Speech processing: how to wreck a nice peach

Richard Harvey
IT Livery Company Professor of Information Technology, Gresham College

Professor of Computer Science, School of Computing Sciences,
University of East Anglia

#richardwharvey

Speech processing: how to wreck a nice beach

Richard Harvey
IT Livery Company Professor of Information Technology, Gresham College

Professor of Computer Science, School of Computing Sciences,
University of East Anglia

#richardwharvey

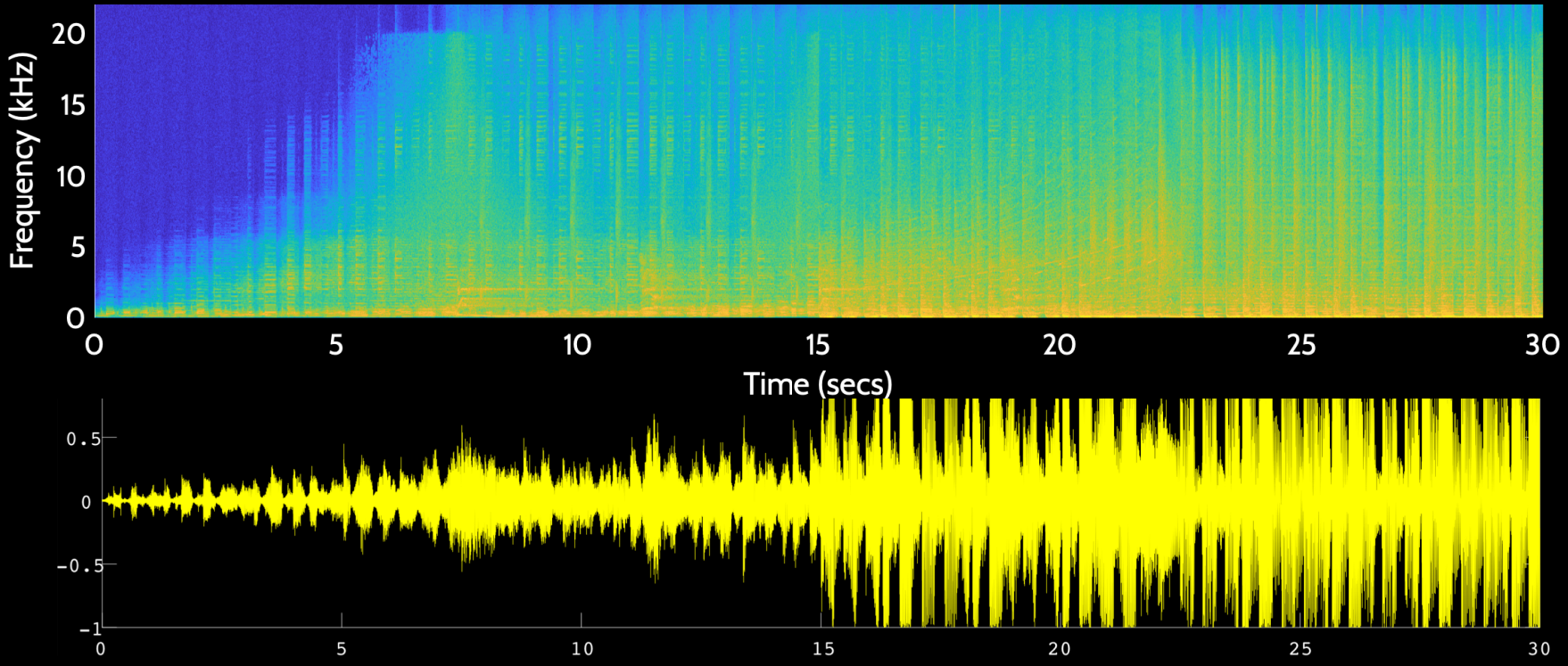
Speech processing: how to recognise speech

Richard Harvey
IT Livery Company Professor of Information Technology, Gresham College

Professor of Computer Science, School of Computing Sciences,
University of East Anglia

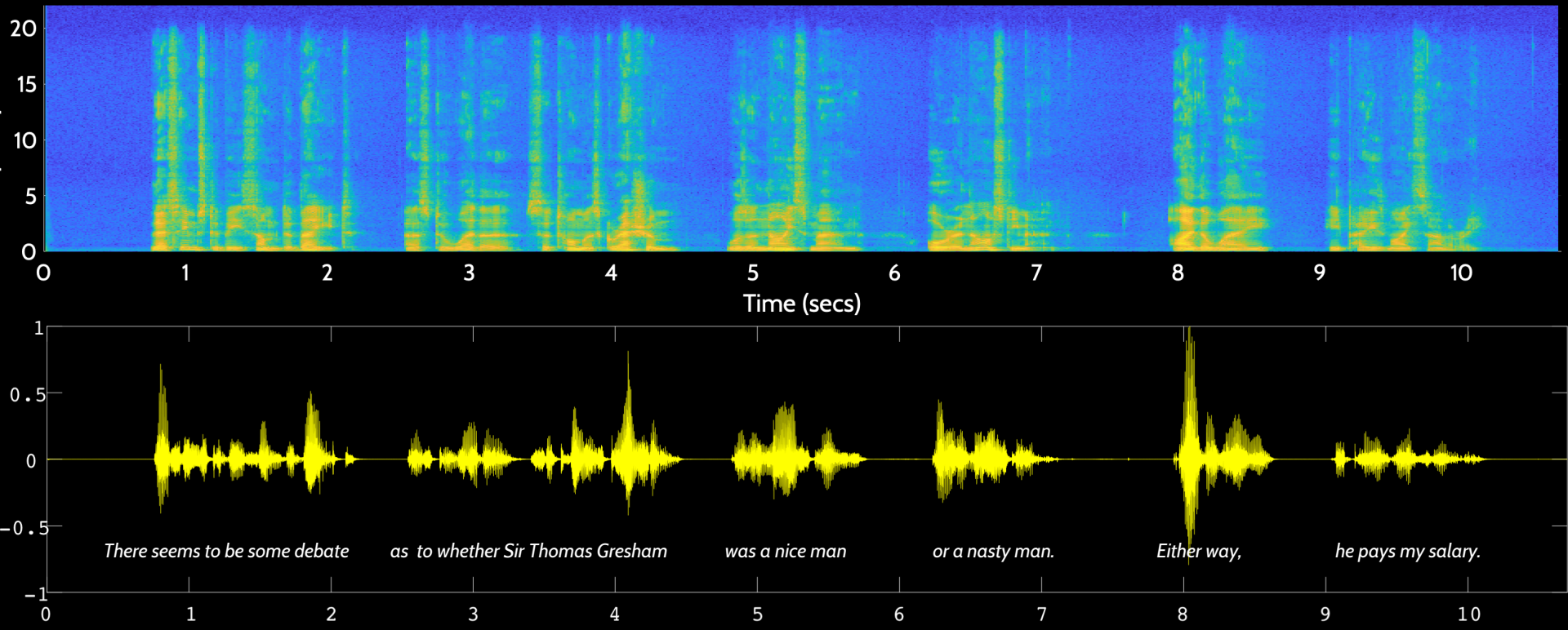
#richardwharvey

What are acoustic signals?



"We love it" Paul Keane, taketones.com

What are acoustic signals?



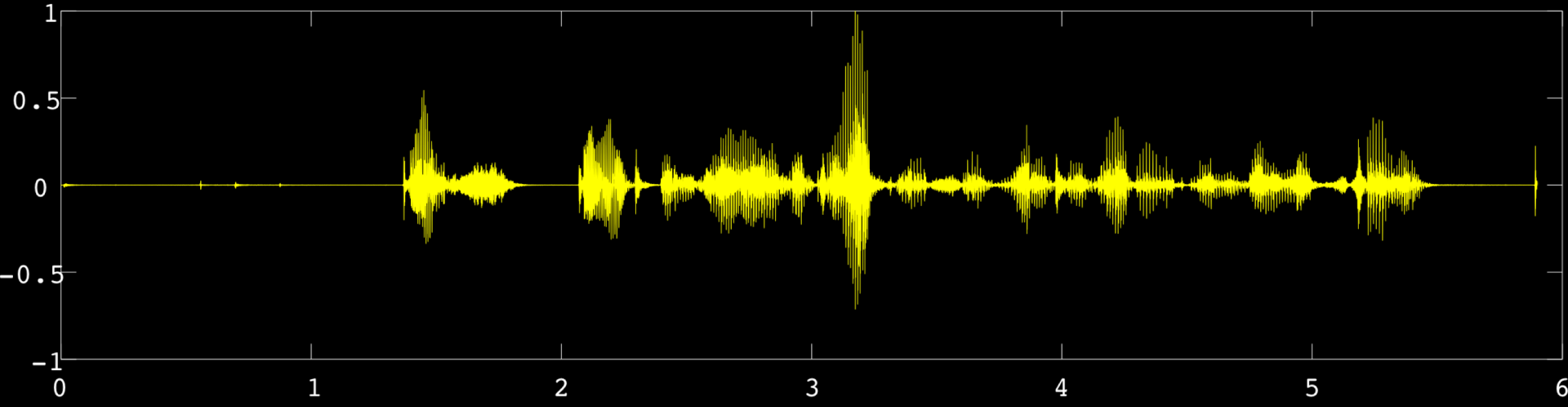
Going digital...

