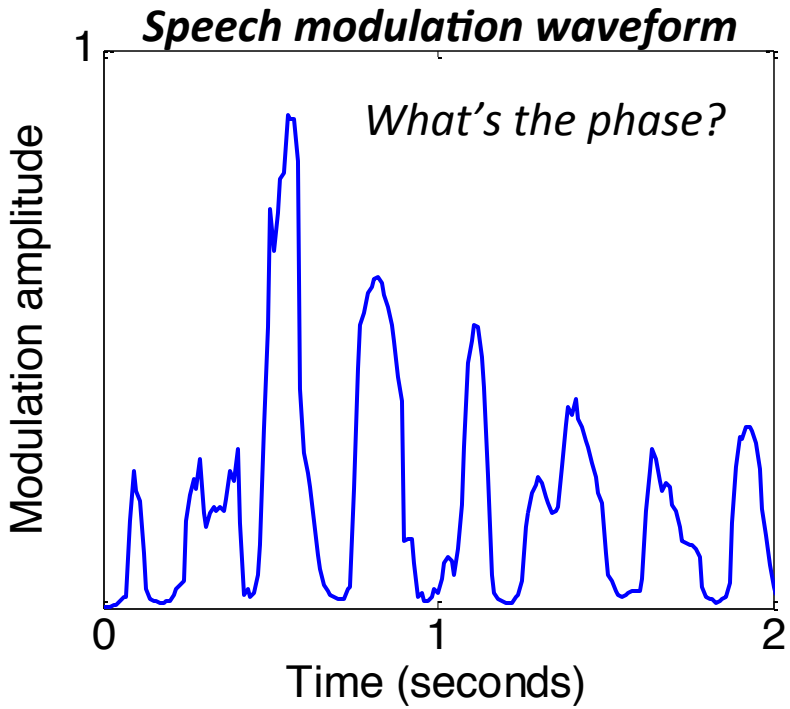
A Physical Descriptor of Speech Rhythm

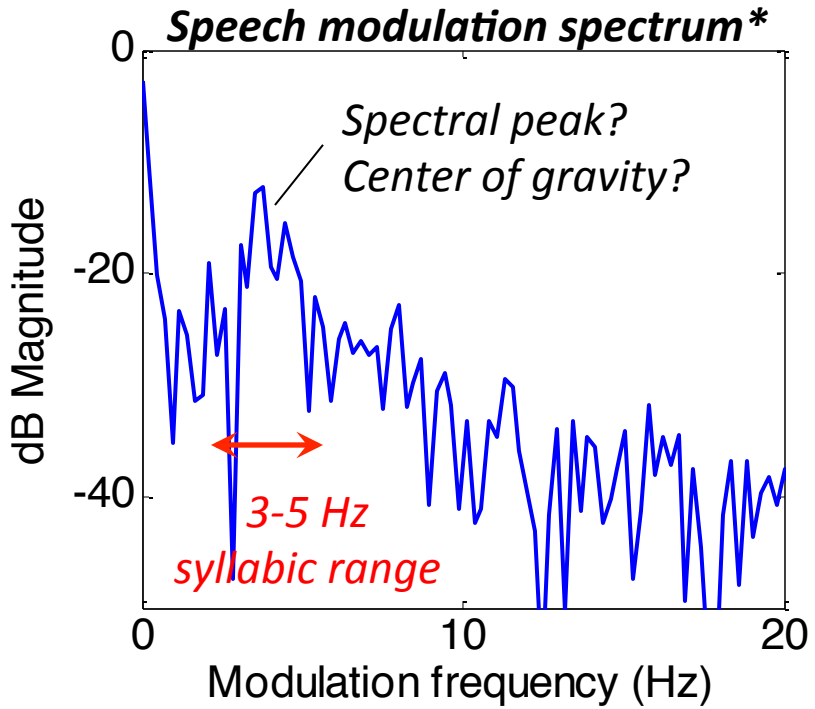




What's Wrong with Fourier?

