Tone Without Pitch

Mark Liberman
http://ling.upenn.edu/~myl

Sources of the material in this talk:

Neville Ryant, Jiahong Yuan, and Mark Liberman,
"Mandarin Tone Classification Without Pitch Tracking",
IEEE ICASSP 2014

Neville Ryant, Malcolm Slaney, Mark Liberman, Elizabeth Shriberg, and Jiahong Yuan,
"Highly Accurate Mandarin Tone Classification In The Absence of Pitch Information",
Speech Prosody 2014

[some unpublished work with Jianjing Kuang]

Terminology: F0 vs. Pitch vs. Tone

F0: An objective (?) physical quantity –
(1) Lowest-frequency (?) local quasi-periodicity in the time domain
Dominant peak in serial cross-correlation function, … , etc.
(2) Greatest common divisor of peaks in spectral fine structure = "harmonics"
Dominant peak in (high-quefrency) cepstrum

Problem: Estimates of these quantities are often ambiguous or unstable --
Small differences in input can lead to big changes in estimates:
(e.g. octave jumps, etc.)

Pitch: A subjective (perceptual) quantity –

Issues: Perceptual "illusions" – pitch is often not = F0
Sounds with inharmonic components
Perception of levels / intervals vs. perception of pitch glides?

Tone: Linguistic (phonetic/phonological) categories or dimensions

Missing:

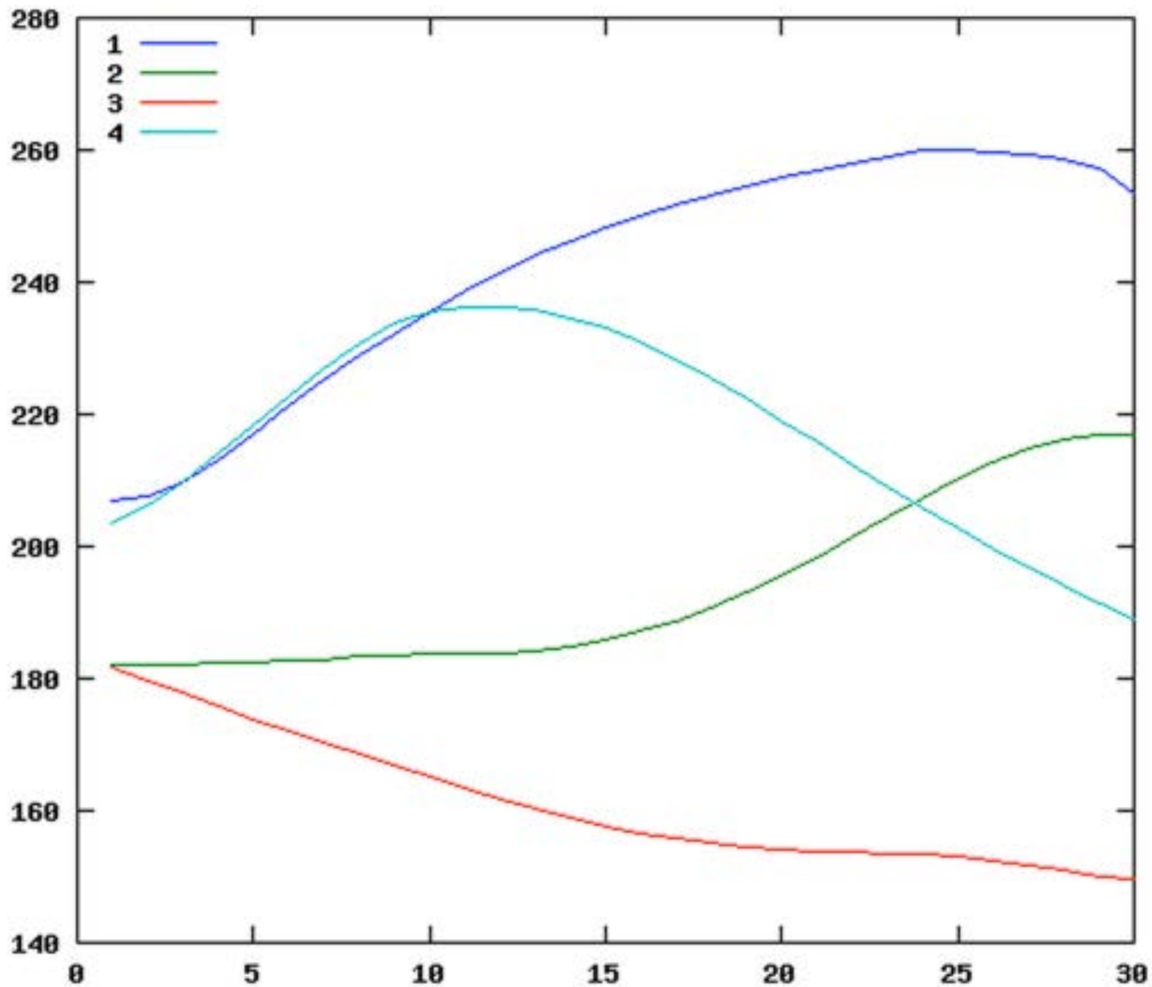
A term for the complex of articulatory adjustments
involved in creating and adjusting laryngeal oscillation:

- subglottal pressure
- supraglottal impedance
- tension of various muscles:
  cricothyroid, interarytenoids, vocalis, geniohyoid, sternothyroid, etc.
- ???

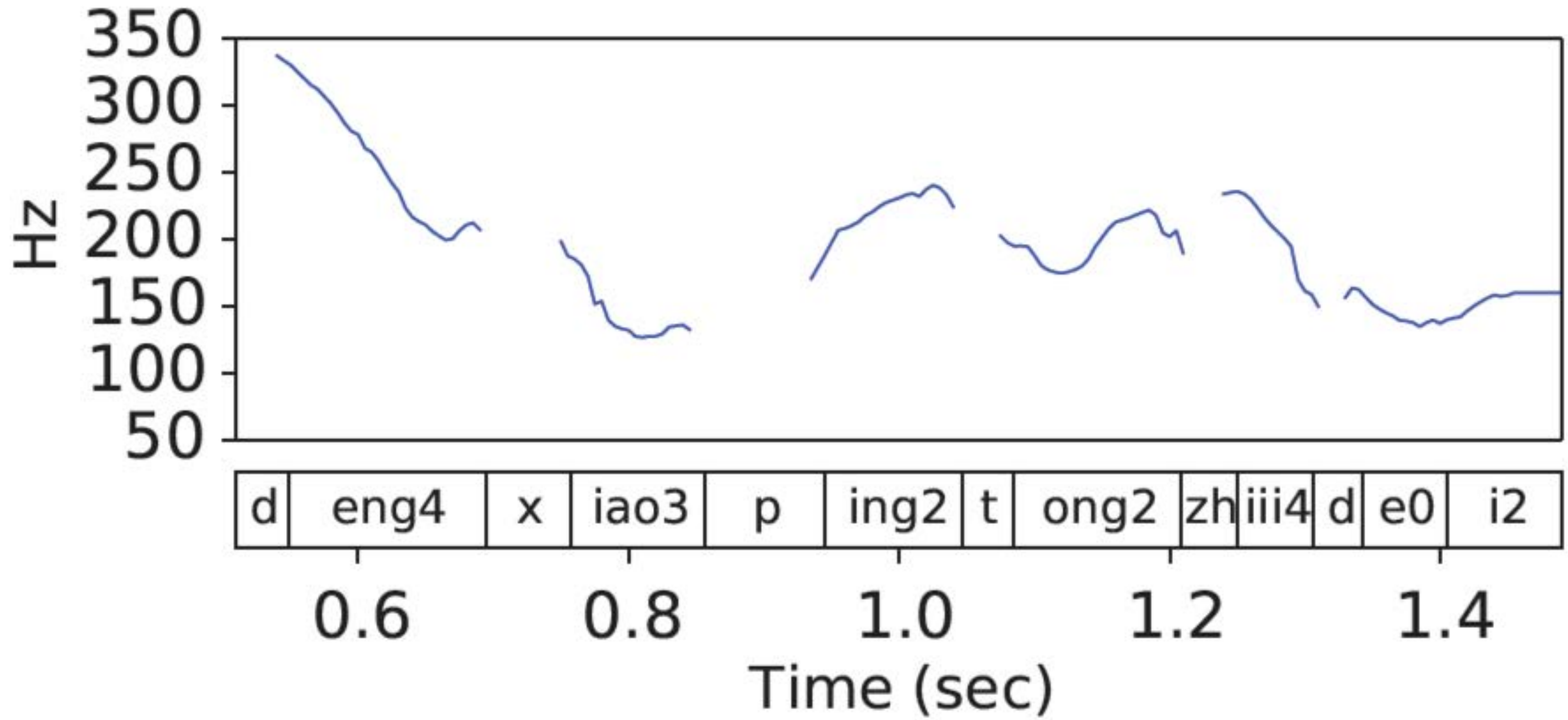
Mandarin tones:

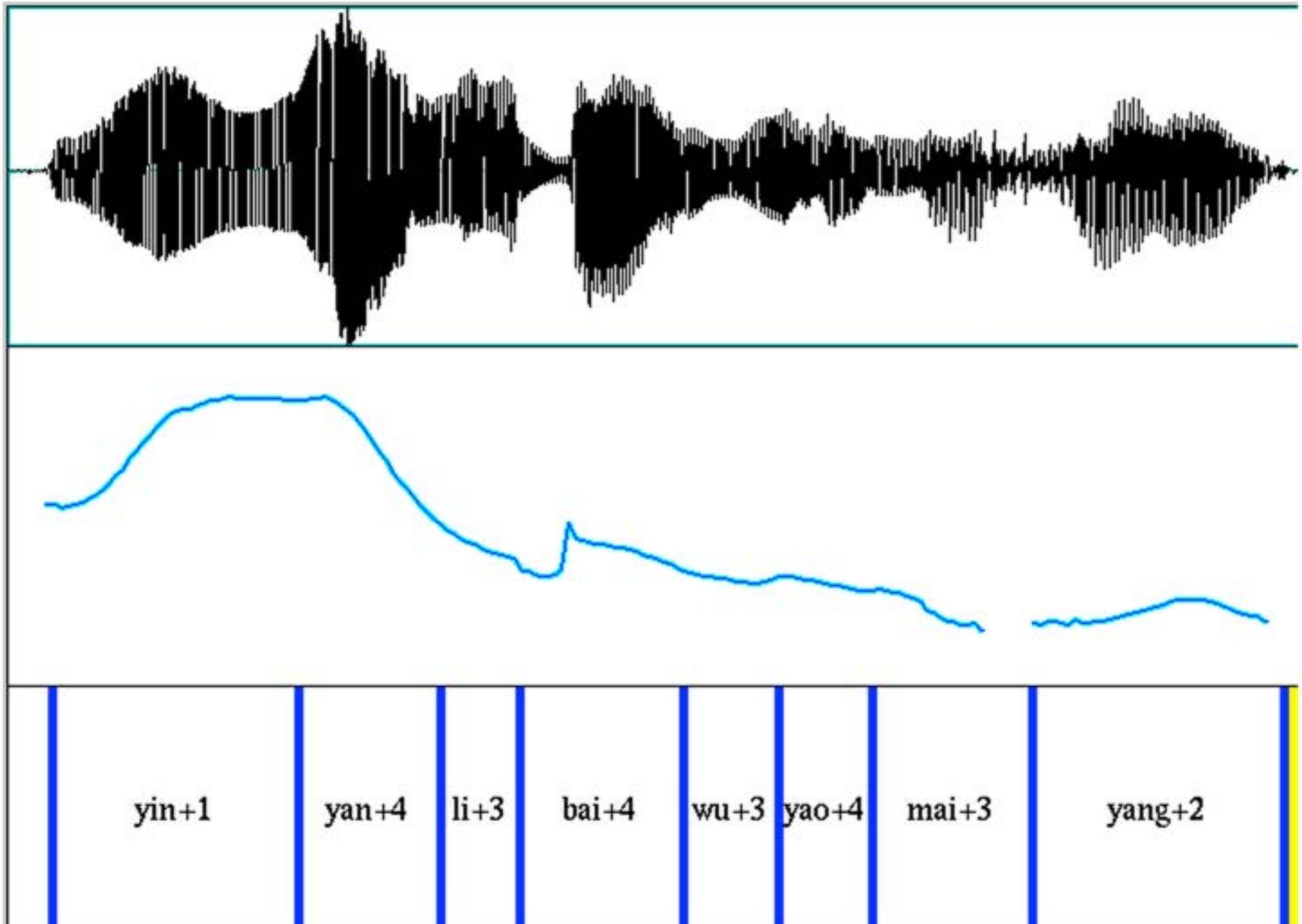
Mean of [7,992 time-normalized syllables]

in sentence context from 8 speakers:



Mandarin tones in context:




yin+1
yan+4
li+3
bai+4
wu+3
yao+4
mai+3
yang+2

Automatic Mandarin Tone Classification – Corpus details:

1997 Mandarin Broadcast News
(LDC98S73 & LDC98T24)

from China Central TV, Voice of America, KAZN-AM

	Speakers	Hours	Utterances	Segments	TBUs
Train	20	6.05	7,549	196,330	96,697
Test	6	0.22	300	7,189	3,464

Segmentation: nonspeech, initials, finals

Tone-bearing units (TBUs) are finals

Test set details:

300 utterances:
  50 chosen at random
    from each of six speakers

Segmentation and tone labeling by hand
 (Chinese native speaker/phonetician)

Training-set details:

Forced alignment of the HUB-4 training utterances
by an HMM-based forced aligner
using the CALLHOME Mandarin Chinese Lexicon and HTK

The aligner employed explicit phone boundary models
and achieved 93.1% agreement within 20 ms
compared to manual segmentation on the test set

[Yuan, Ryant & Liberman,
"Automatic Phonetic Segmentation in Mandarin Chinese", ICASSP-2014]

We also checked tone labels
for 1,252 syllables in 100 training utterances –
15 (1.2%) syllables had the wrong tone.