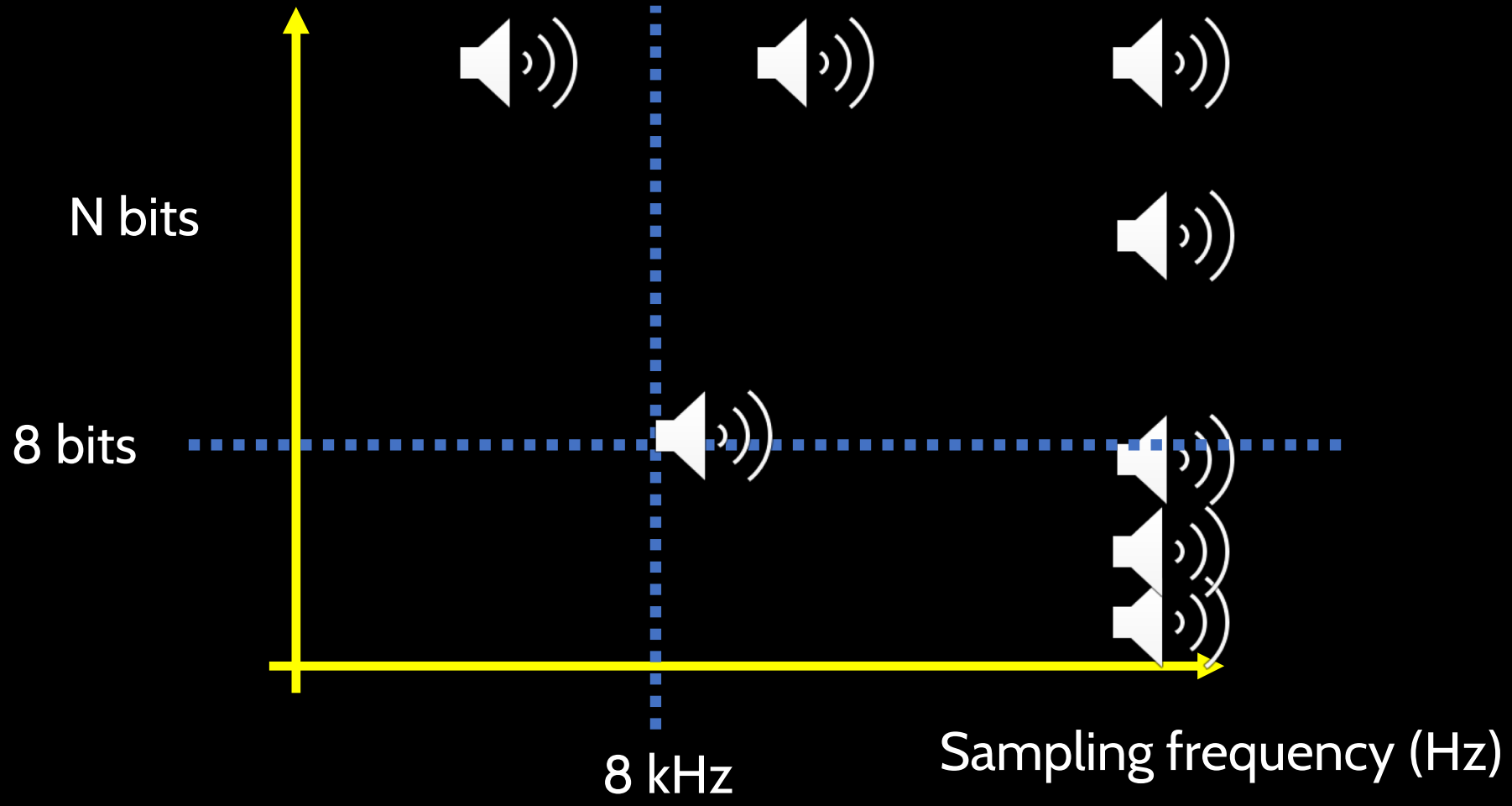
Smoothly varying acoustic signals are

sampled

quantised

Sampling and quantisation




N bits
8 bits
8 kHz
Sampling frequency (Hz)

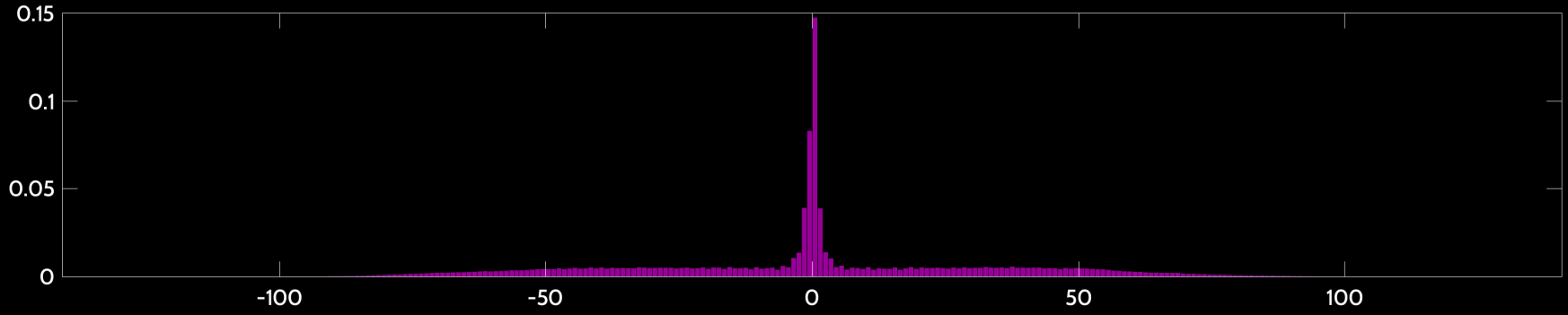
The Nyquist sampling theorem

If a signal has a bandwidth of B Hz

then it can be completed reconstructed by sampling at least $2B$ Hz.

Companding

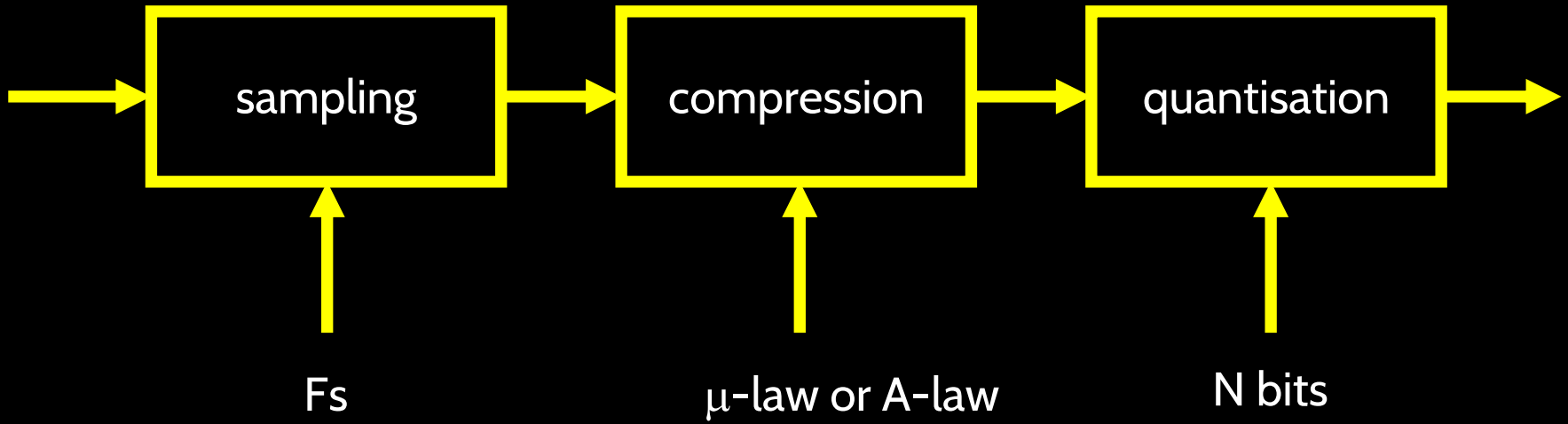
8 bits = 2^8 levels = 256 levels (±128) levels



Companding

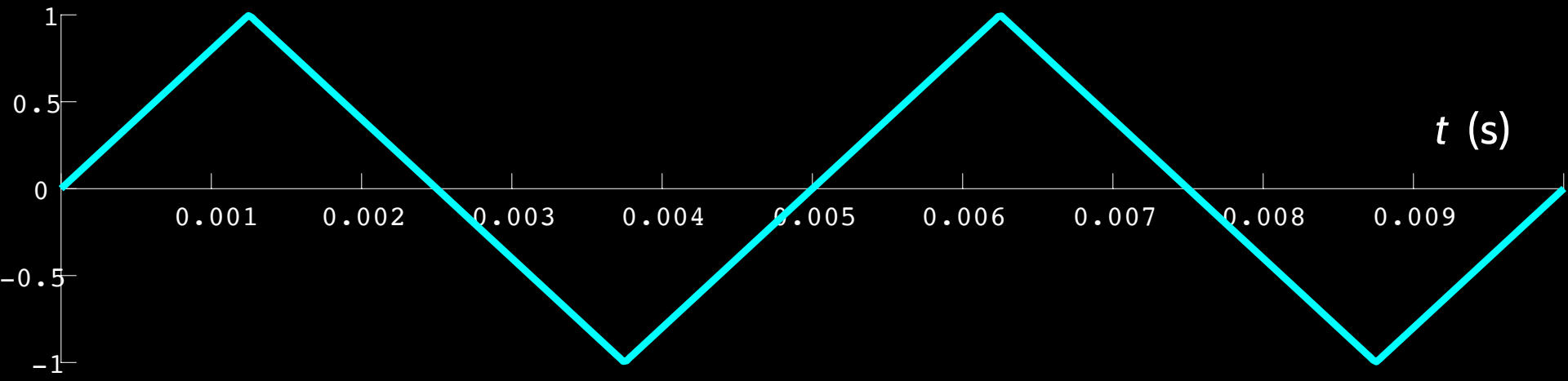
Also frequently used as an artistic effect...