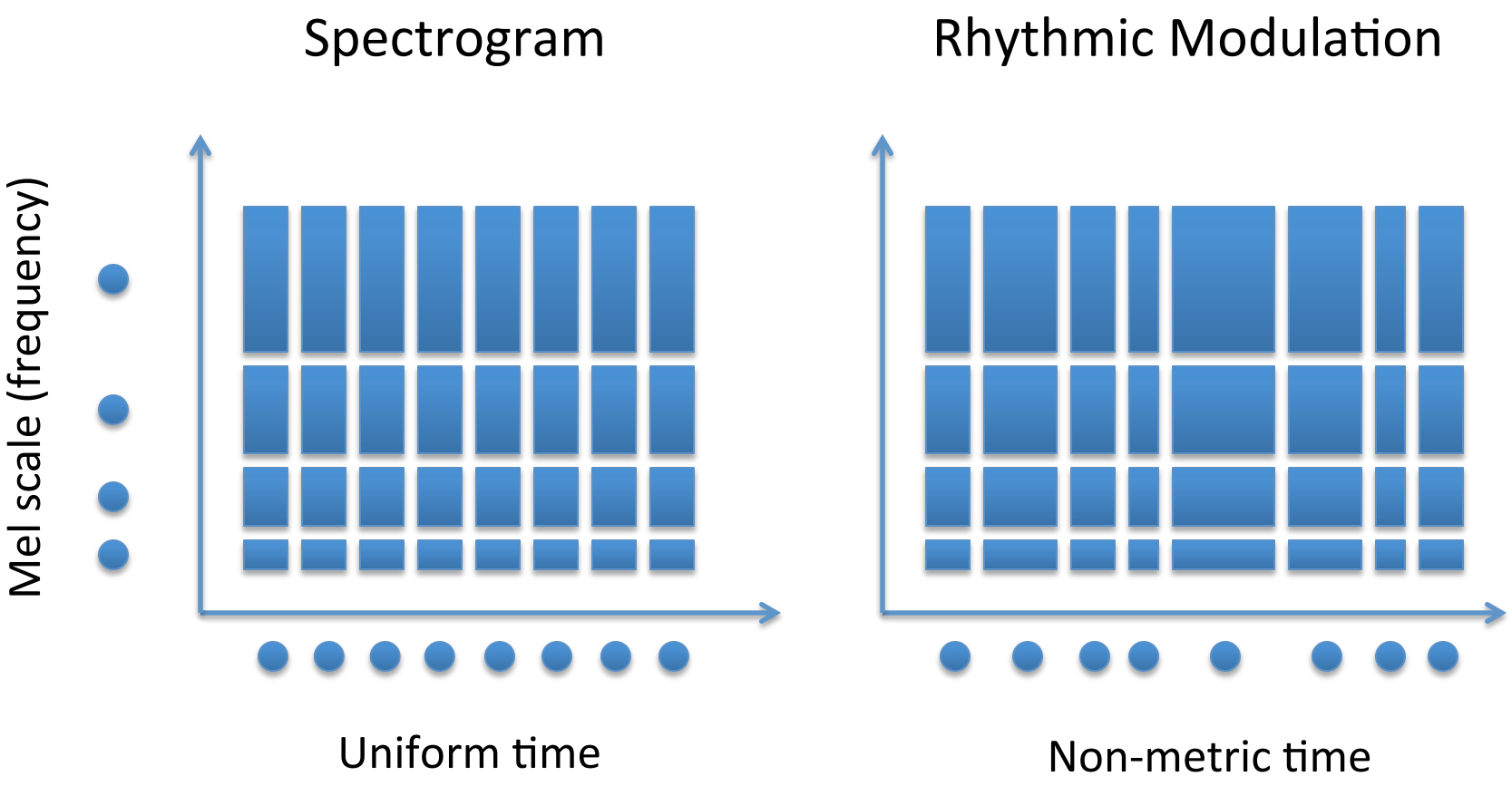
Speech is not periodic.