Input Features:

40 Mel-Frequency Cepstral Coefficients (MFCCs)
  25-ms Hamming window
  1024-point DFT
  40 triangular filters evenly spaced on mel scale 0-8000 Hz
  per-utterance cepstral mean-variance normalization

MFCCs for 21 frames at offsets of -100, -90, …, +100 ms
  = 840-dimensional input vector

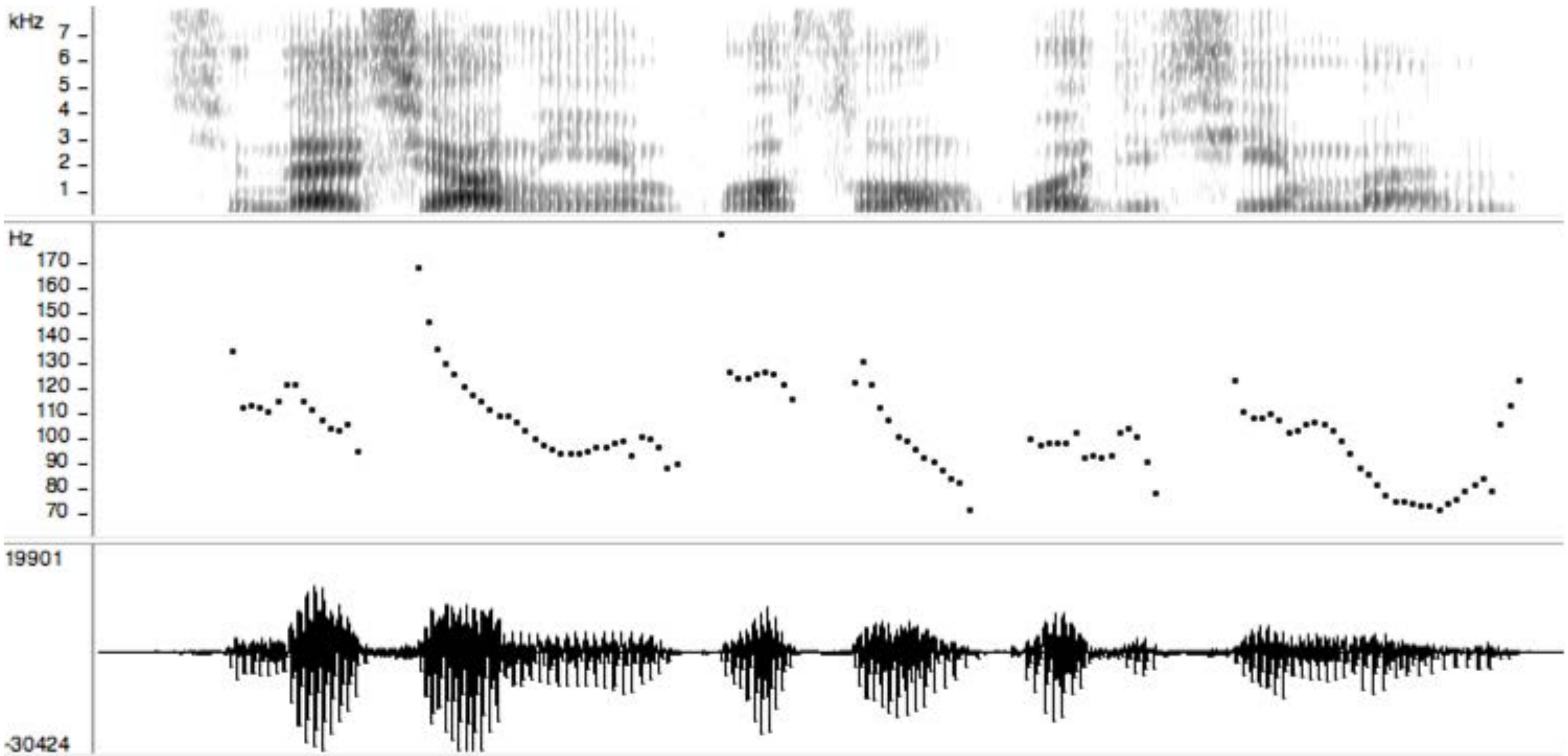
Advanced in 1-ms steps

= “coarse spectrogram”

Mel vs. linear scale doesn’t really seem to matter
“Cepstrum” (i.e. cosine transform of log power spectrum) also doesn’t matter

Brief excursus on “F0”

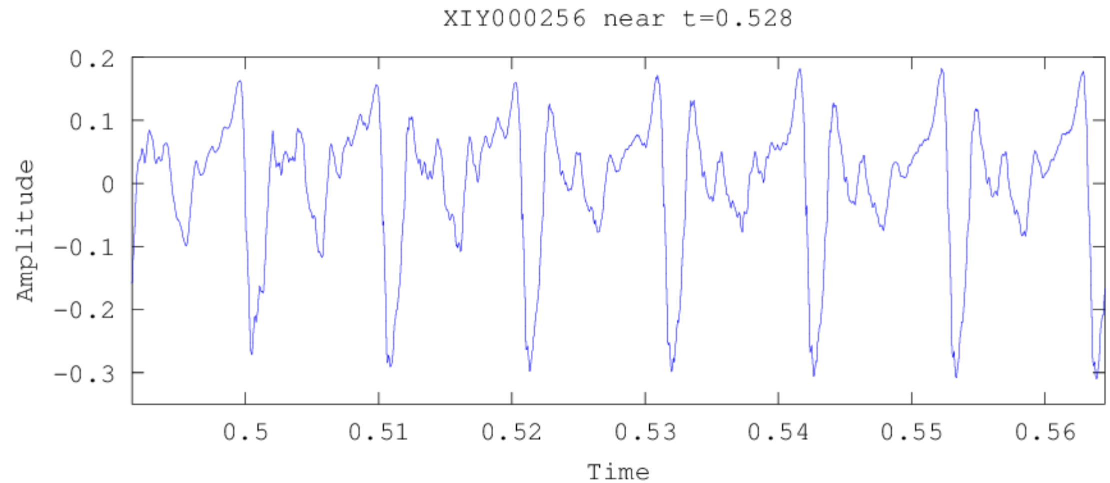
and on the input to our classifier…



kHz
7
6
5
4
3
2
1
Hz
170
160
150
140
130
120
110
100
90
80
70
19901
-30424

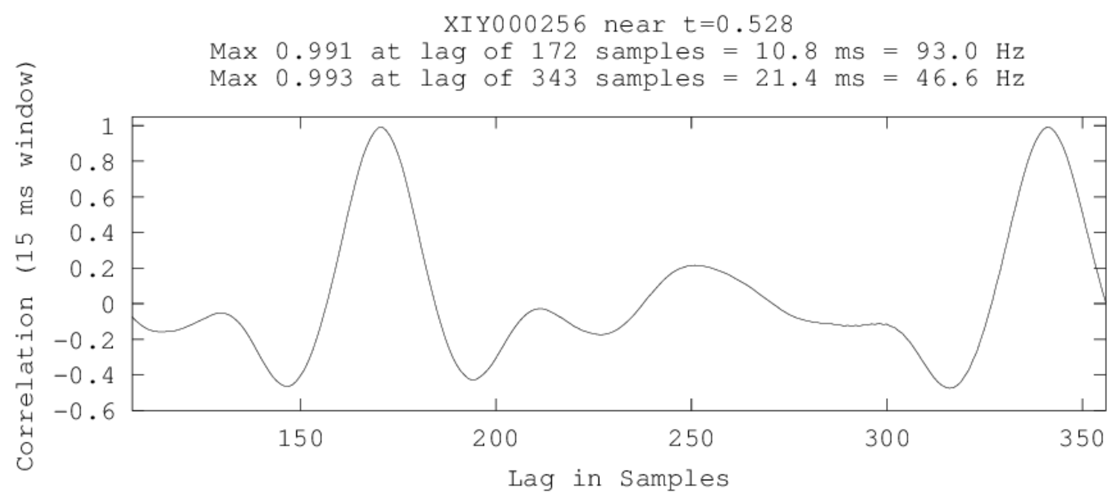
F0 =

time domain quasi-periodicity:



“Serial cross-correlation” is one way to find time-domain periodicity:

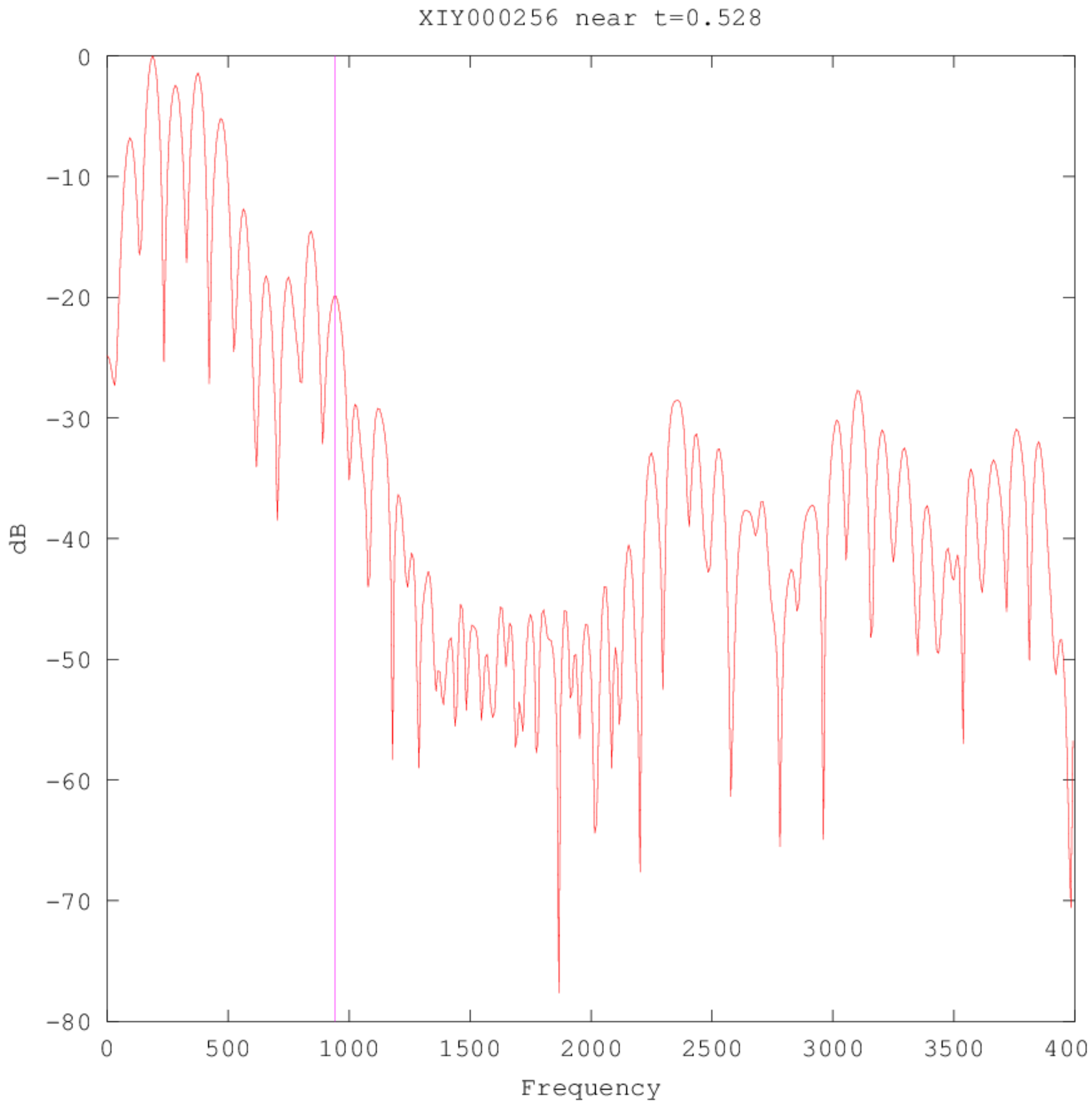
(note period doubling)



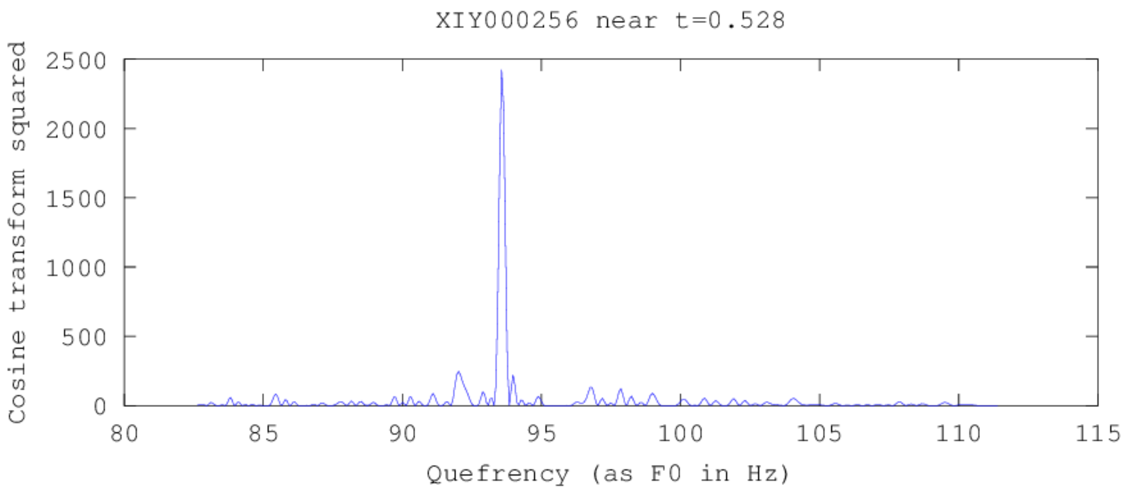
F0 is also periodicity
in the frequency domain
(“spectral fine structure”),
visible if the analysis window
includes at least two periods.