So ...in your device...

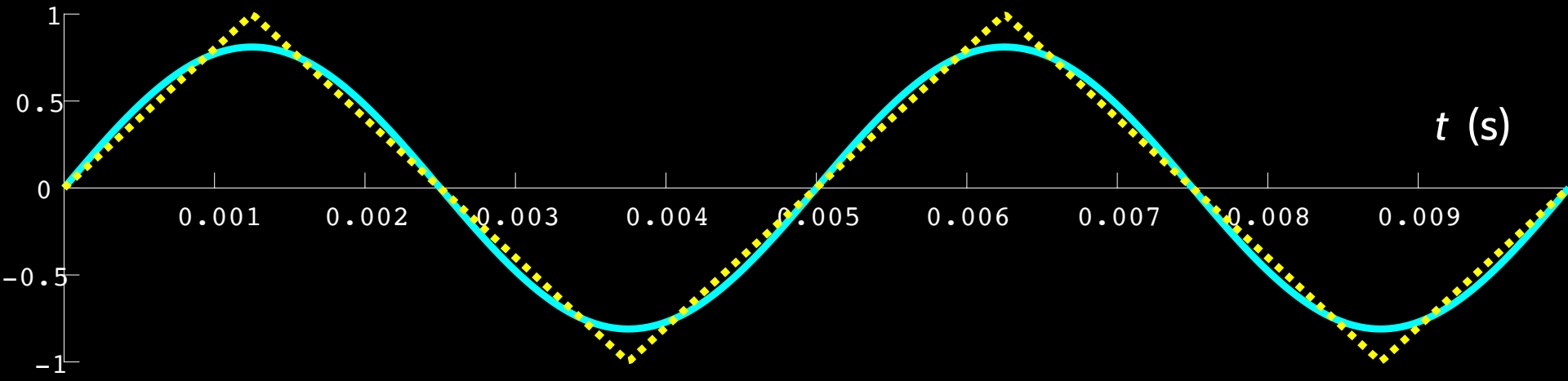


Fourier representations

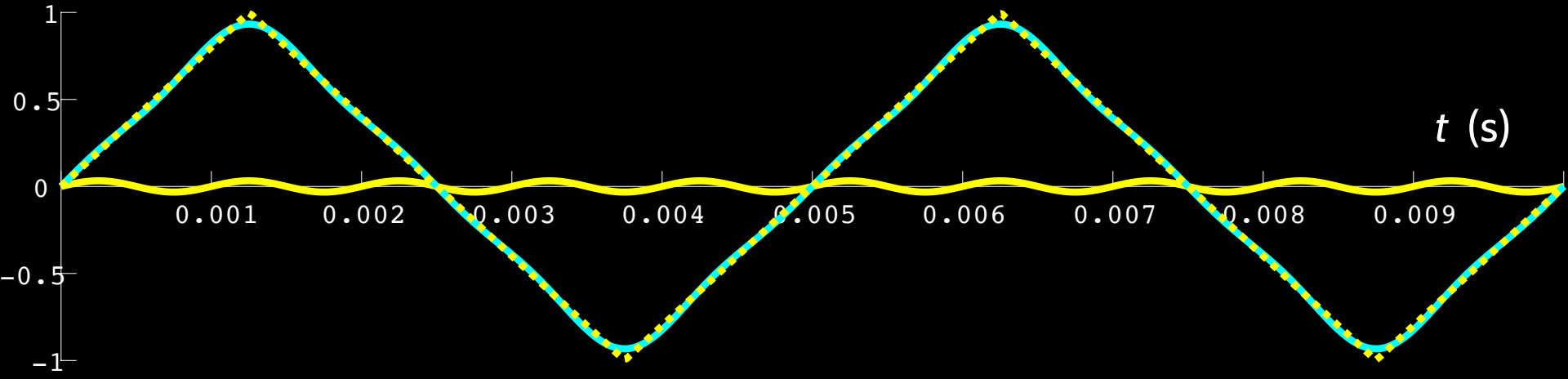
f = 200Hz



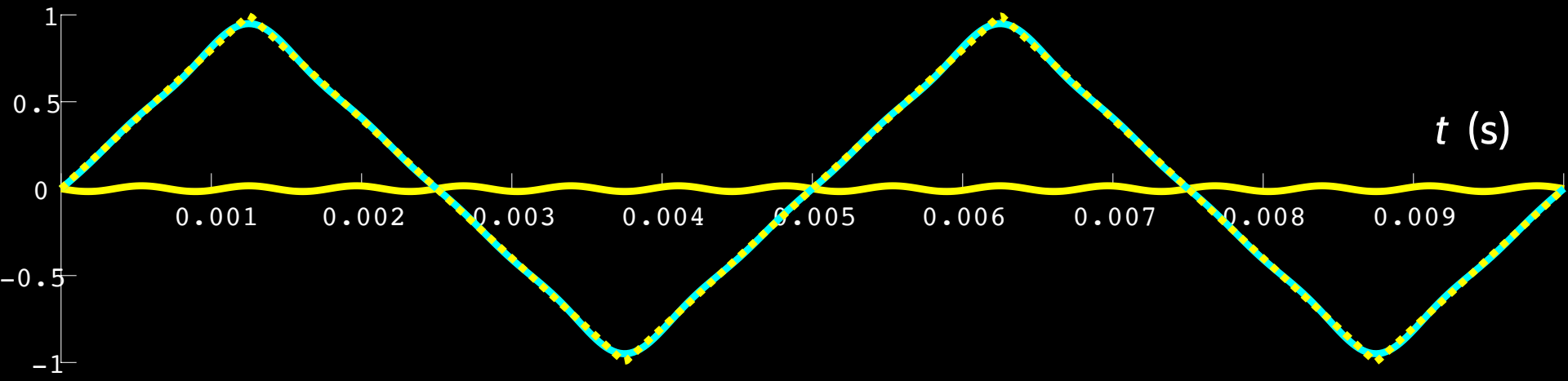
f = 200Hz



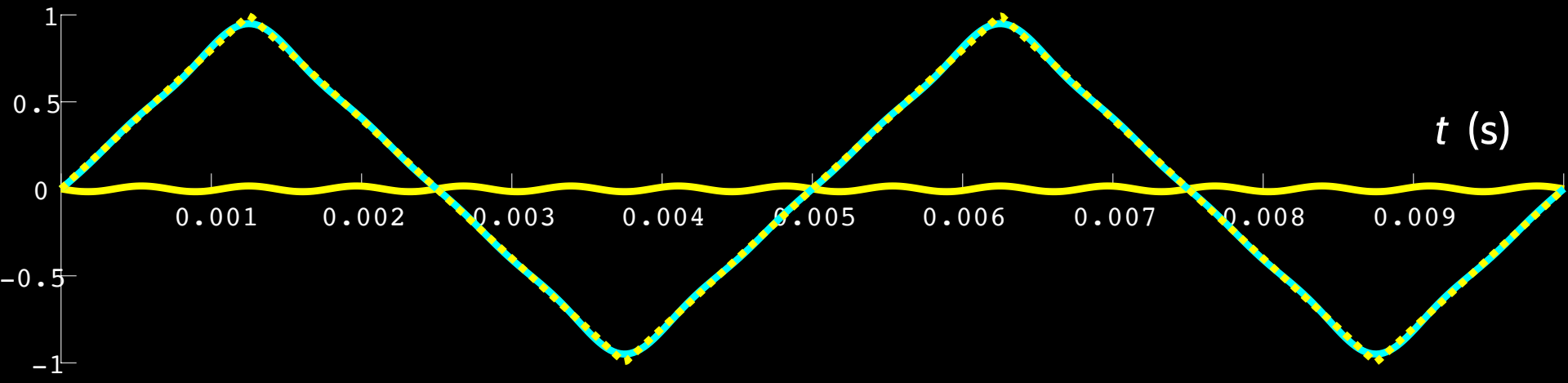
$f = 3 \times 200\mathrm{Hz}$

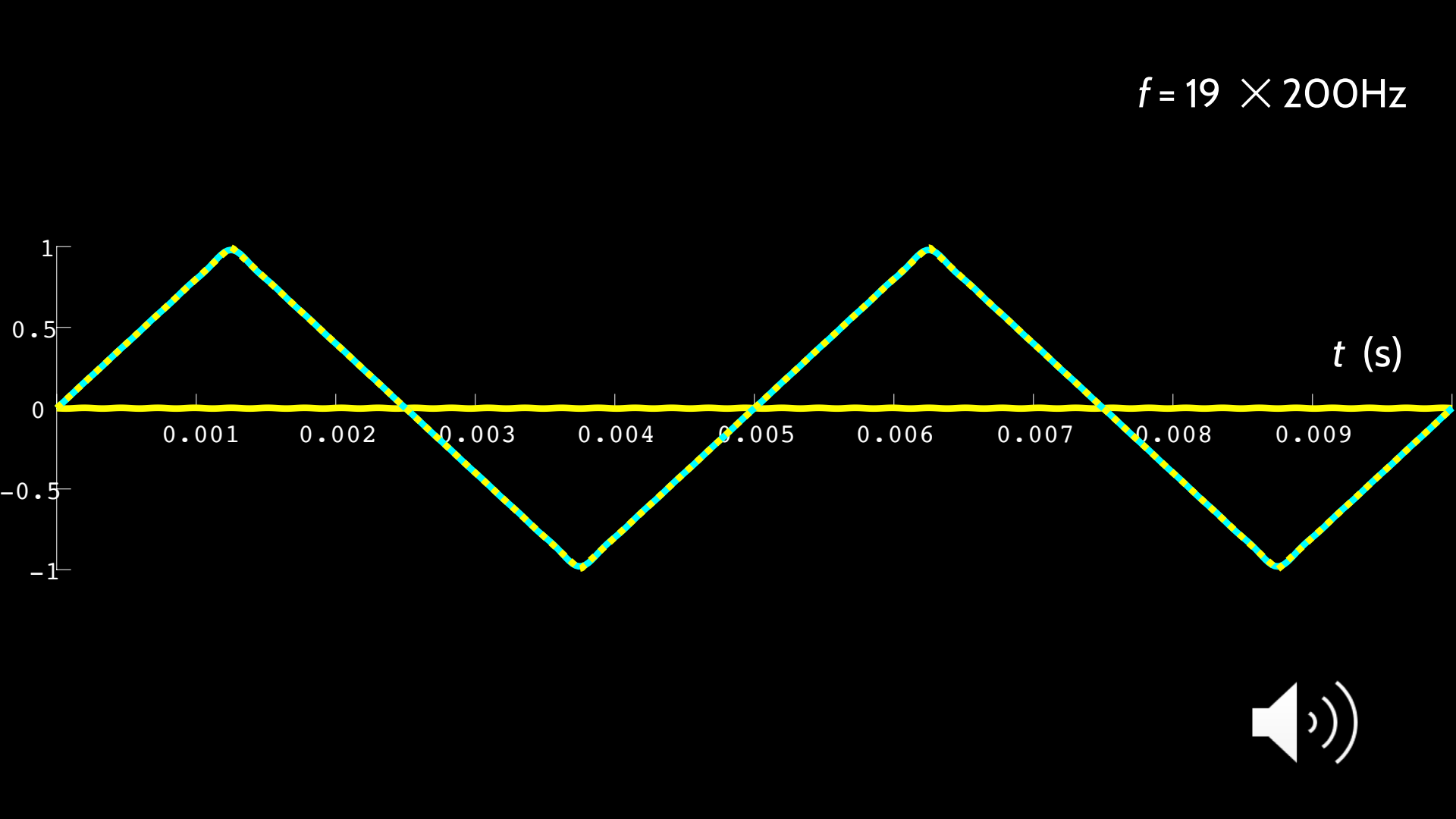


$f = 5 \times 200\mathrm{Hz}$

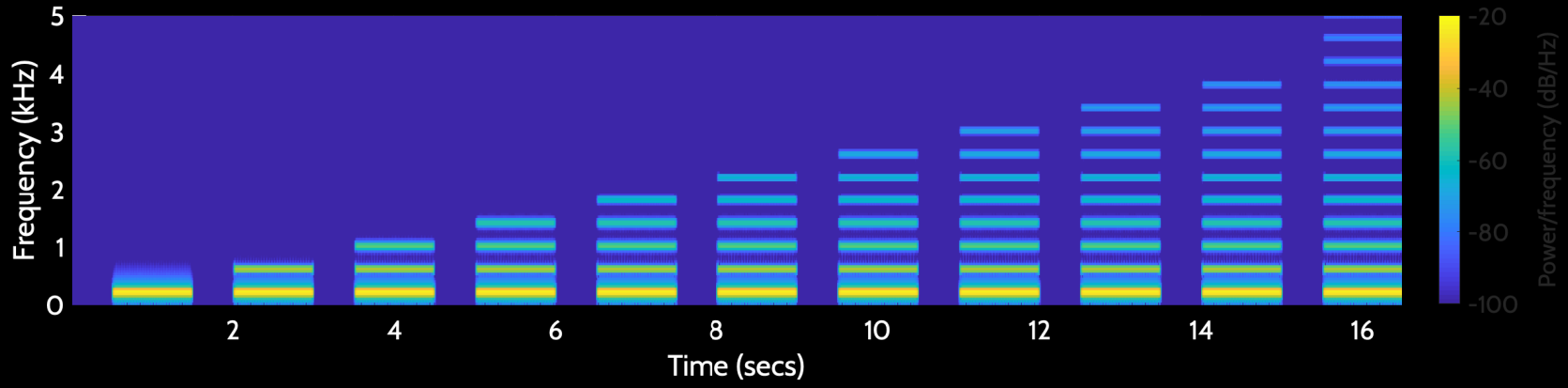


$f = 7 \times 200\mathrm{Hz}$




f = 19 × 200Hz
1
0.5
0
−0.5
−1
0.001
0.002
0.003
0.004
0.005
0.006
0.007
0.008
0.009
t (s)