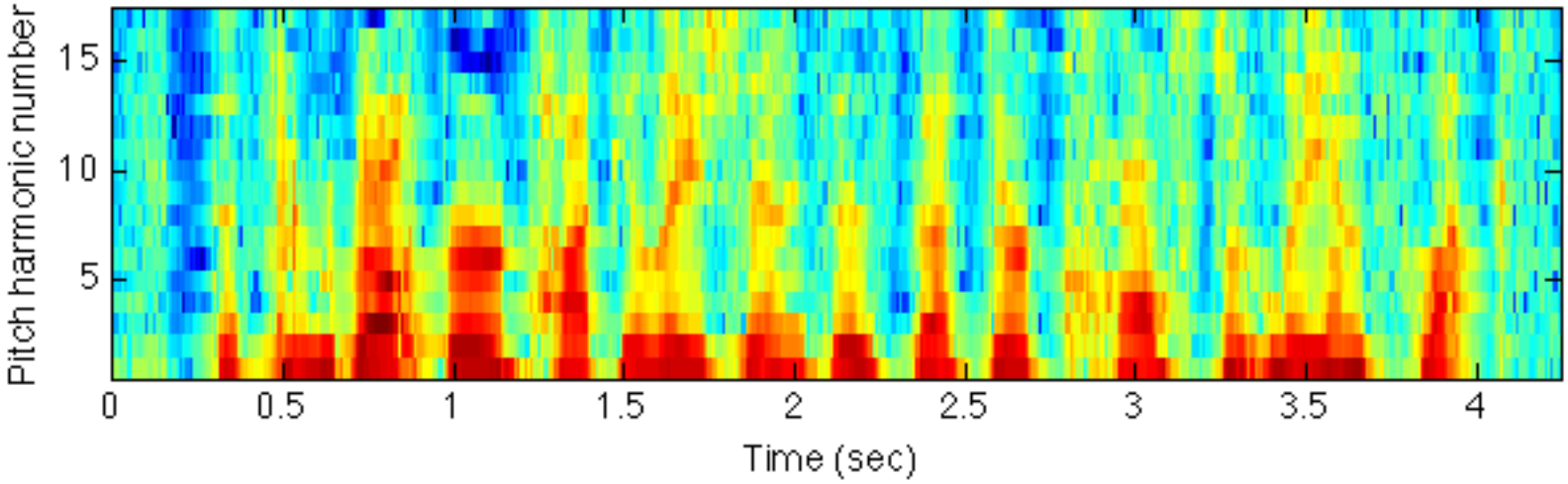
** Houtgast and Steeneken, 1985; Drullman, et al., 1994; Hermansky and Morgan, 1994*

Rhythm is Instead Event-Driven

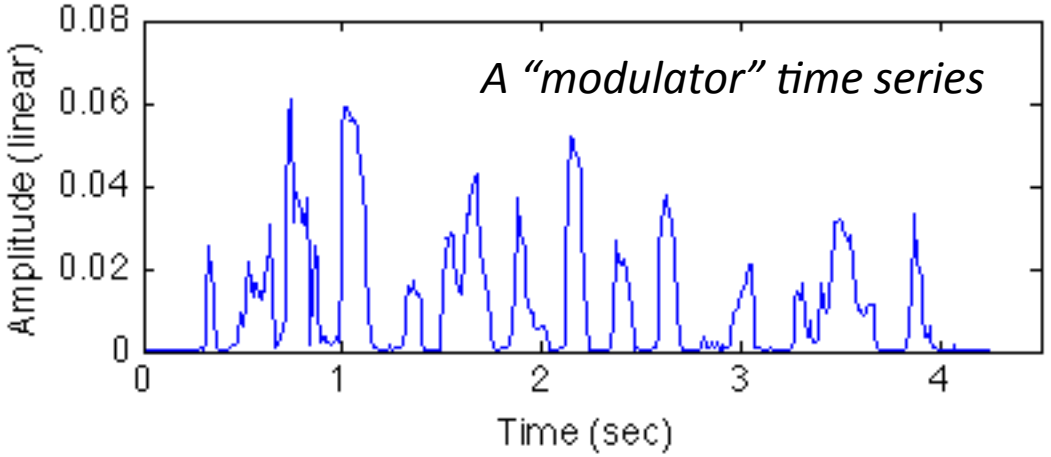


Rhythmic Demodulation (Part 1 of 3)

Smooth time-frequency representation (pitch-adaptive here):