Here this
also suggests 93 Hz:

thus 10^{th} harmonic = 930 Hz

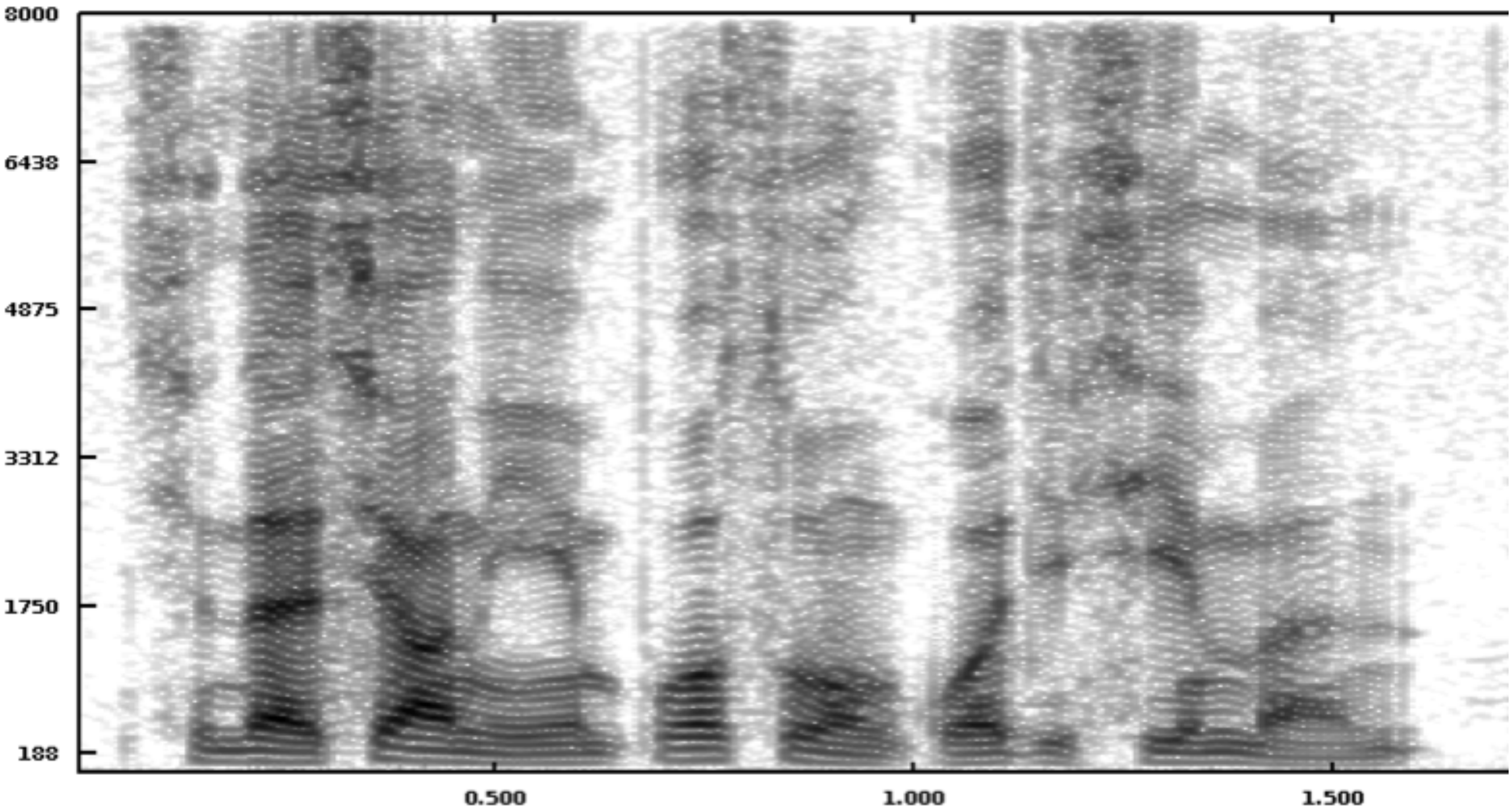


The cepstrum (i.e. cosine transform of log power spectum)
can be used as a technique to find frequency-domain periodicity --
e.g. here a peak at quefrency corresponding to 93 Hz:

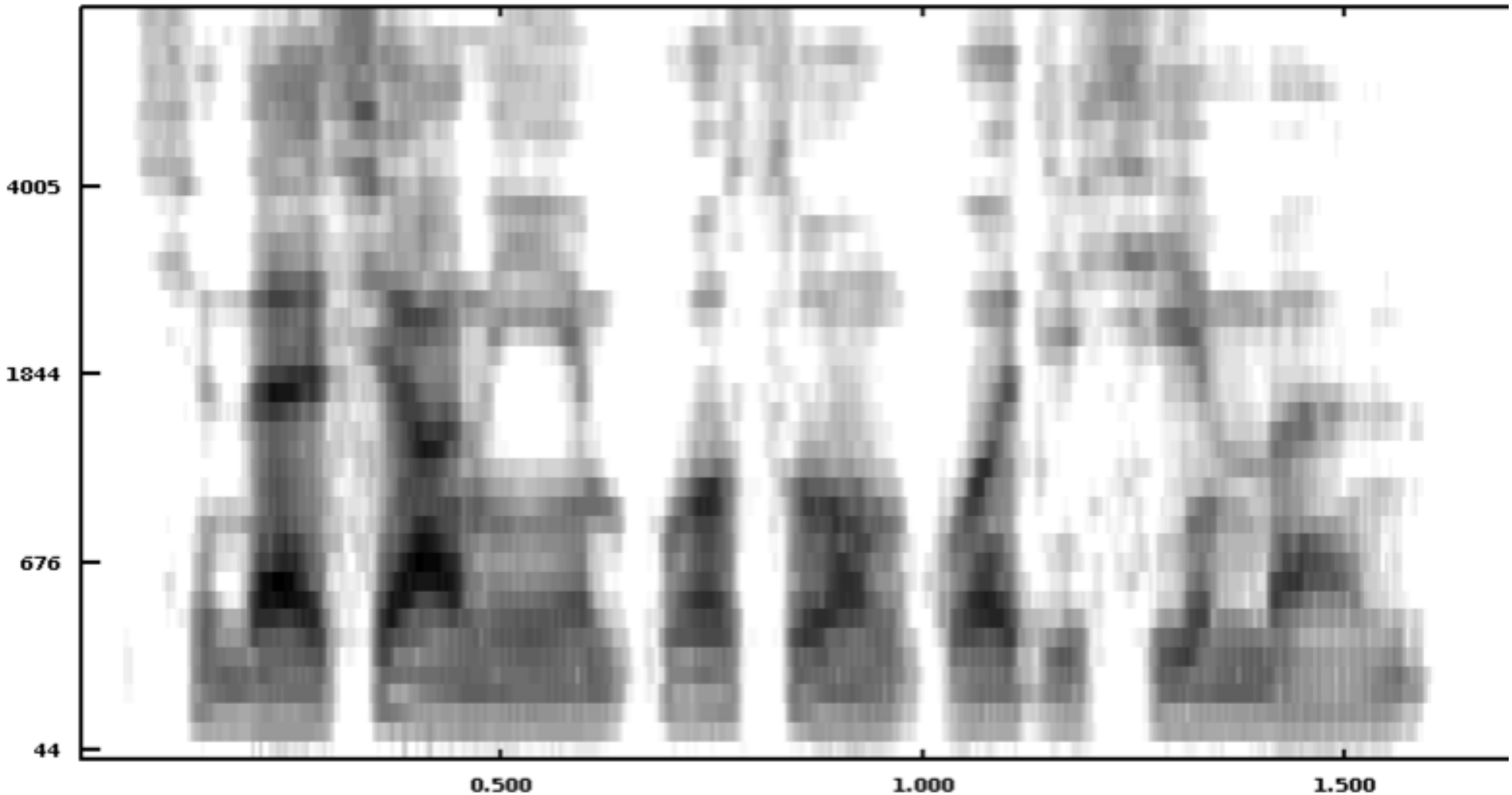


So what about the input to our classifier?

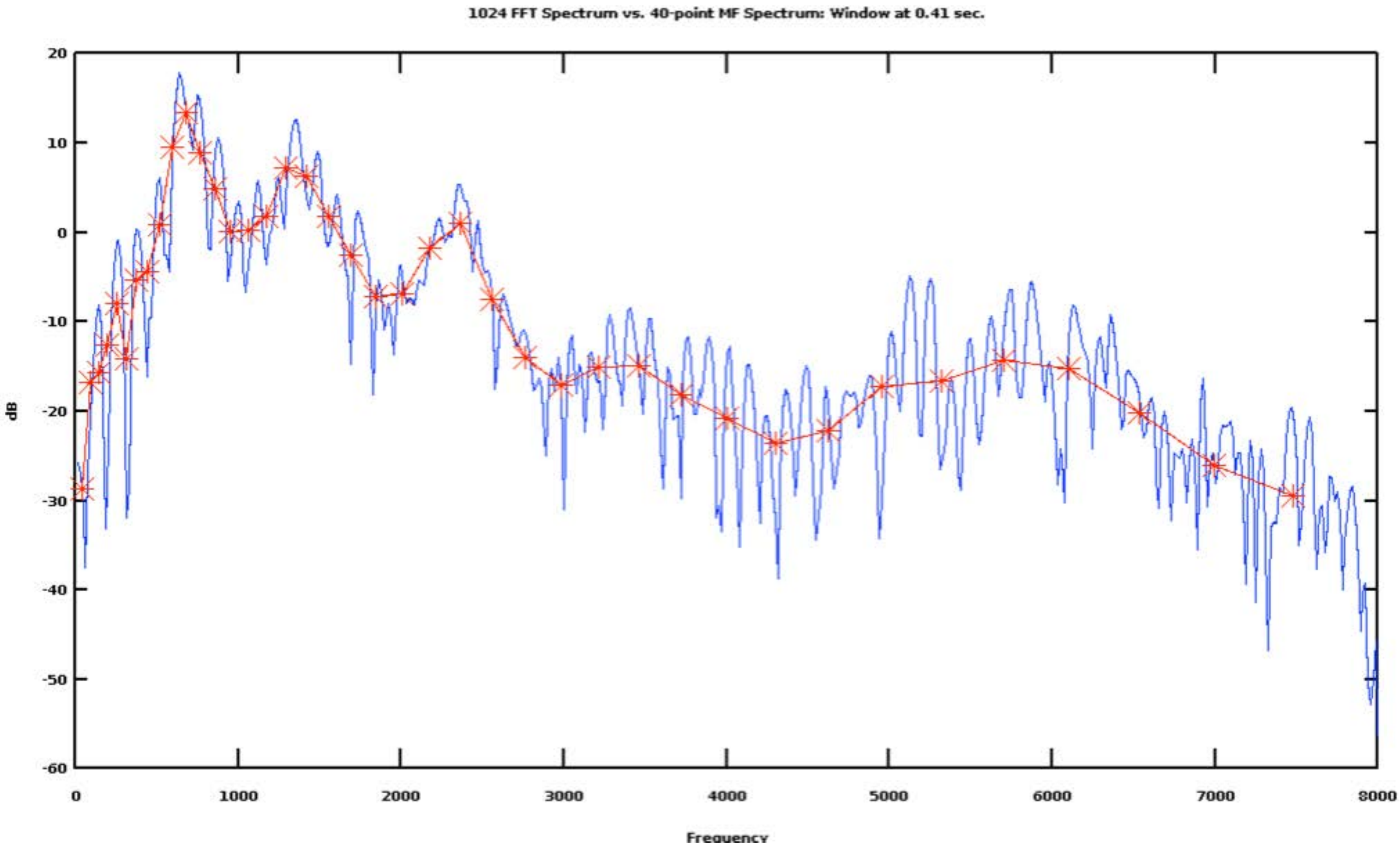
We start with linear-frequency, 1024-point FFT power spectra based on 25-msec analysis windows – = “narrow-band spectrogram”:



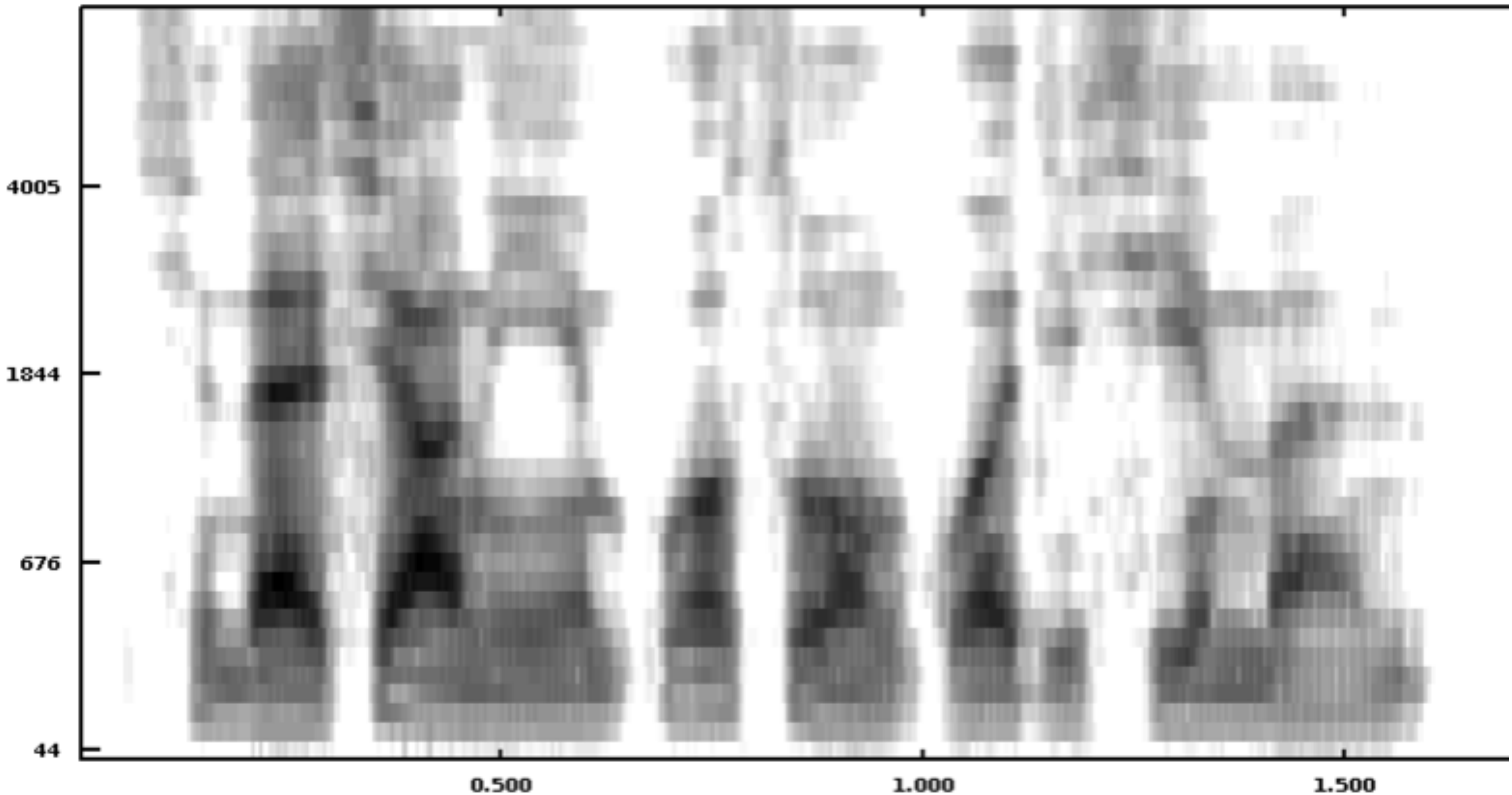
Then we smooth each spectral slice via 40 triangular filters,
evenly spaced on the mel scale –
This yields a coarsely-pixelated spectrogram in which F0 is not visible
either in the time domain or in the frequency domain:



Here's what one spectral slice looks like --
In the frequency domain, the spectral fine structure is smoothed out
(and likewise in the time domain given the 25-msec analysis window):



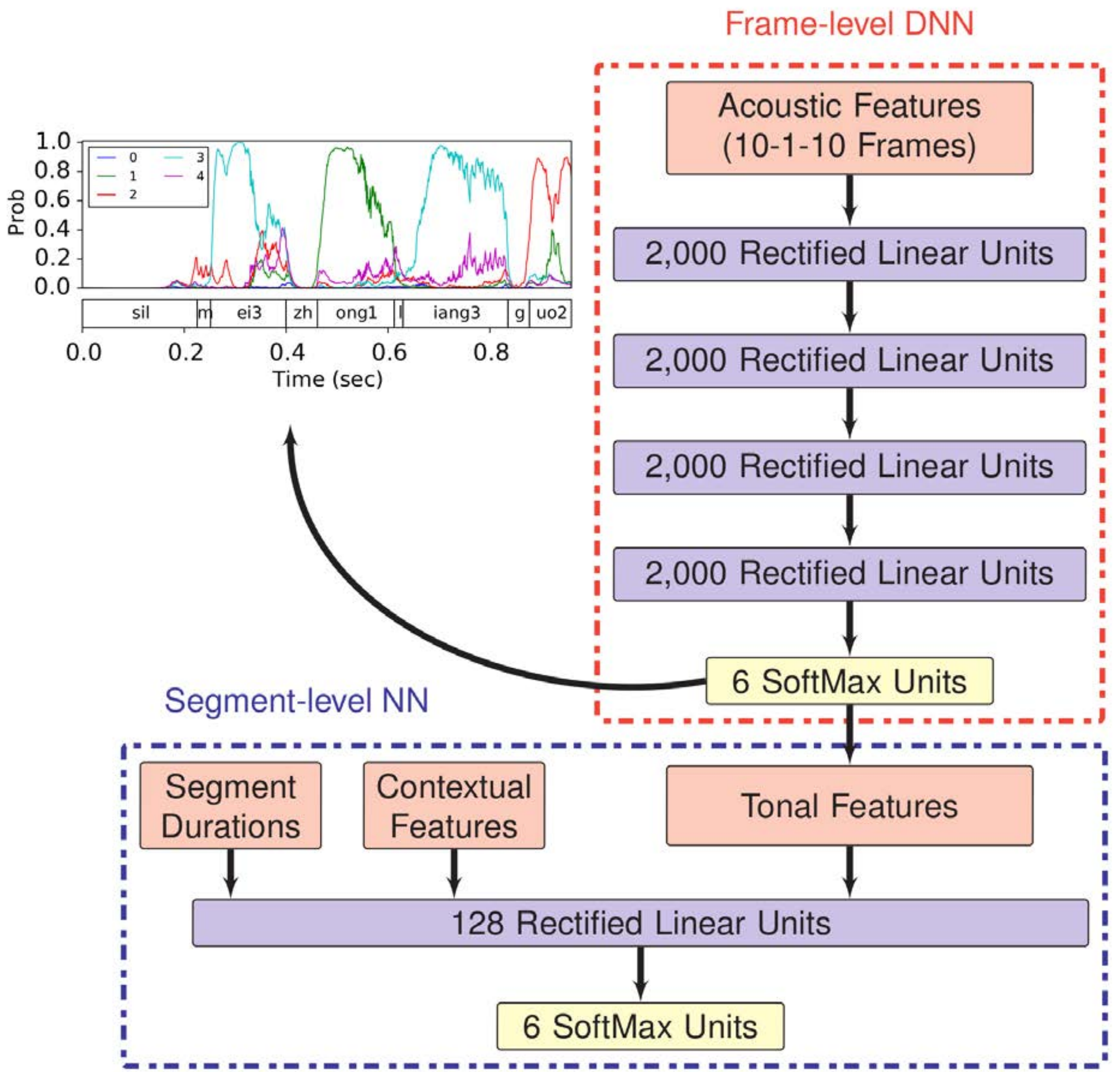
F0? Not so much…



Malcolm Slaney:

"Um, this is the worst design for a pitch tracker
that I've ever seen."

System architecture

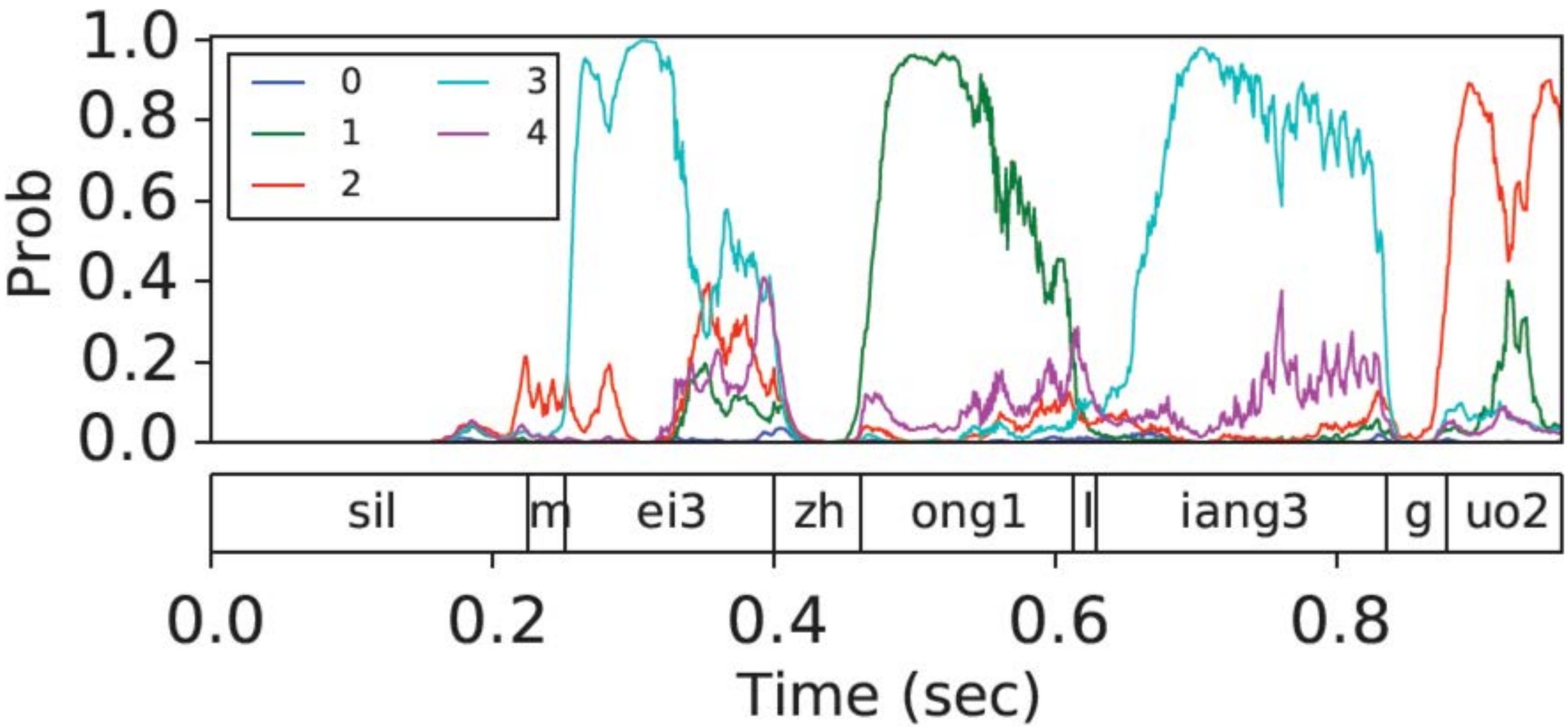


DNN details:

Input layer of 840 features
    40 MFCCs every 10 ms in 21 frames (10+1+10)
Four hidden layers of 2,000 Rectified Linear Units (ReLUs)
Output layer of 6 softmax units = 6 pseudo-probabilities:
    Tone 1-4, neutral tone (= tone 0), nonspeech
140 epochs of training, 250,000 examples each
Dropout: 20% input layer, 40% hidden layers
Initial learning rate η = 0.5, decaying after each epoch
Fixed momentum of 0.5

…Sacrificial first-born goat was white with black spots
  Offerings made to Athena and Eris
  For incantations see Hinton et al., Necr. Inf. Proc. 2013…

Example of Output --
6 pseudo-probabilities per 1-ms analysis:
(only five tonal categories shown)



Two notes on scoring:

(1) Silence is easy to recognize in this material,
so results can be artificially raised
depending on how much silence there is in the test set.

Therefore here we report "Frame Error Rate"
as only the percent incorrect classification of frames
in test regions that are within an utterance
according to the ground-truth segmentation

(2) In addition, we report "Segment Error Rate"
as the percent incorrect classification
of overall tone-bearing units in the test set.

From frame-wise classification to segment-wise classification:

Input features:

Tonal features of segment (= 6 mean DNN outputs)
Duration of segment
Tonal features + durations of adjacent two segments

Baseline: Class with highest posterior probability given “tonal features”

Three supervised methods:

L2-regularized logistic regression
SVM with RBF kernel
NN with single hidden layer of 128 ReLUs

MFCC System Results:

Within-utterance Frame Error Rate (FER): 16.36%
Within-TBU Frame Error Rate (FER): 27.38%
Baseline TBU Segment Error Rate (SER): 17.73%

Features	Logistic	SVM	NN
Tonal features + dur.	16.98	16.98	16.57
+ 1 segment context	16.69	16.75	15.96
+ 2 segments context	16.31	16.43	**15.62**

Best previously-reported results on Chinese Broadcast News:
FER 19.7% Lei et al. 2005 [17% relative reduction) [% silence comparable?]
SER 23.8% Lei et al. 2006 (34% relative reduction)

Simple pilot experiment to estimate human performance:

Even with steep low-pass cutoff at 300 Hz, listeners could often guess words,
so we synthesized sounds from F0 and amplitude contours
(RAPT pitch-tracking via get_f0, frame step = 5 ms,
fundamental + 9 overtones at 1/F amplitudes,
modulated to match measured F0 and amplitude in voiced regions)

15 utterances selected at random from each of the 6 test speakers, 982 tones total.
Chinese native speaker used Praat, with syllabic segmentation indicated,
to classify syllable tones interactively from AF-modulatation resynthesis.

Subsequent checking determined that 39 of the 982 syllables had synthesis problems
(mainly failure to detect voicing), so we retained judgments for 943.

The overall score was 71.79% correct = **29.21 SER**

% Correct for T0: 22.22%; T1: 78.87%; T2: 75.33%; T3: 59.35%; T4: 77.68%.
Without T0 (889 tones), the overall accuracy was 74.8% = **25.2% SER**

Now with F0 and amplitude as inputs:

Same DNN topology and training procedure –
but MFCCs input features are replaced with F0 and energy.

F0 computed with RAPT implemented in ESPS get_f0
per-utterance normalized to mean 0 and variance 1
in voiced regions.

Log-energy in 25 ms Hamming window
per-utterance normalized to mean 0 and variance 1.

Same 21-analysis-frame input, so 42-dimensional input vectors.

F0 System Results:

Within-utterance Frame Error Rate (FER): 24.22% (vs. 16.26%)
Within-TBU Frame Error Rate (FER): 40.05% (vs. 27.38%)
Baseline TBU Segment Error Rate (SER): 31.64% (vs. 17.73%)

SER results – MFCC system in parens for comparison:

Features	Logistic	SVM	NN
Tonal features + dur.	30.31 (16.98)	29.36 (16.98)	29.27 (16.57)
+ 1 segment context	29.04 (16.69)	26.56 (16.75)	24.83 (15.96)
+ 2 segments context	27.51 (16.31)	26.33 (16.43)	**22.66 (15.62)**

F0 system is much worse than MFCC system:

Within-utterance FER	24.22/16.26 = 49% worse
Within-TBU FER	40.05/27.38 = 46% worse
Best SER	22.66/15.62 = 45% worse

WHY?

Uninteresting hypothesis #1:

get_f0 is not a good enough pitch tracker

Test: try Dan Ellis's SAaC and Malcolm Slaney's SACD
(two recent "soft" estimators of F0 and F0 change…)

Result: F0 features still lose.

Uninteresting hypothesis #2:

We’ve somehow managed to build
a really terrific DNN-based pitch tracker….

Test: Use the same inputs and DNN architecture to predict F0
rather than tone class

Result: Didn’t work very well.

Uninteresting hypothesis #3:

Our DNN is recognizing phone sequences
and inferring tones from them

Test #1: Cheating experiment with perfect knowledge of true pinyin

Result: Not good enough, even with true pinyin:

	Overall FER	TBU FER	SER
Oracle (mono)	28.28	52.14	51.96
Oracle (tri)	11.54	21.27	20.76

And when we train our system as a pinyin recognizer, it has ~30% phone error rate…

Interesting hypotheses:

1. Tone is not (just) pitch

   There is other relevant information in the spectrum

2. And/or pitch is not (just) F0
   = dominant period of laryngeal oscillation
   = greatest common multiple
     of spectral fine structure frequencies)

   …either in production or in perception:
   Aspects of timbre also matter.

It's easy to show that both of these are true.