Fourier synthesis and analysis

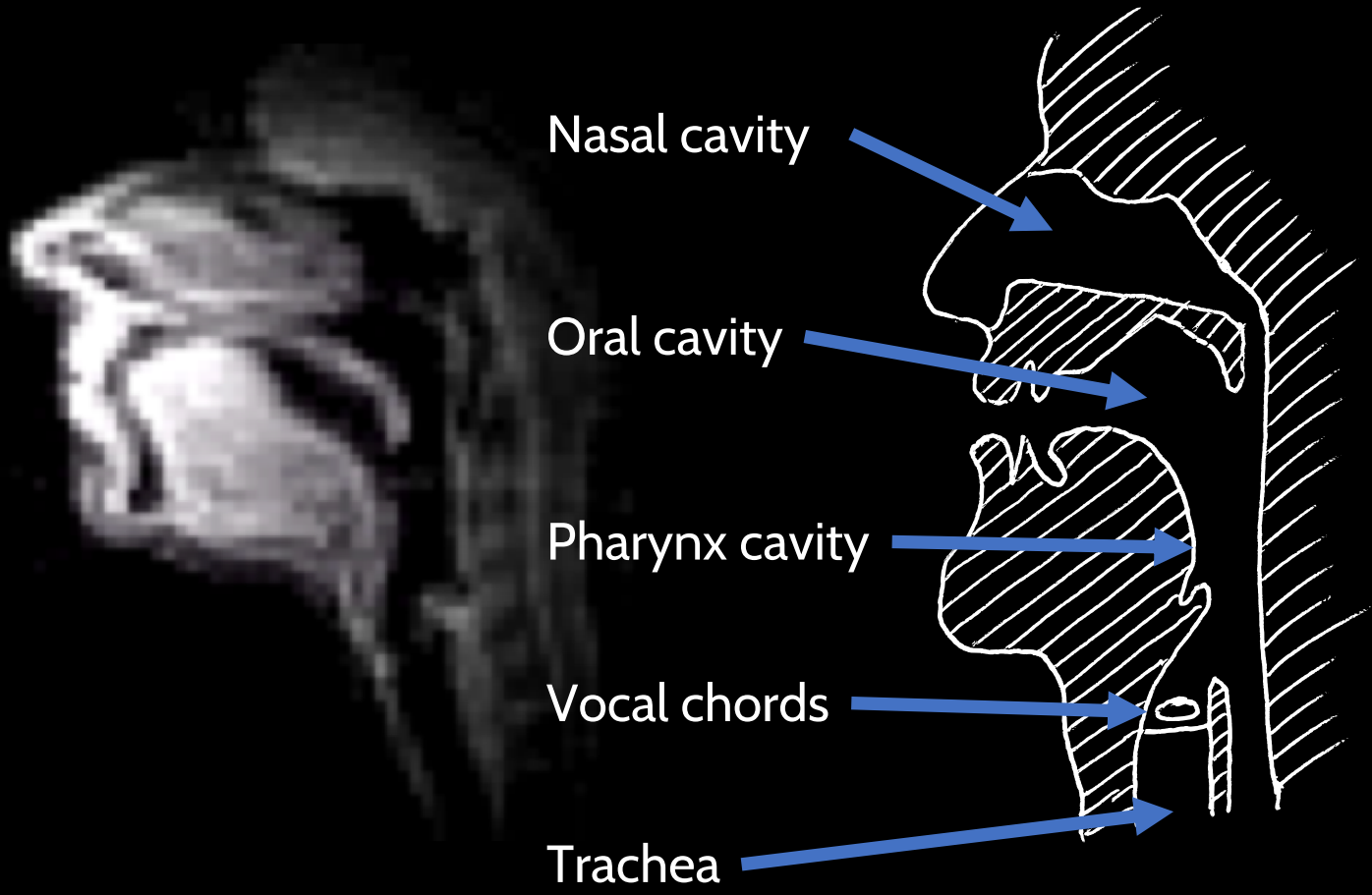


Fourier synthesis and analysis

Time domain

Frequency domain

$x(t)$  $X(f)$



Shrikanth Narayanan, Asterios Toutios, Vikram Ramanarayanan, Adam Lammert, Jangwon Kim, Sungbok Lee, Krishna Nayak, Yoon-Chul Kim, Yinghua Zhu, Louis Goldstein, Dani Byrd, Erik Bresch, Prasanta Ghosh, Athanasios Katsamanis, Michael Proctor, "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC)", The Journal of the Acoustical Society of America, vol. 136, no. 3, pp. 1307-1311