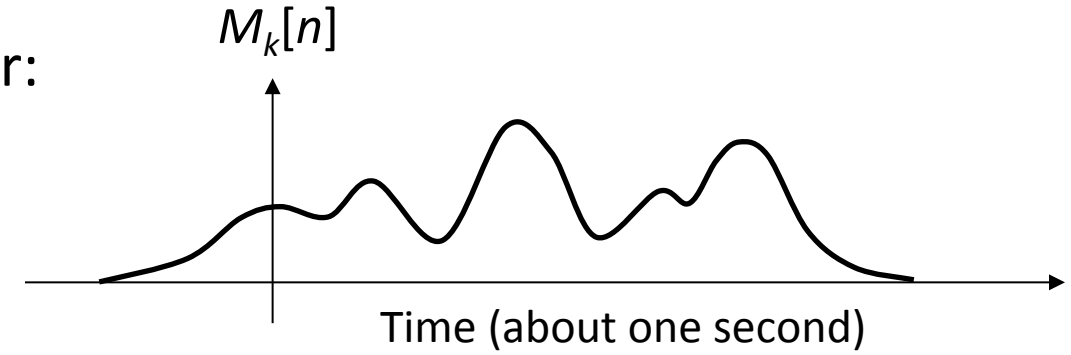


Take one row:

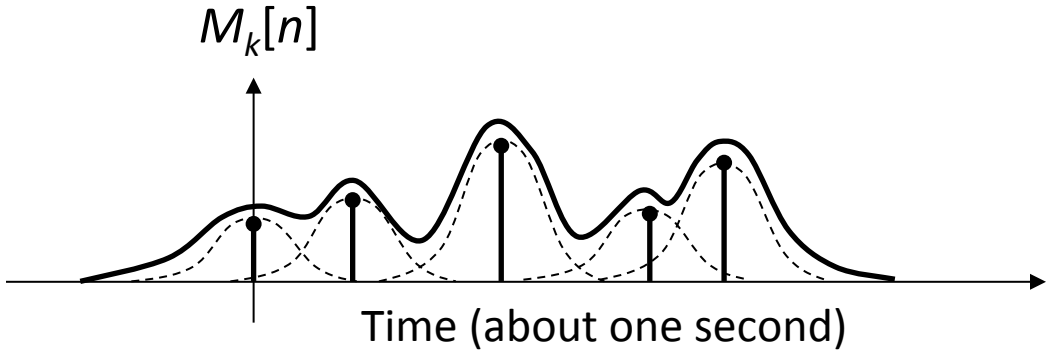


Rhythmic Demodulation (2 of 3)

Idealized modulator:

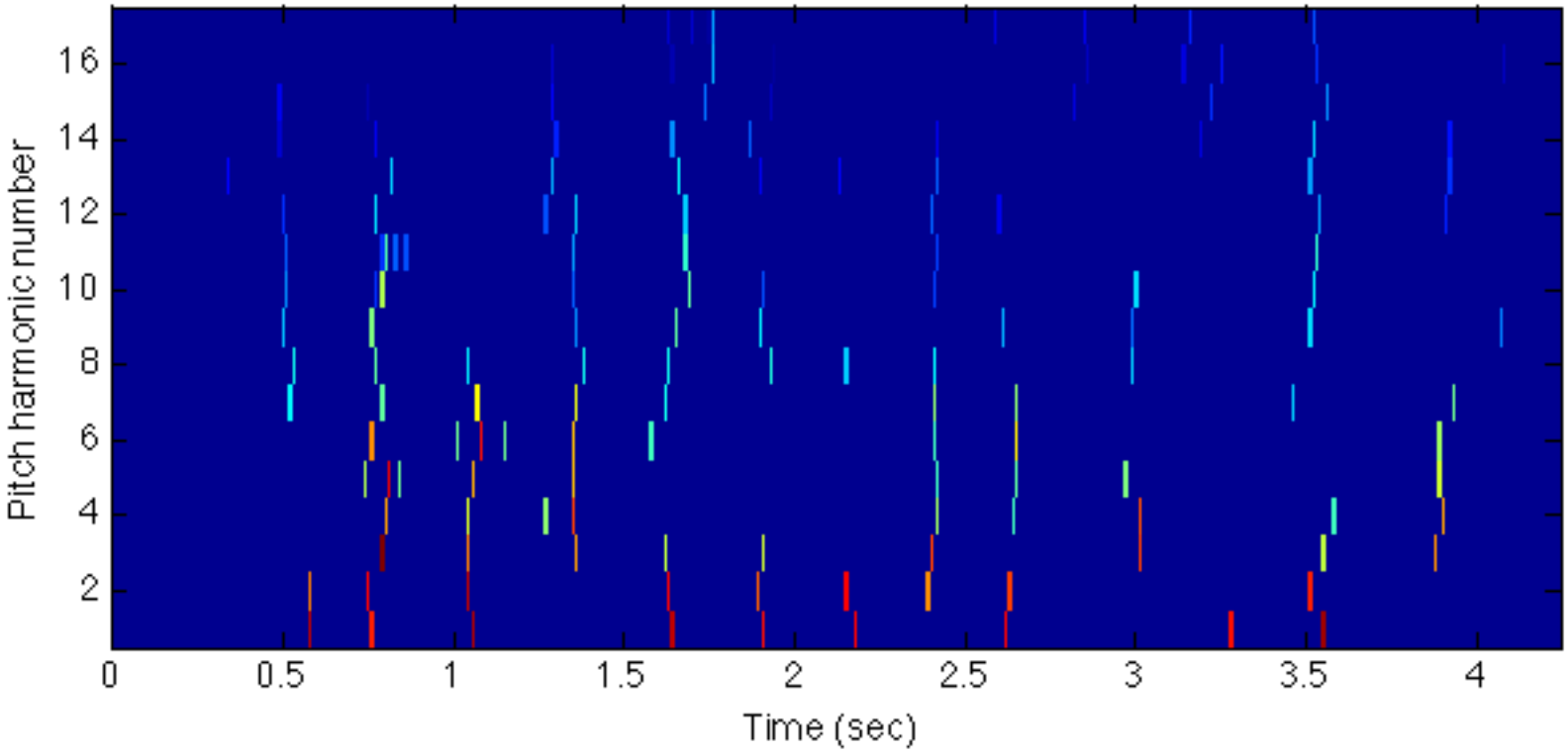


Rhythmic model with **sparse activations** and Gaussian signal basis (*"rhythmogram" approach, Lee and Todd, 2004*)



Rhythmic Demodulation (3 of 3)

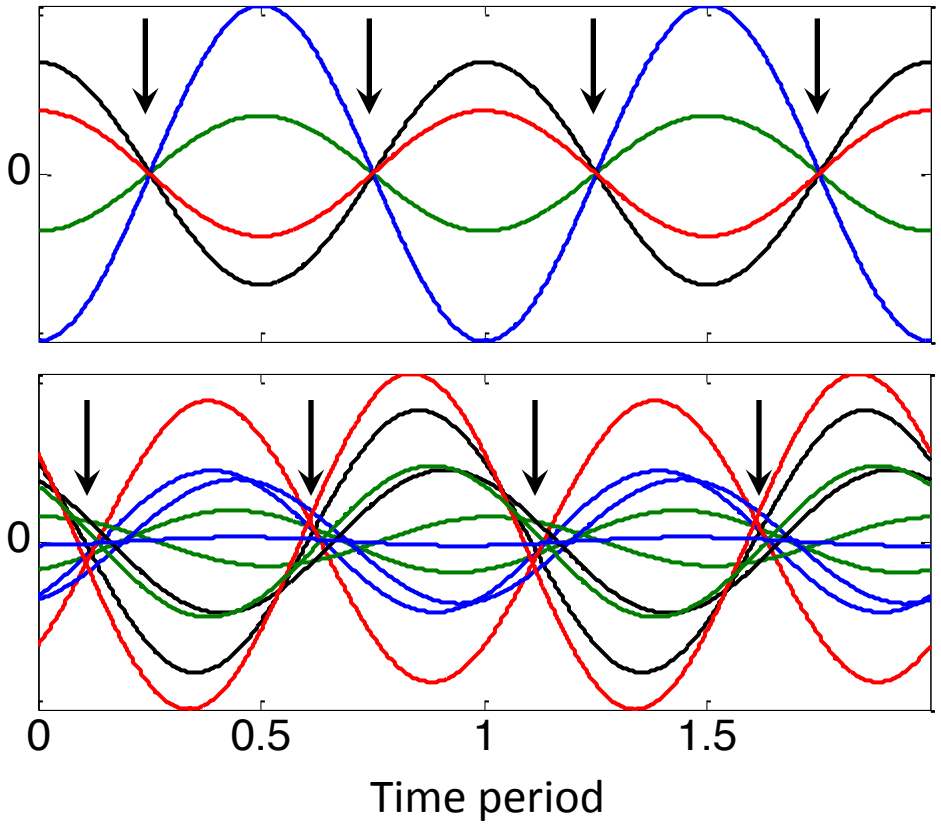
A systematic decoder based on matching pursuits (Mallat and Zhang, 1993), showing all activations after 20 iterations per row:



Principal Components of Rhythm

Rhythmic random process: $m[n] = \int a(\omega)\cos(\omega n) - b(\omega)\sin(\omega n)\, d\omega$

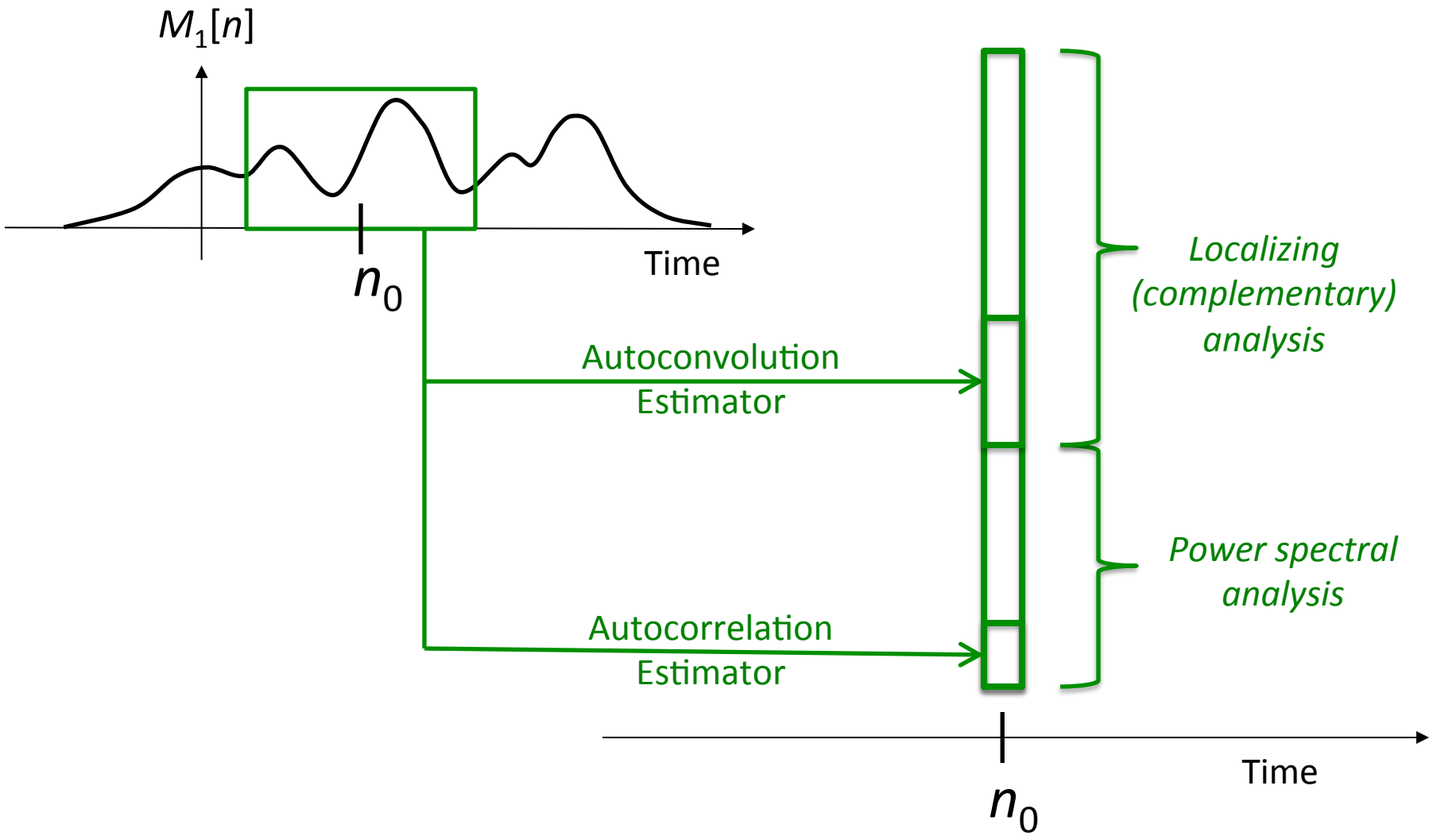
Gaussian random variables



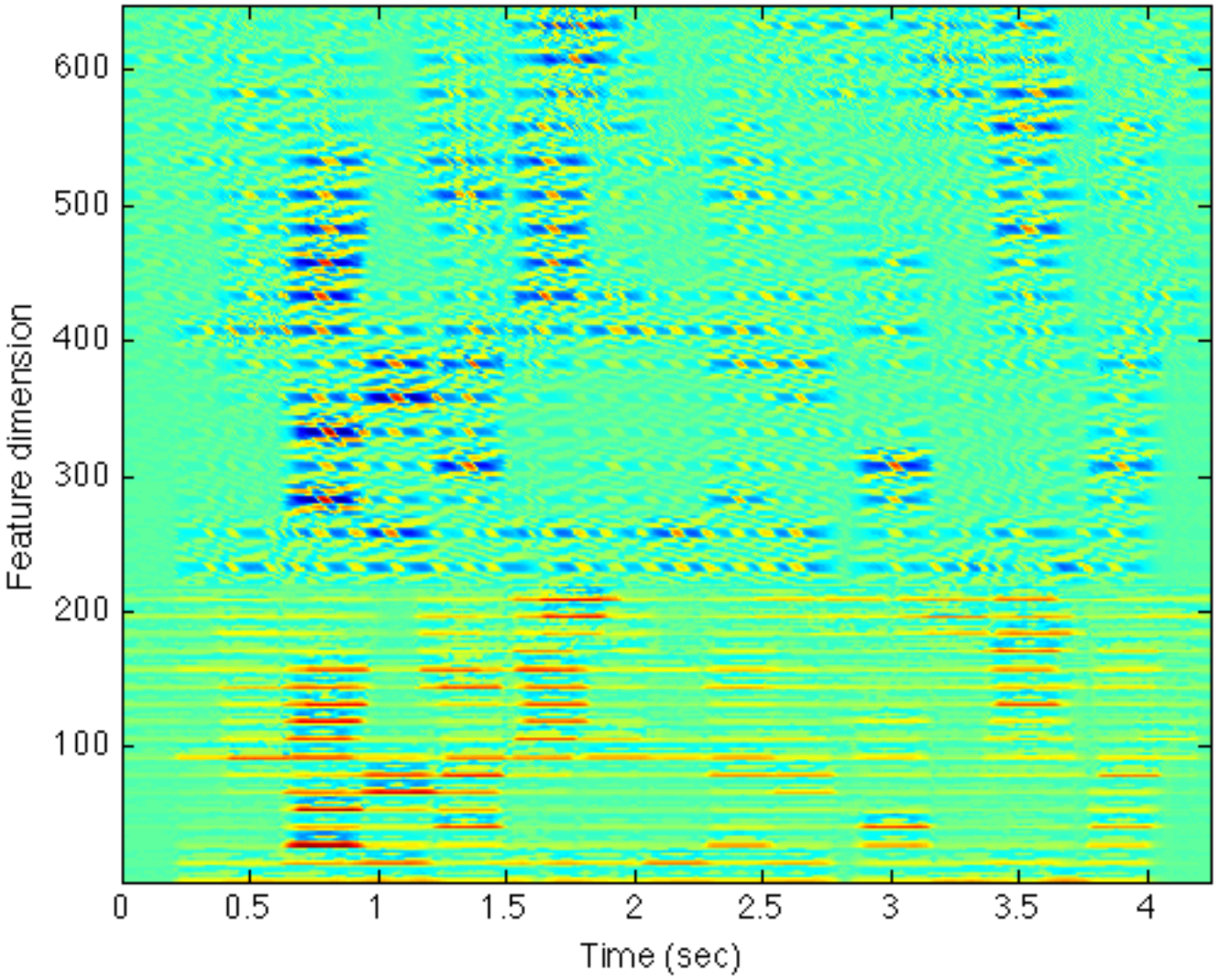
$$\text{var}\{a\} = \text{var}\{b\}$$
$$E\{ab\} \neq 0$$

$$\text{var}\{a\} \neq \text{var}\{b\}$$
$$E\{ab\} \neq 0$$

Super-Vector Formation



Speech Super-Vector Example



Conclusion

Operational definition of syllabic rhythm:

- Sparse activations
- Non-uniform timing
- Non-periodic due to local variation

Localized deconvolution reveals underlying pattern of syllabic activations.

Possible speaker invariance?