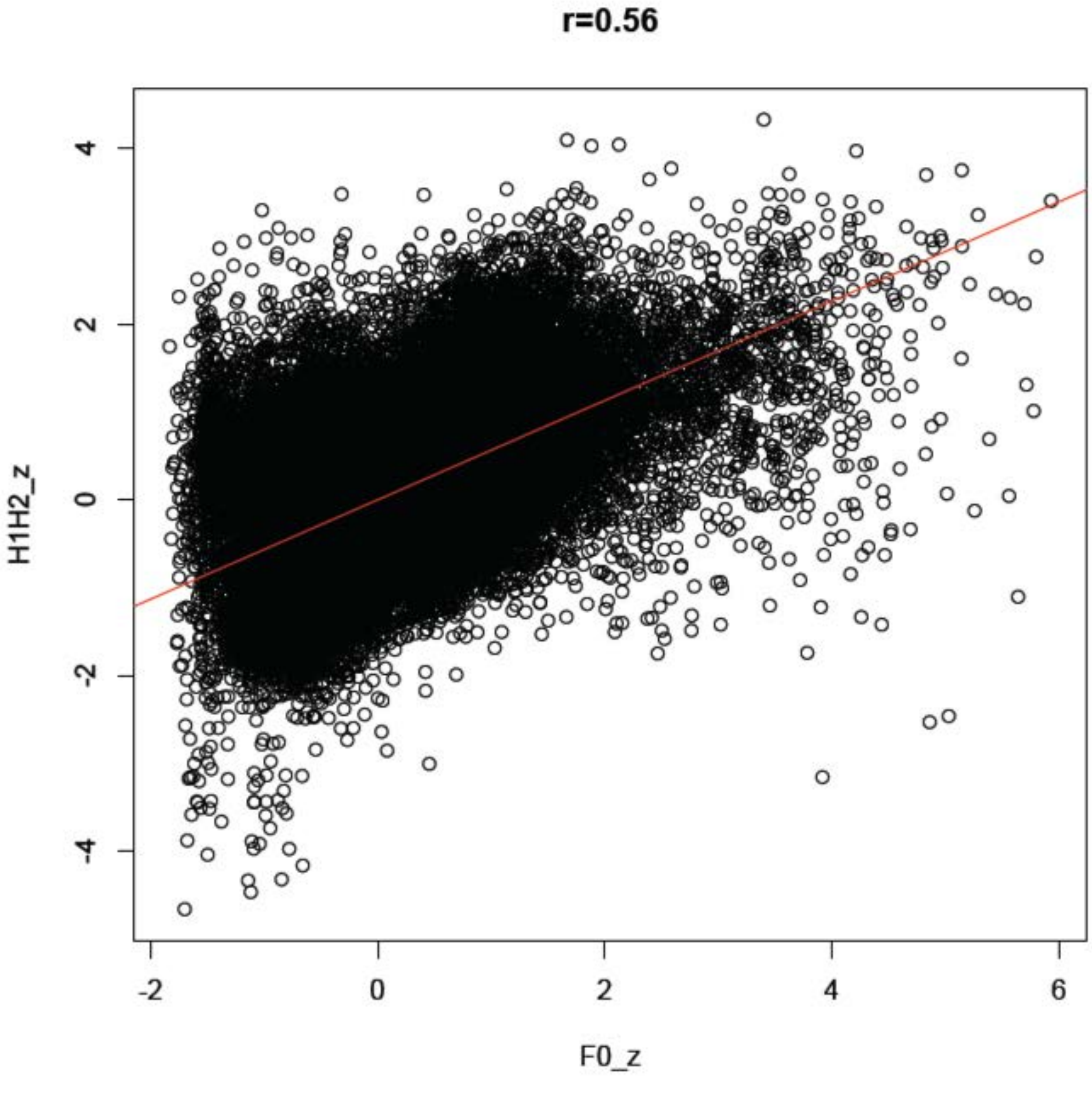
Some correlations from TIMIT…

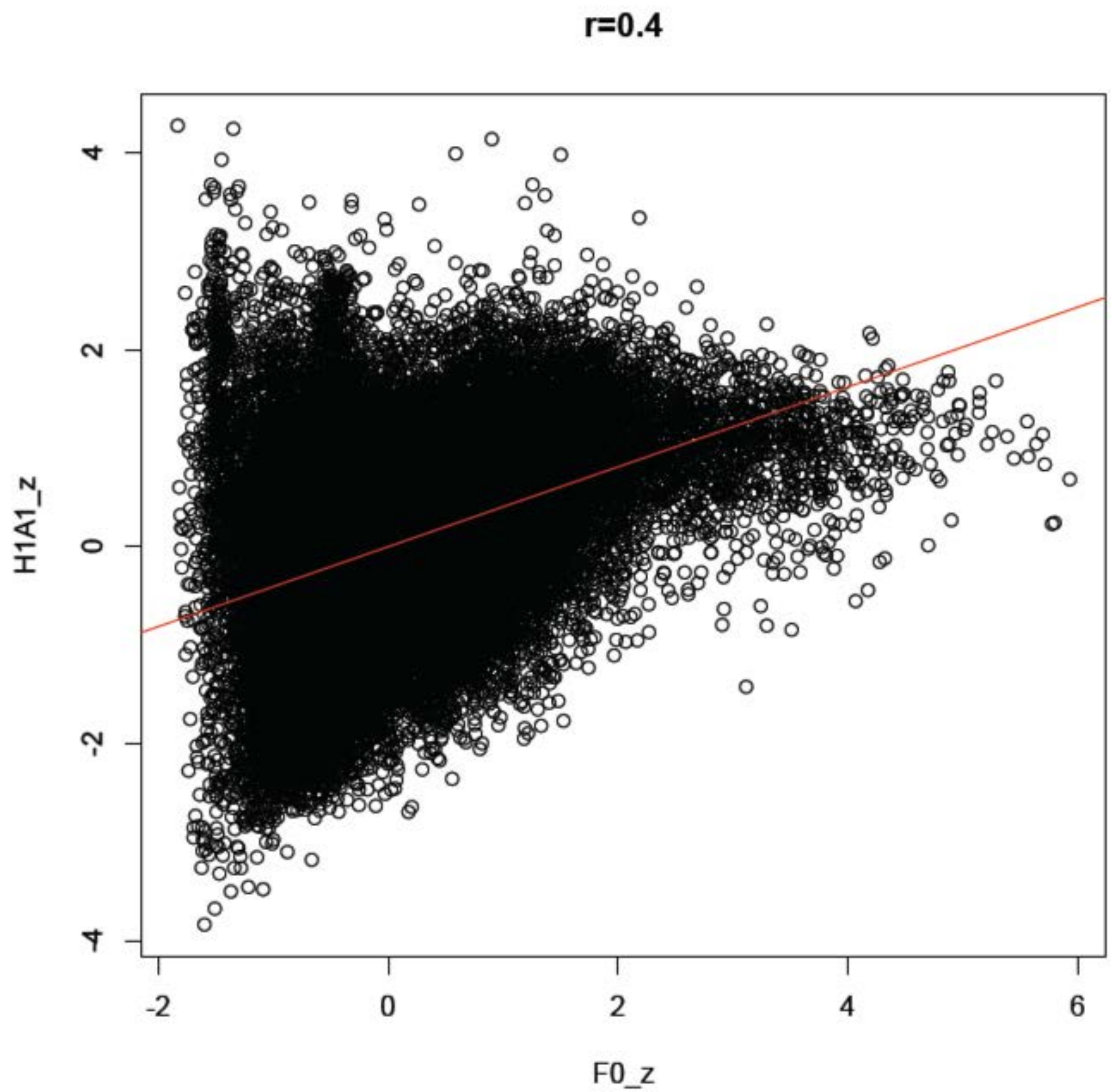
H1 & H2 are the amplitudes of the first two harmonics
(= F0 and 2*F0)

A1 is the amplitude of F1

Relationship of H1 to other spectral amplitudes
is commonly used as a measure of voice quality

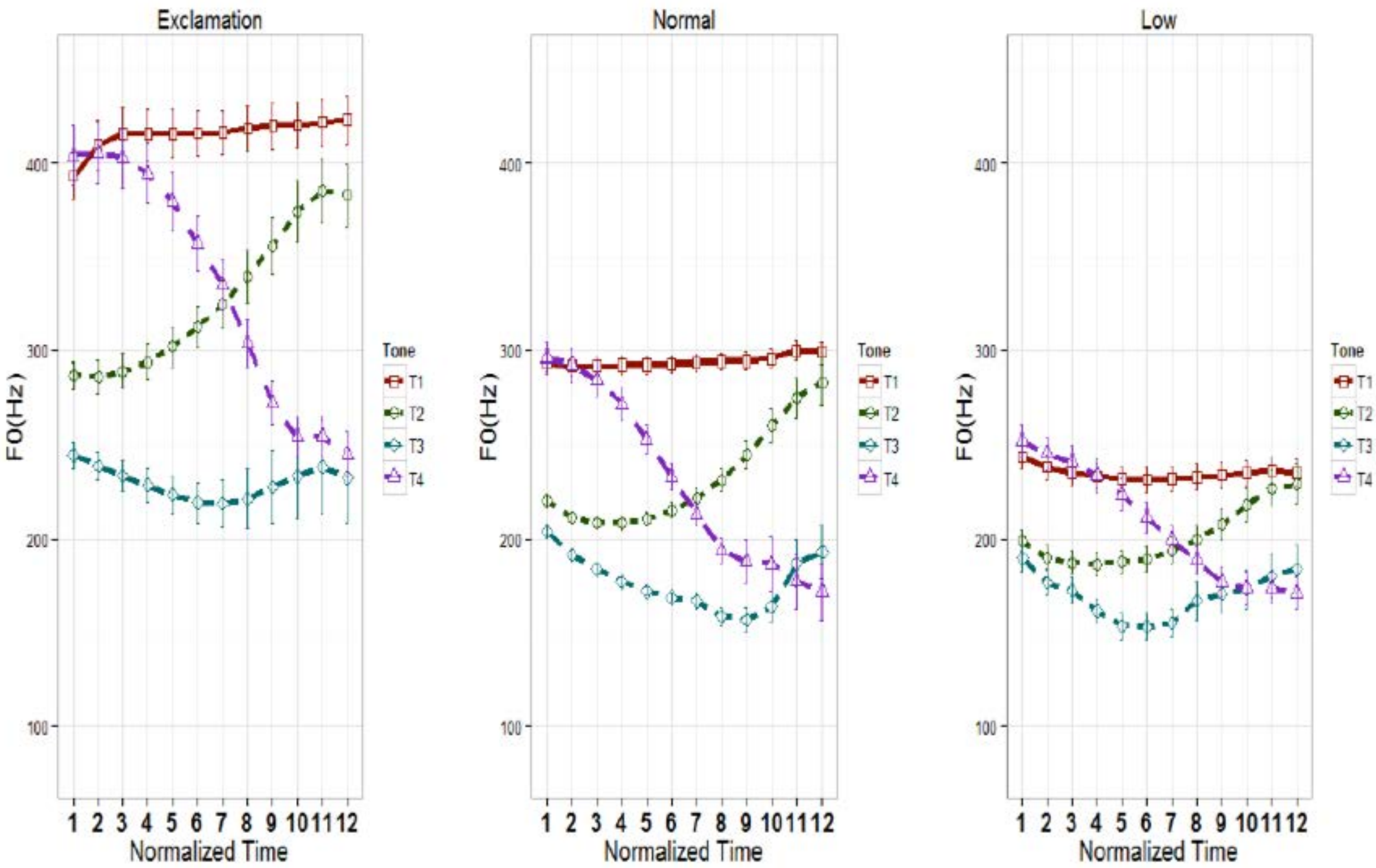
(Data from Jianjing Kuang)


r=0.56
H1H2_z
F0_z
4
2
0
-2
-4
-2
0
2
4
6

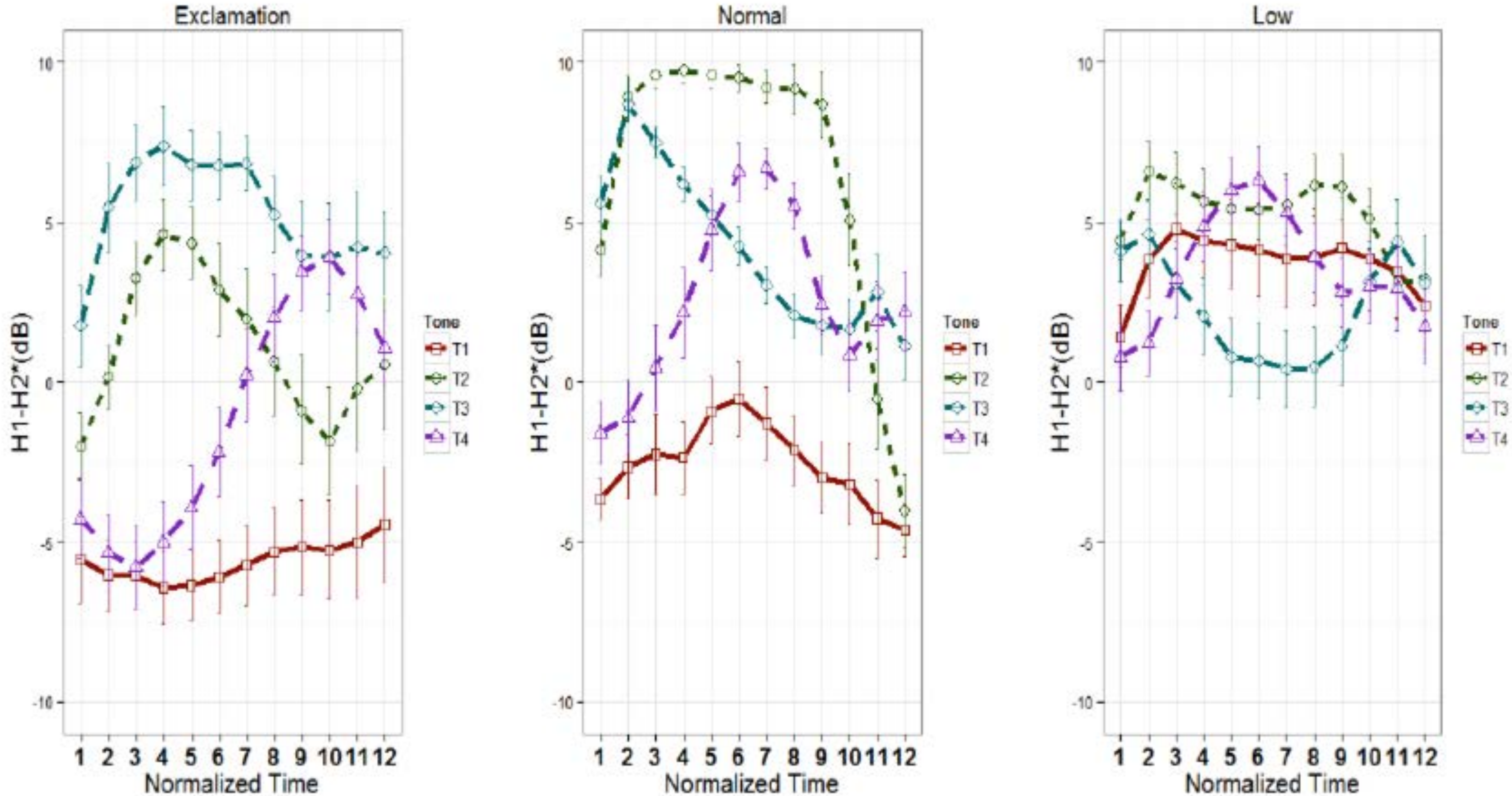


r=0.4
H1A1_z
F0_z
4
2
0
-2
-4
-2
0
2
4
6

Mean F0 from another Mandarin tone experiment:



And H1-H2 (in dB) from the same data:



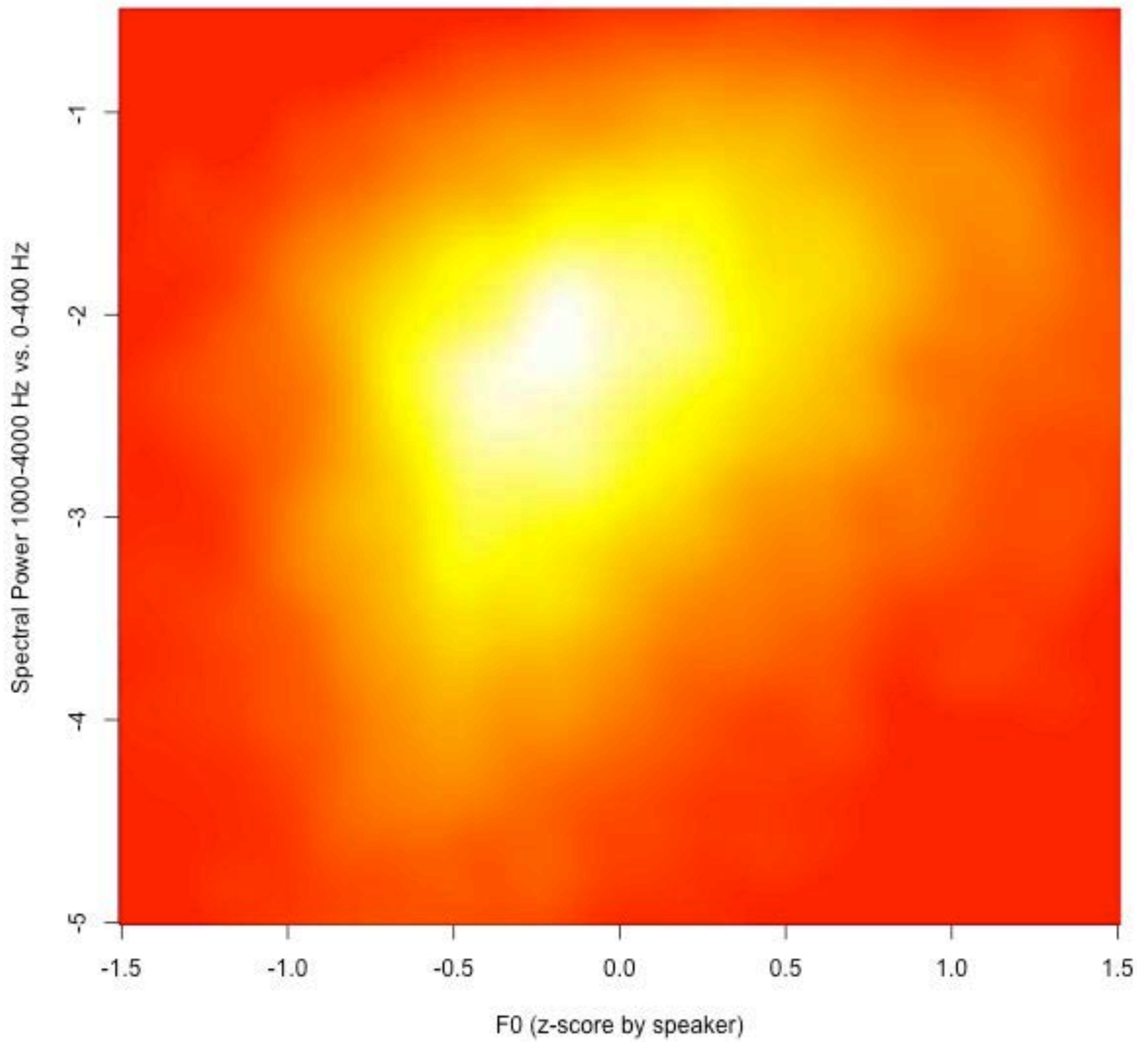
H1, A1, etc.

Have the same problems as F0 tracking:

Small differences in the input
lead to large differences in the output

But more robust measures look similar:

F0 vs. spectral balance 1000-4000/0-400 in TIMIT:



So maybe, even if we ignore spectral fine structure
(i.e. overtone spacing = high quefrencies),

the combination of "supra-laryngeal" spectral structure
(mid quefrencies = formant-like content)
and "voice quality" spectral structure
(lower quefrencies, sub-F1 energy, etc.)

gives a reliable measure of vocal effort
and local changes in (articulatory) pitch.

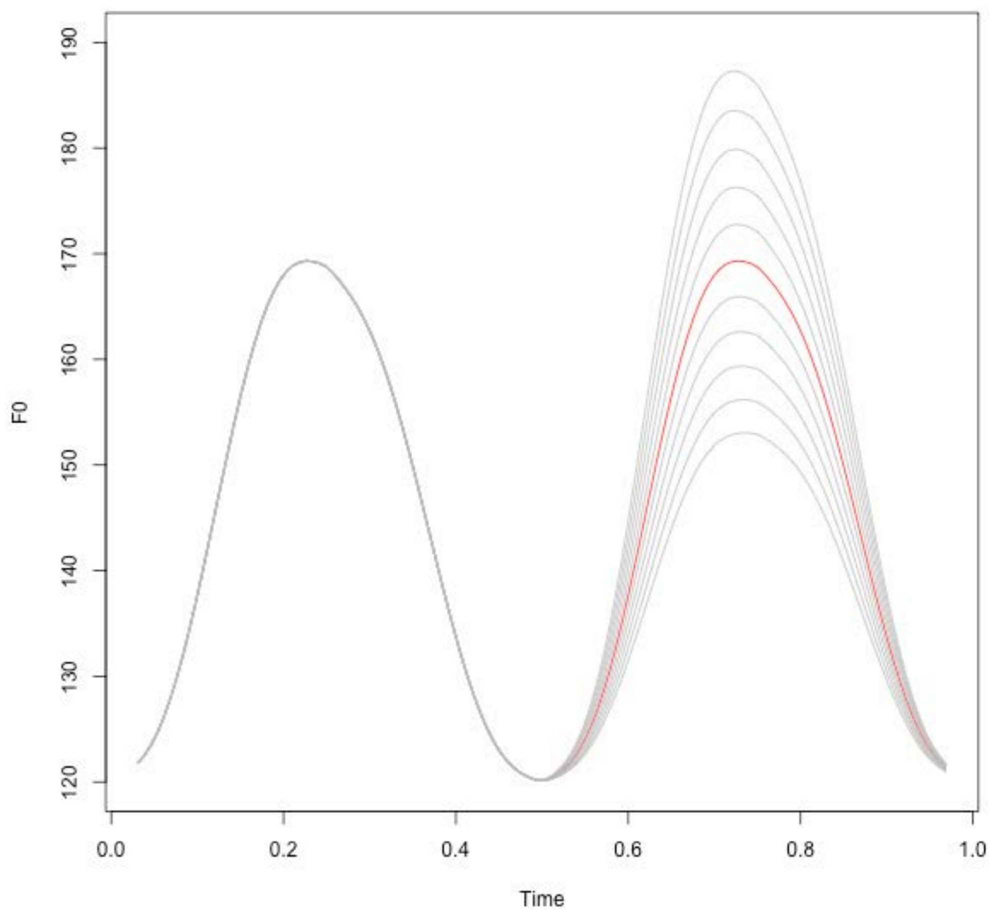
But it's a surprise
that gross spectral measures
are a better basis for tone classification
than F0 estimates are!

Given this,
we should expect to see
perceptual confusion in pitch perception
between F0 and timbre…

And we do!

[Joint work with Jianjing Kuang]

Frequency-modulated overtone series – 15 overtones, 11 F0 contours:



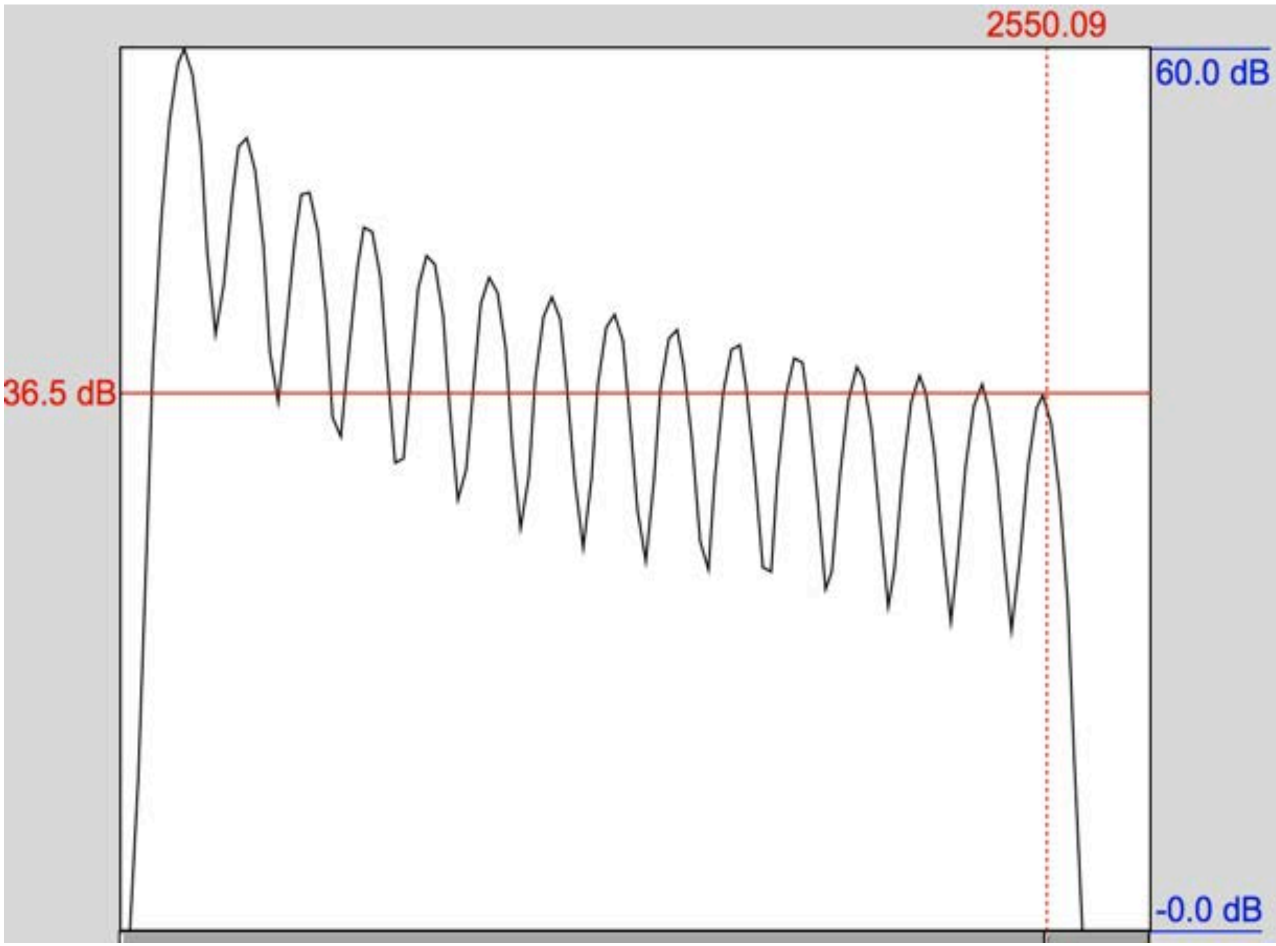
Four timbre conditions:

(A) Overtone amplitude varies as 1/F

(B) Constant-amplitude overtones

(C) First part = (A), second part = (B)

(D) First part = (B), second part = (A)



A B C D

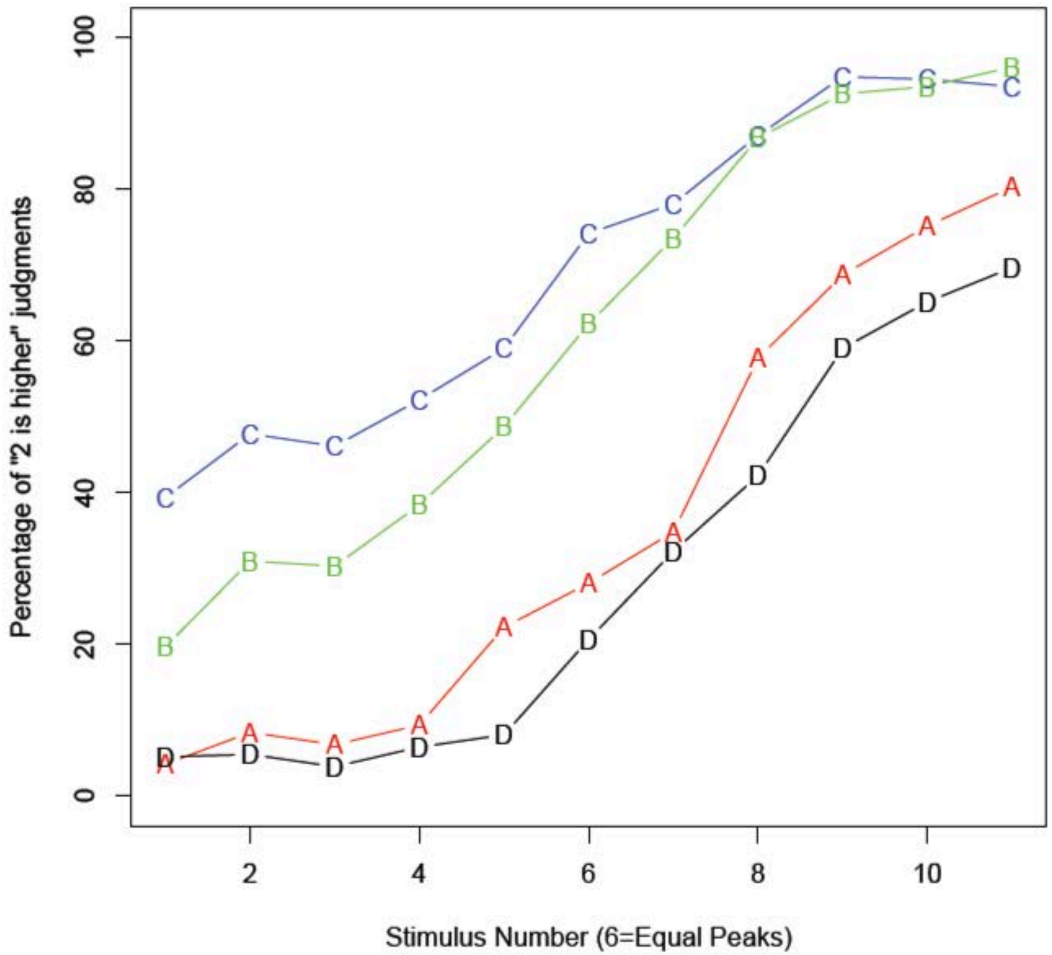
11 F0 steps X 4 timbre conditions = 440 stimuli

10 copies of each,
presented in random order to each subject

Forced-choice classification:
Is second peak higher or lower in pitch?