Mental model of speech

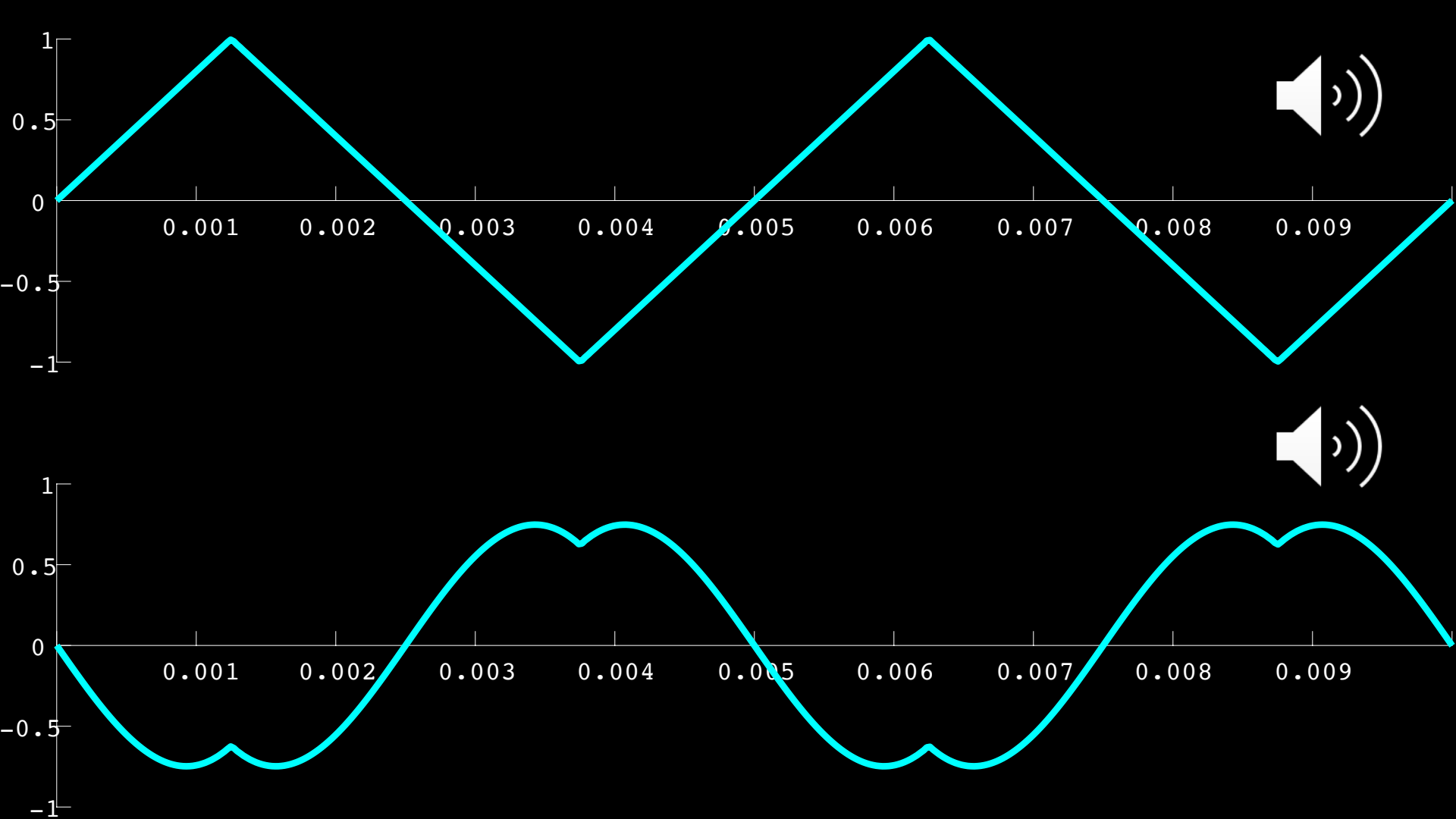
$S(f)$

Modulation, $H(f)$

$S(f) = G\,E(f)\,H(f)$

Excitation, $E(f)$

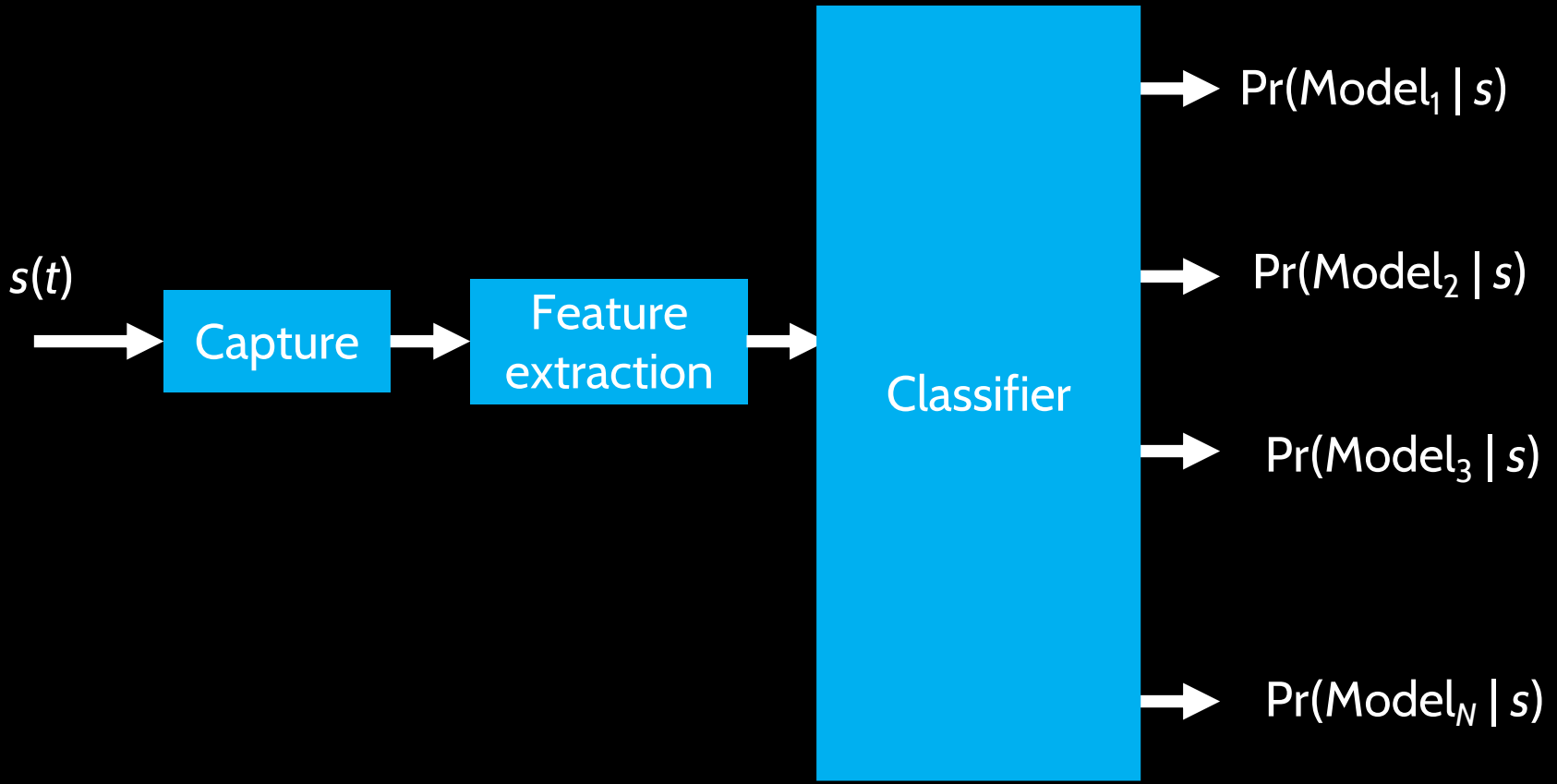
Energy, G


1
0.5
0
-0.5
-1
0.001
0.002
0.003
0.004
0.005
0.006
0.007
0.008
0.009
1
0.5
0
-0.5
-1
0.001
0.002
0.003
0.004
0.005
0.006
0.007
0.008
0.009

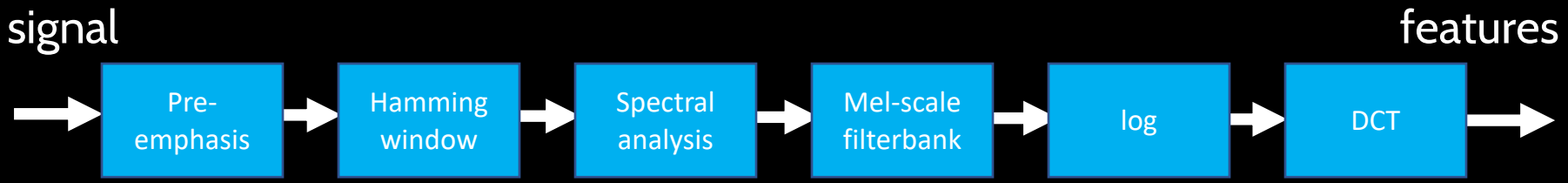
Fred Jelinek

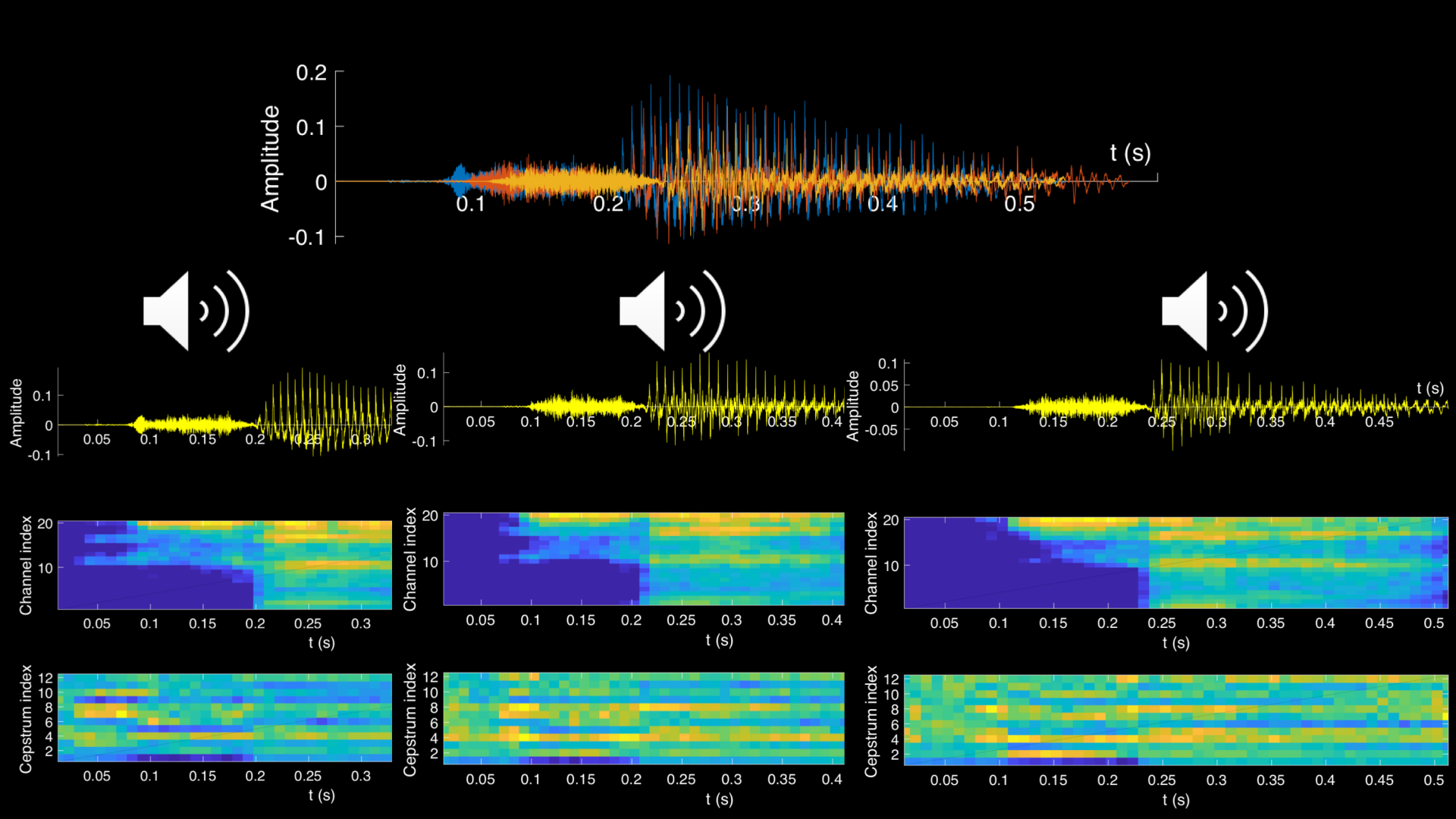
“Every time I fire a linguist, the performance of my speech recogniser improves”

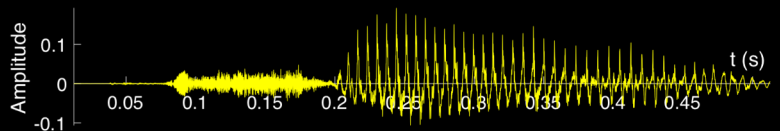
General classification architecture

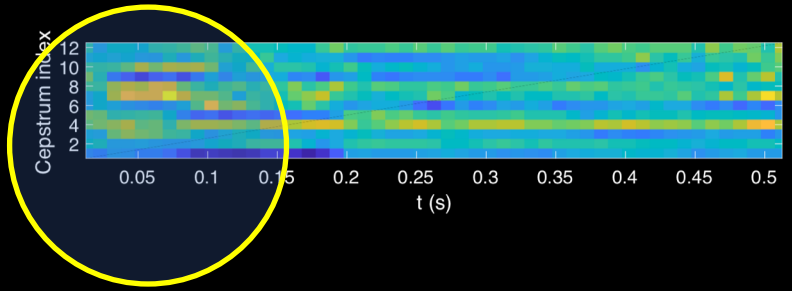


Feature extraction: MFCCs

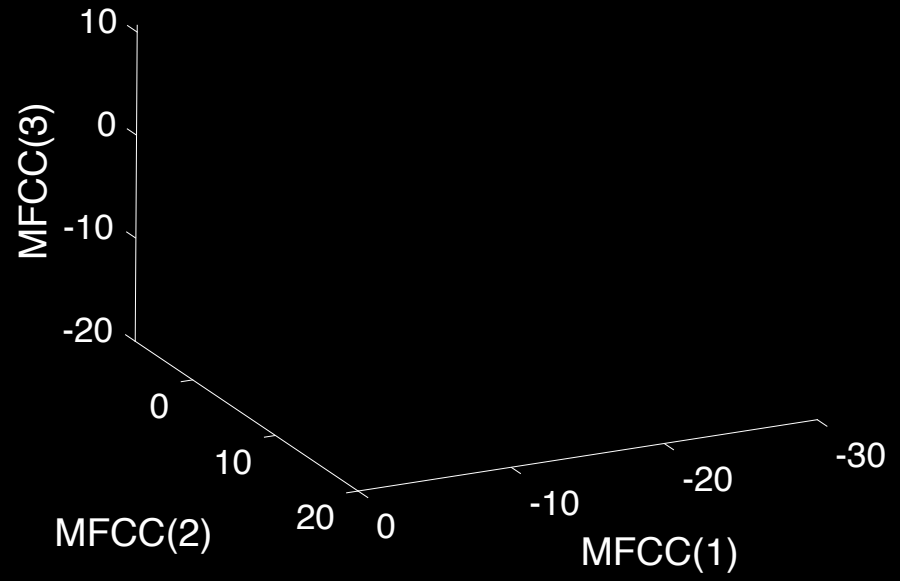



Amplitude
t (s)
Channel index
Cepstrum index





-8.3741
-0.2704
-3.4736
...



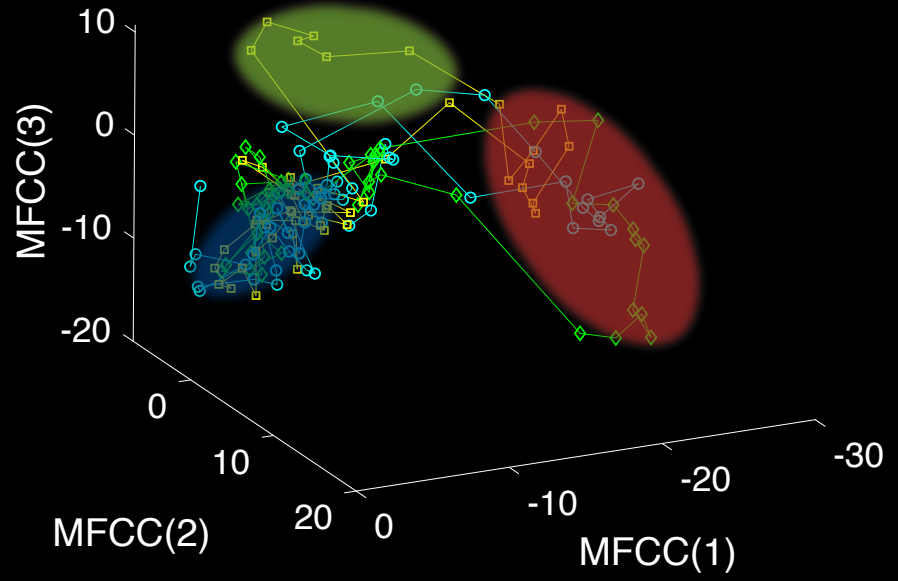
Temporal modelling

Path 1: BBBGRRR

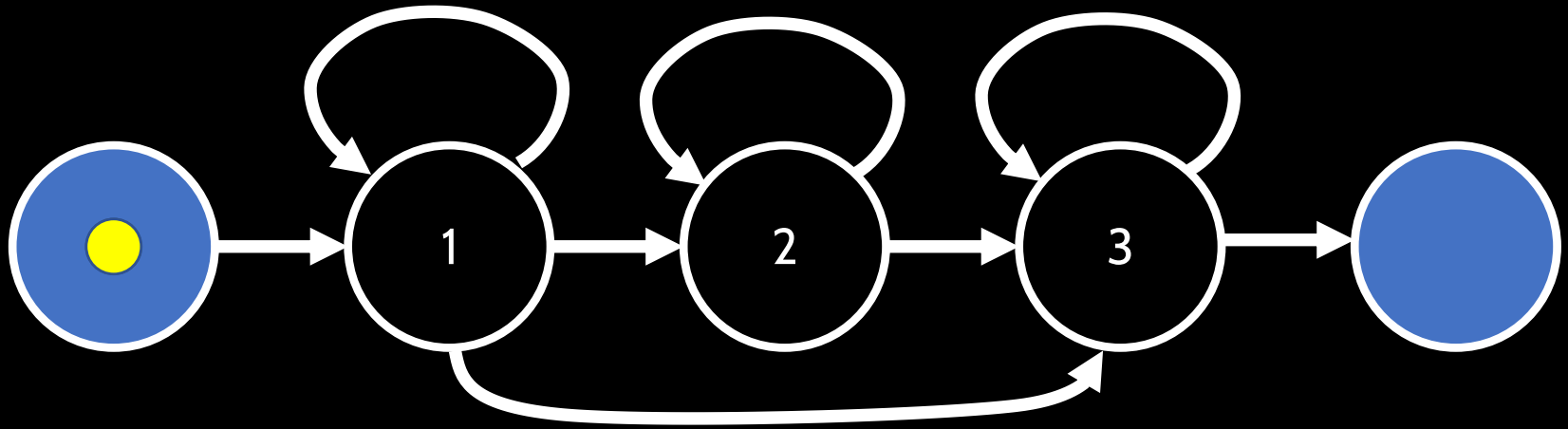
Path 2: BGGRRRRRRR

Path 3: BBBBBBGGRR

...



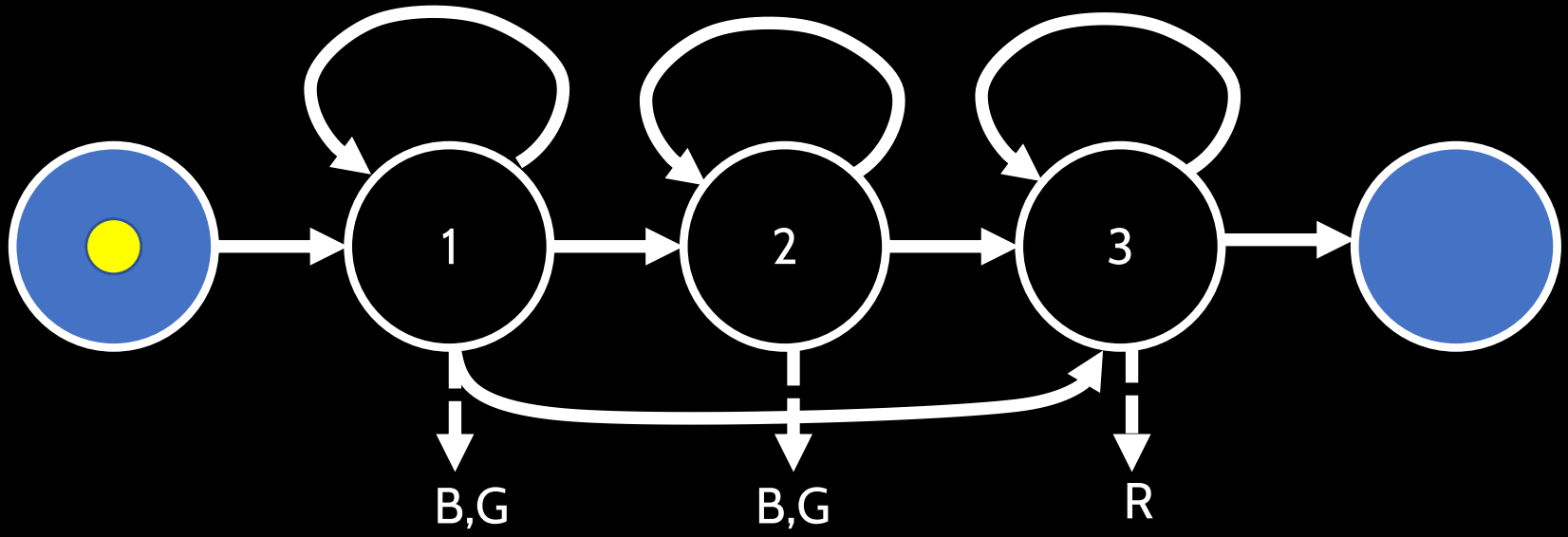
Markov model



State sequence: 1 1 2 2 3

State sequence: 1 1 1 3 3

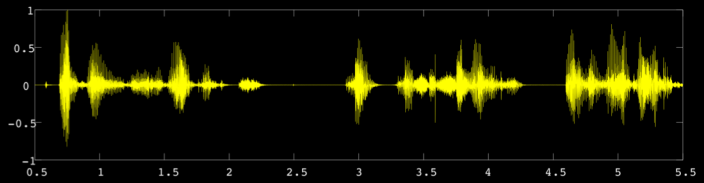
Hidden Markov model



Output sequence: BBGR

Output sequence: BBGRR

Putting it all together

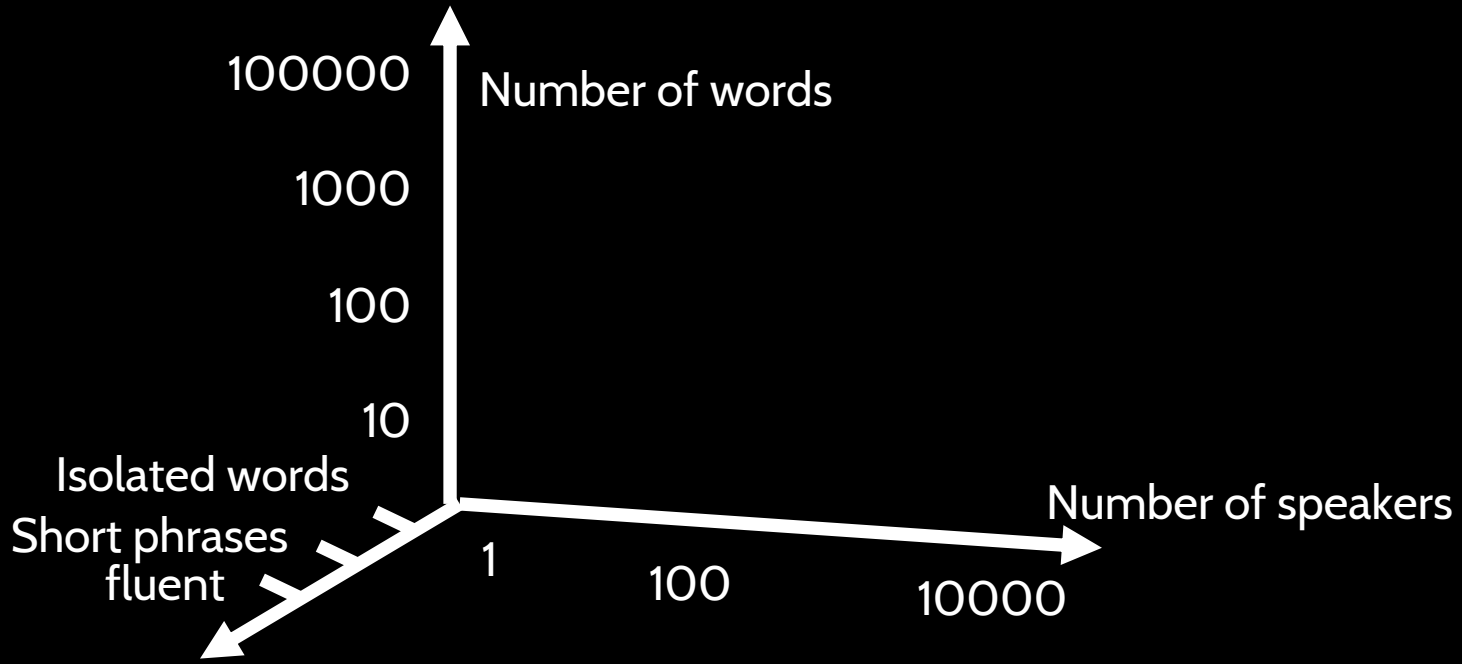


aɪ ˈpleɪɪŋ ɔːl ðə raɪt nnɔːts nɒt ˈnɛsɪsərɪli ɪn ðə raɪt ˈɔːdə

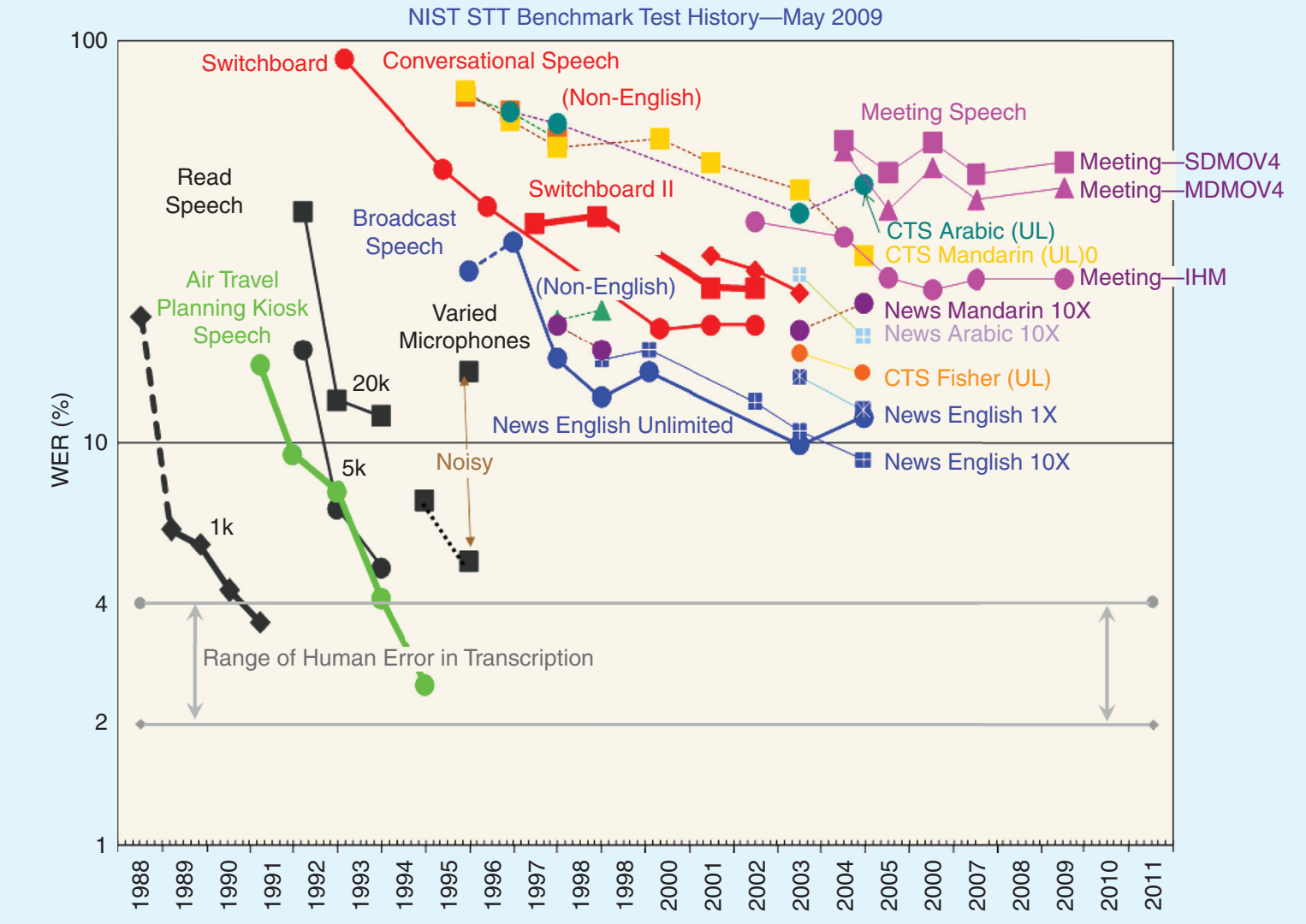
Phone sequence is usually full of errors!

Solutions:
Viterbi coding
Nbest sequences
Ngram models
Language modelling
....
Many many more

Why is speech recognition tricky?



Not shown: noise conditions, accent



Xaiodong He and Li Deng, Speech Recognition, Machine Translation and Speech Translation - A Unified Discriminative Learning Paradigm, IEEE Signal Processing Magazine (1), Sept 2011, pp 126--133

Quo vadis speech recognition?

Huge efforts from Apple, IBM, Microsoft, Google, Amazon, Xiaomi

”Deep learning ” and “end to end” recognition

Plenty of problems still to be solved in language though...

Ladle Rat Rotten Hut

as read by Vivian Altman

Wants pawn term, dare worsted ladle gull hoe lift wetter murder inner ladle cordage, honor itch offer lodge dock florist. Disk ladle gull orphan worry ladle cluck wetter putty ladle rat hut, an fur disk raisin pimple colder Ladle Rat Rotten Hut.

Wan moaning, Rat Rotten Hut's murder colder inset, "Ladle Rat Rotten Hut, heresy ladle basking winsome burden barter an shirker cockles. Tick disk ladle basking tutor cordage offer groin-murder hoe lifts honor udder site offer florist. Shaker lake! Dun stopper laundry wrote! An yonder nor sorghum-stenches, dun stopper torque wet strainers!" ...

What about visual speech?

Benjamin Franklin, in 1785 writing about his new invention, bifocal spectacles ...

By this means, as I wear my spectacles constantly, I have only to move my eyes up or down, as I want to see distinctly far or near, the proper glasses being always ready. This I find more particularly convenient since my being in France, the glasses that serve me best at table to see what I eat, not being the best to see the faces of those on the other side of the table who speak to me; and when one's ears are not well accustomed to the sounds of a language, a sight of the movements in the features of him that speaks helps to explain,

so that I understand French better by the help of my spectacles.

Beer is proof that God loves us and wants us to be happy.

McGurk effect

Reasons why lip-reading is tricky?

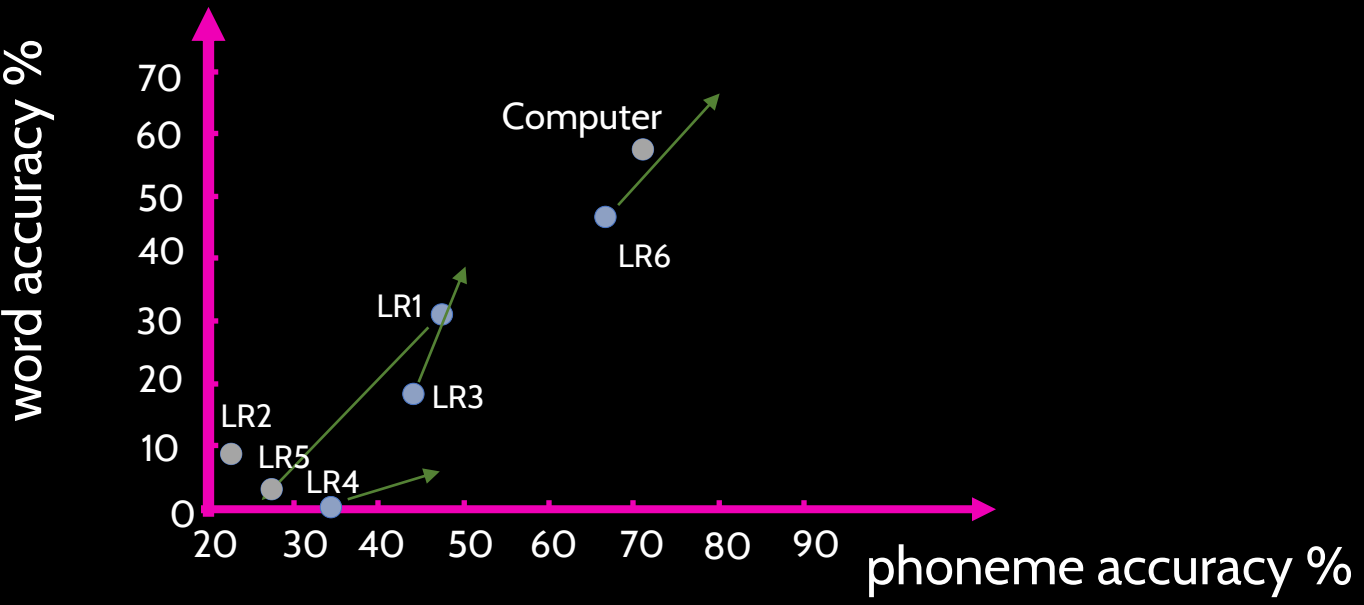
- Not all speech information appears on the lips
- Tracking the lips is tricky
- Extracting features is tricky
- Even humans find it difficult

But

very useful in noisy situations

Keyword spotting

Lip-reading performance 2012



Summary

- Slow and incremental progress has made speech technology practical
- Most effective implementations are in the hands of commerce
- Further work needed on
  - the connection with machine understanding
  - visual speech

- But visual speech implies a whole new modality – vision.

Next lectures:

vision (12th Feb)
learning (19th March)
text (16th April)
creativity (28th May)

Credits

Professor Stephen Cox, UEA

Dr Ben Milner, UEA

The Worshipful Company of IT Professionals

The Exploratorium, San Francisco

Words for speech recogniser

Compressor, compression, expander, compander, Fourier, sampling, Nyquist, Shannon, analog-to-digital converter, Worshipful Company of Information Technologists, Sir Thomas Gresham, lip-reading, Benjamin Franklin, bifocals,