55 subjects from Penn's undergrad subject pool

Results (all subjects):



A B C D

The 55 subjects can be divided into three groups:

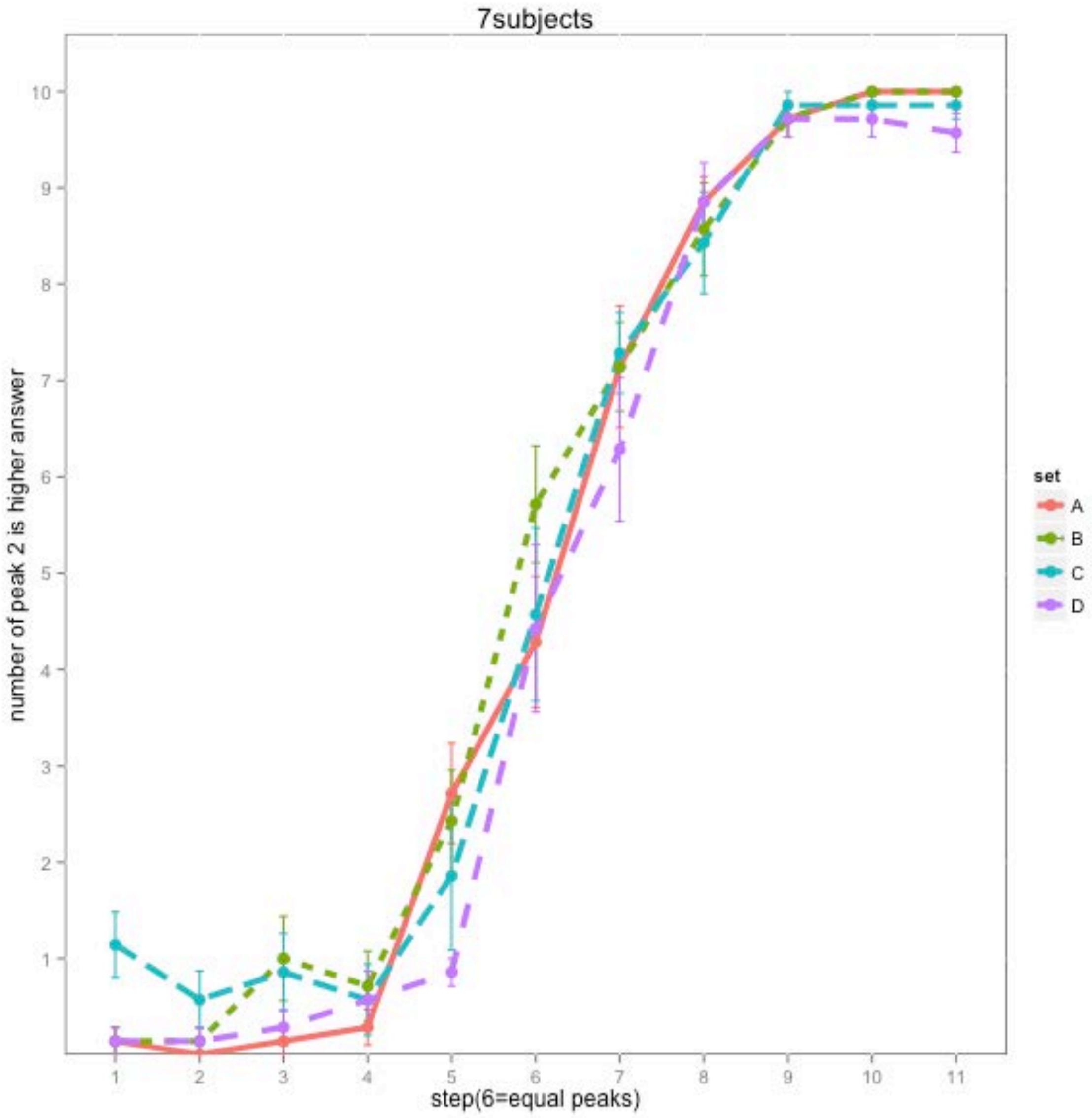
Those who attended only to F0 (7 subjects)

Those who attended only to timbre (8 subjects)

Those who attended to both (40 subjects)

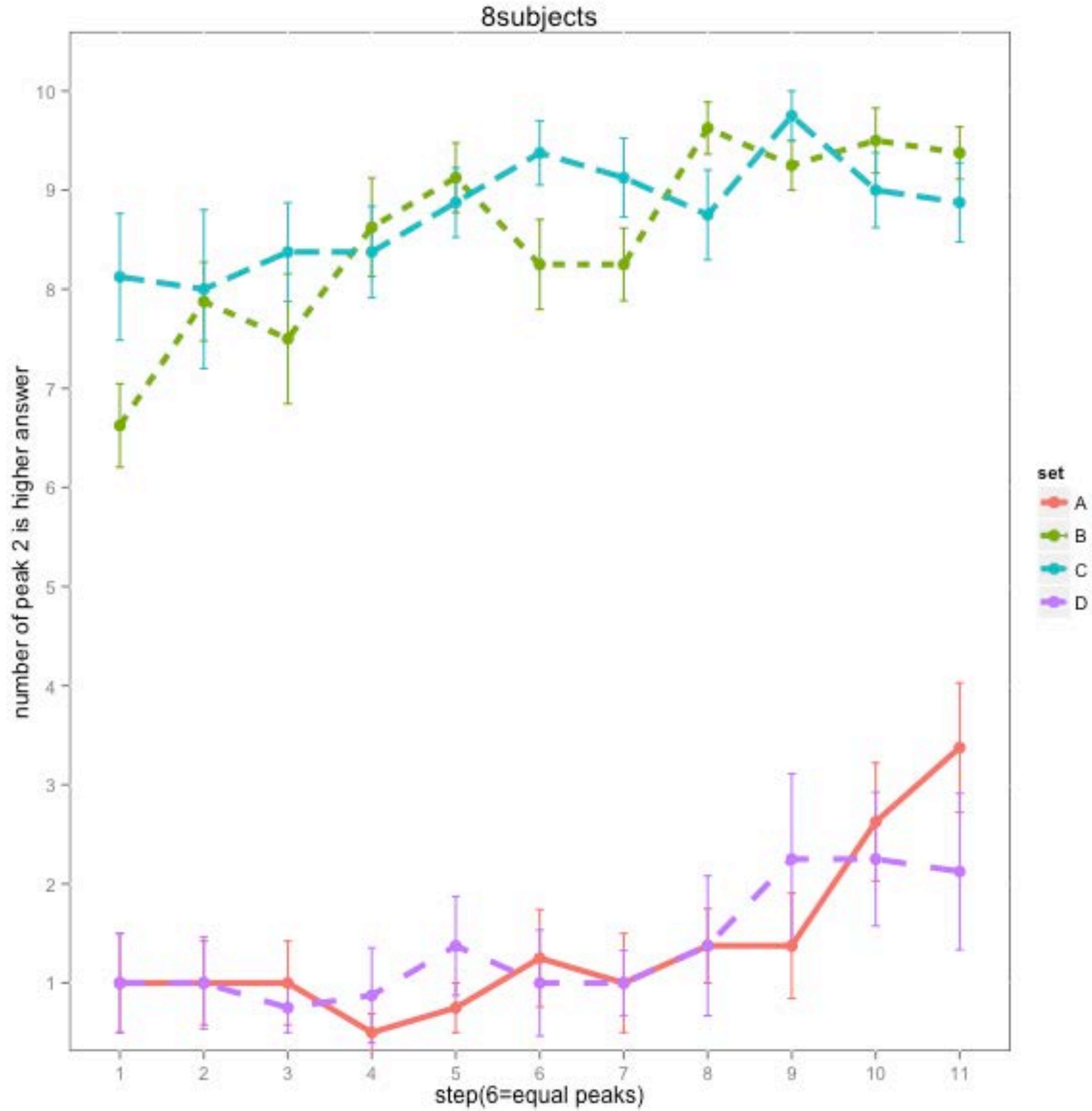
Attending only to F0 –

7 subjects:



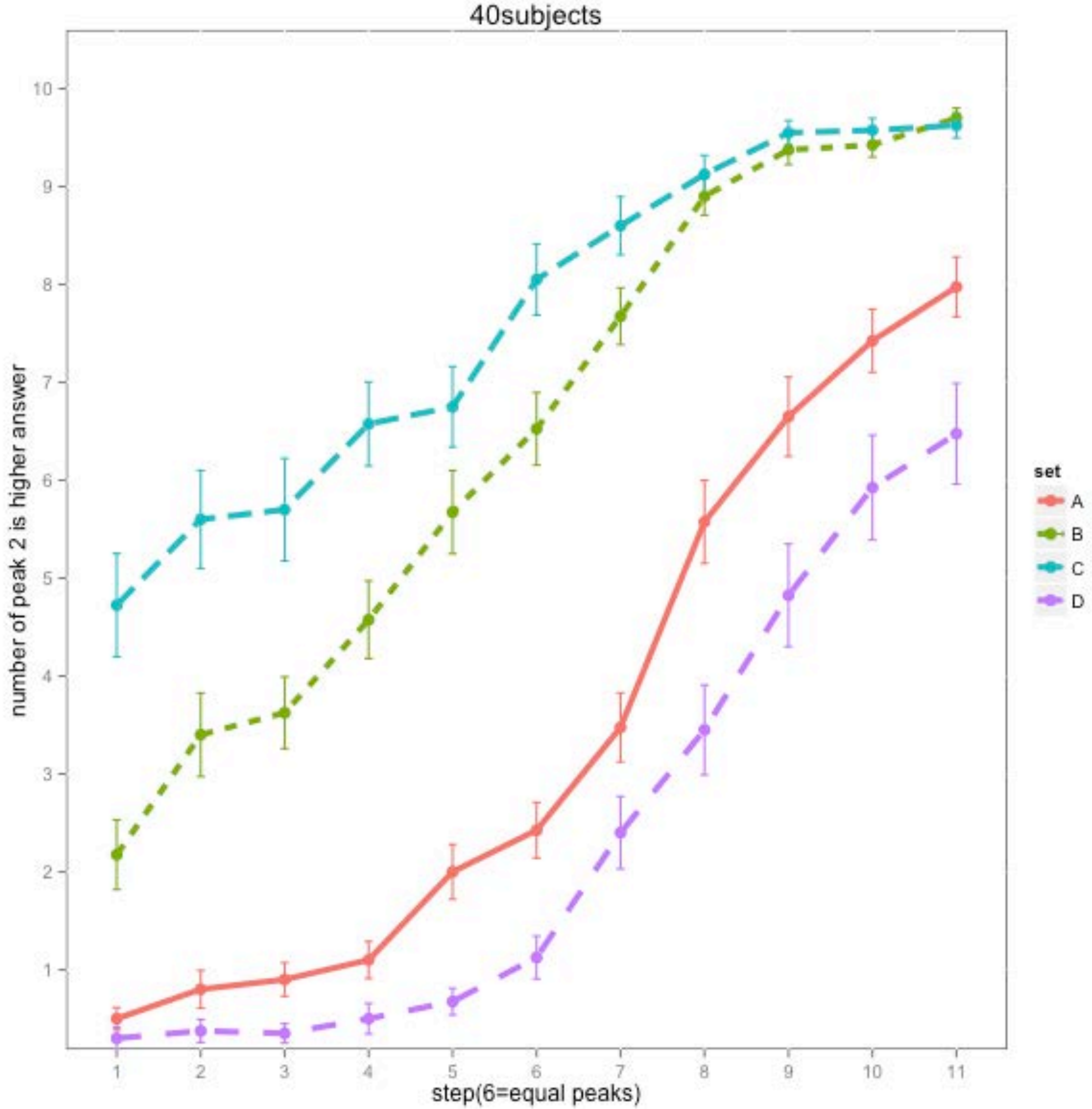
Attending only to timbre—

8 subjects:



Attending to both F0 & timbre—

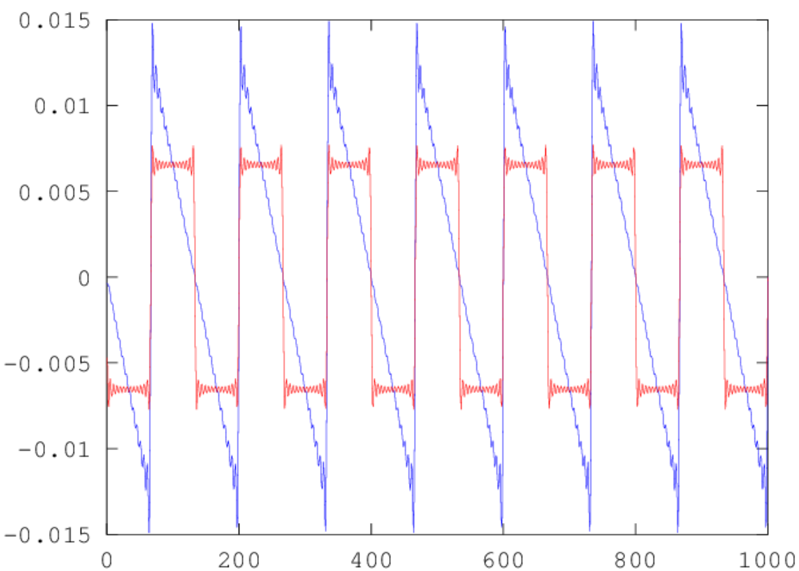
40 subjects:

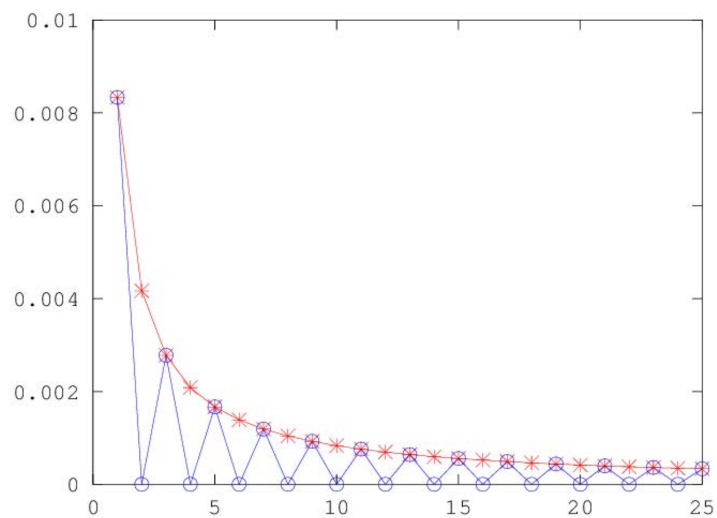


Another simple demonstration of the psychological interaction of pitch and timbre:

S1 is a 750-msec tone at 120 Hz, made of 25 overtones with 1/F amplitudes.
S2 is the same, except that overtones at 2*F0, 4*F0, 6*F0, etc., are missing.

So S1 is a sawtooth and S2 is a square wave:





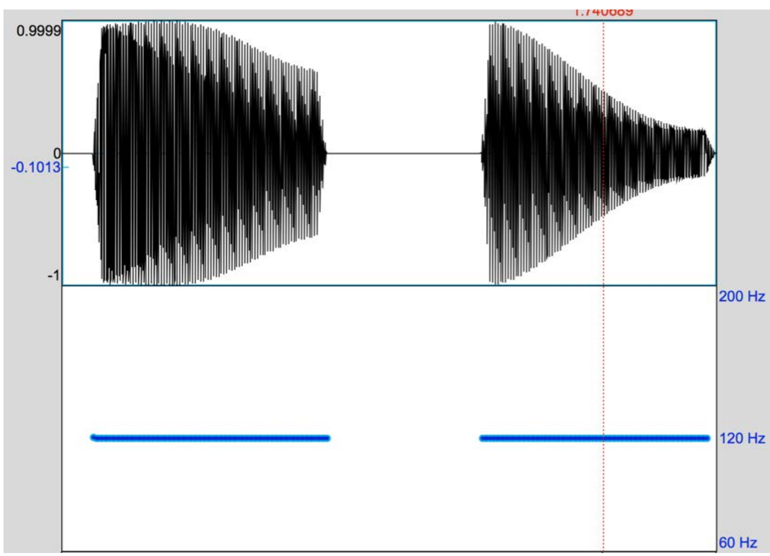
“F0” is the same in both cases – but S2 tends to sound an octave higher…

Now we make S1-2 as a blend of S1 and S2,
starting with 100% S1 and ending with 100% S2.

And S2-1 is a similar blend in the opposite order,
starting with 100% S2 and ending with 100% S1.

(Also added: onset and offset ramps, and an overall falling envelope…)

S1-2: S2-1:



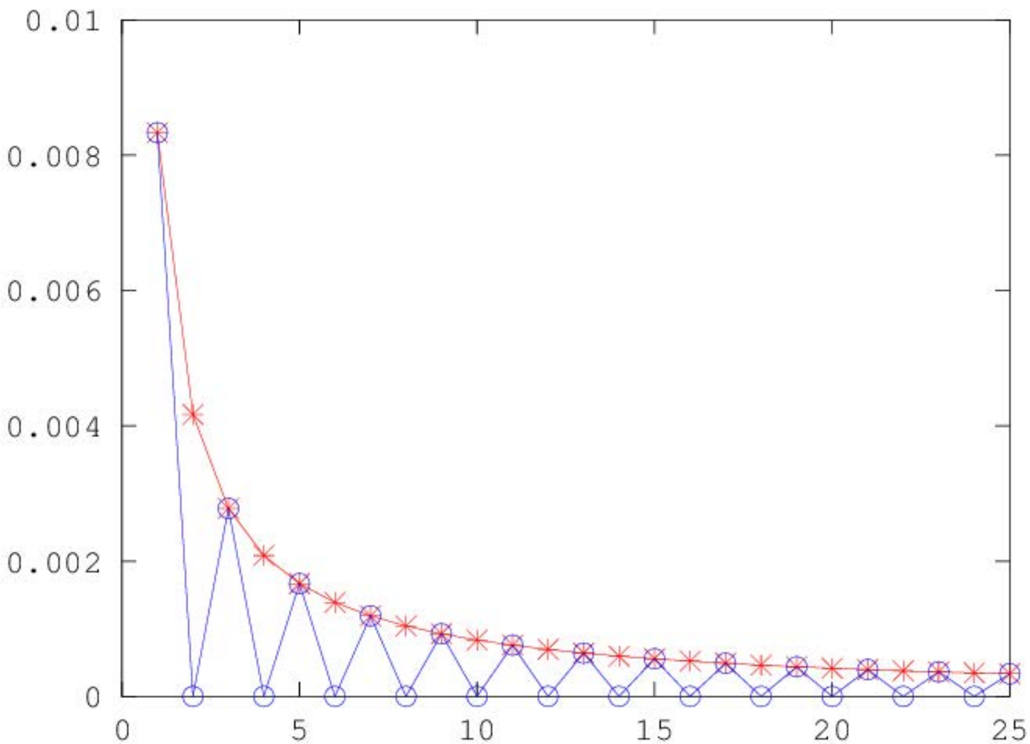
Praat thinks the pitch is constant –

… because the F0 is indeed constant!

Why does this happen?

With missing even harmonics, there's a paradox:
The spacing between the harmonics is 2*F0 –
but all the expected overtones of such a fundamental, n*2*F0, are absent.

Also, omitting all the even harmonics, given a 1/F spectrum,
tilts the overall spectral balance towards higher frequencies.



SO:

Tone is not (just) pitch,

And pitch is not (just) F0 --
either in production
or in perception

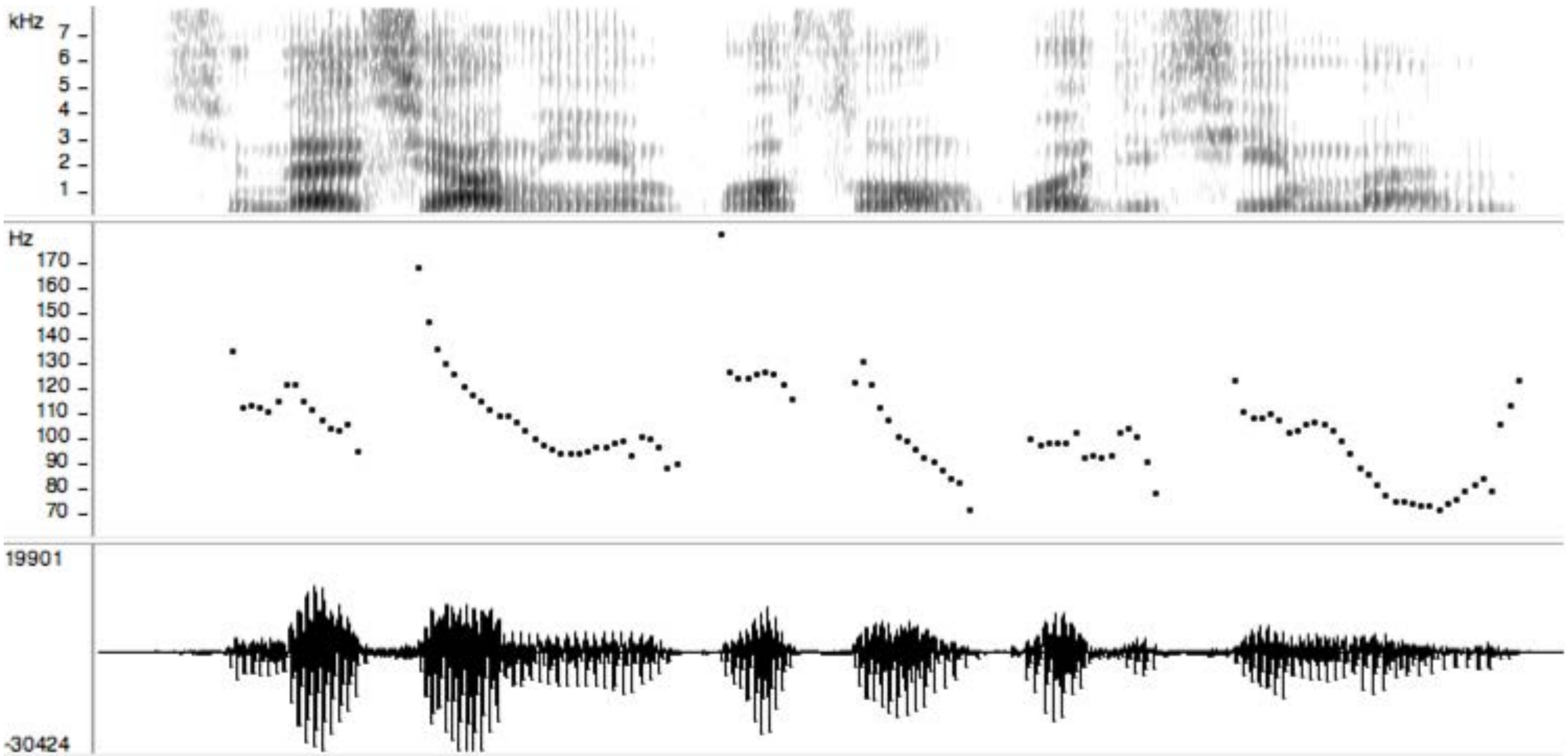
Maybe F0 is like formants:
A useful low-dimensional proxy for a phonetic concept --
But not a genuinely plausible model
(in physical, psychological, or practical terms)

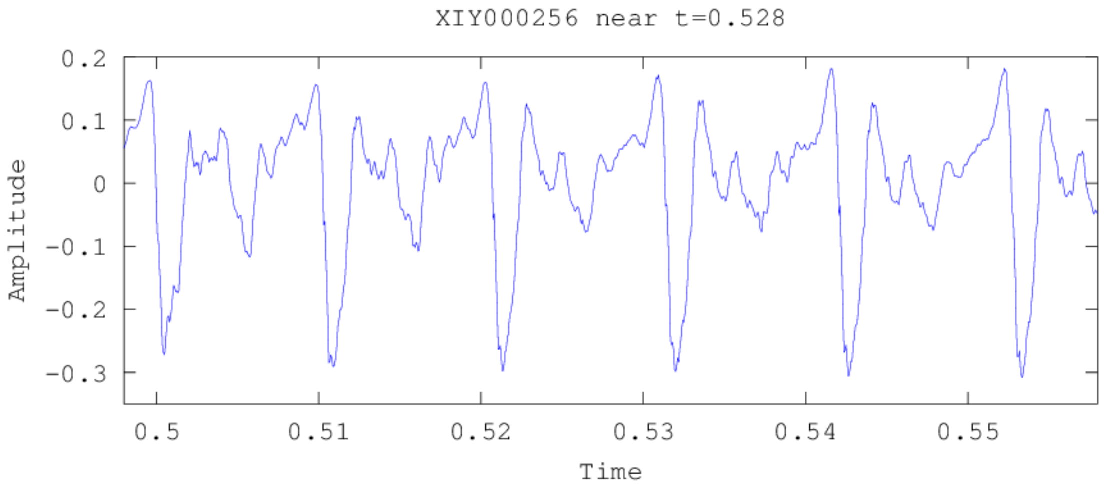
And now…

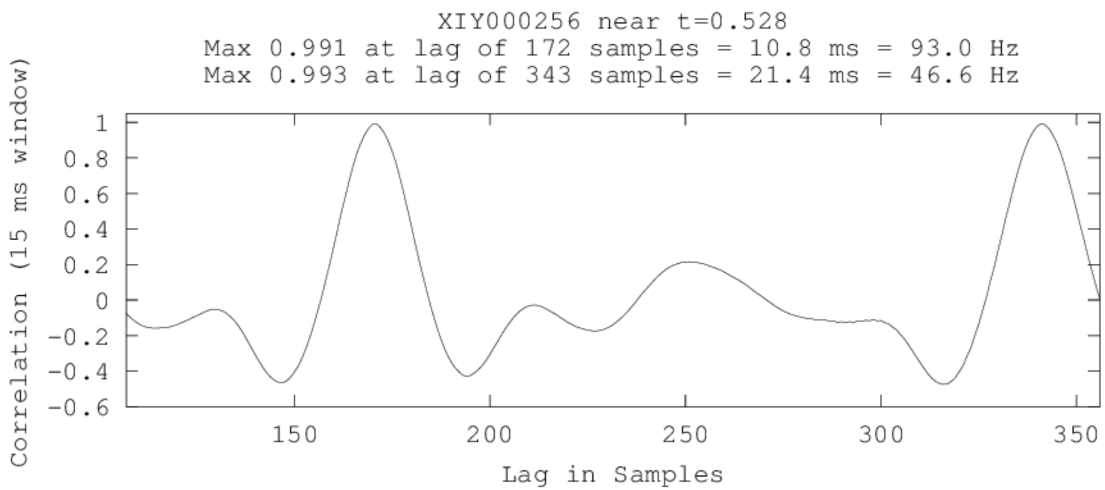
?

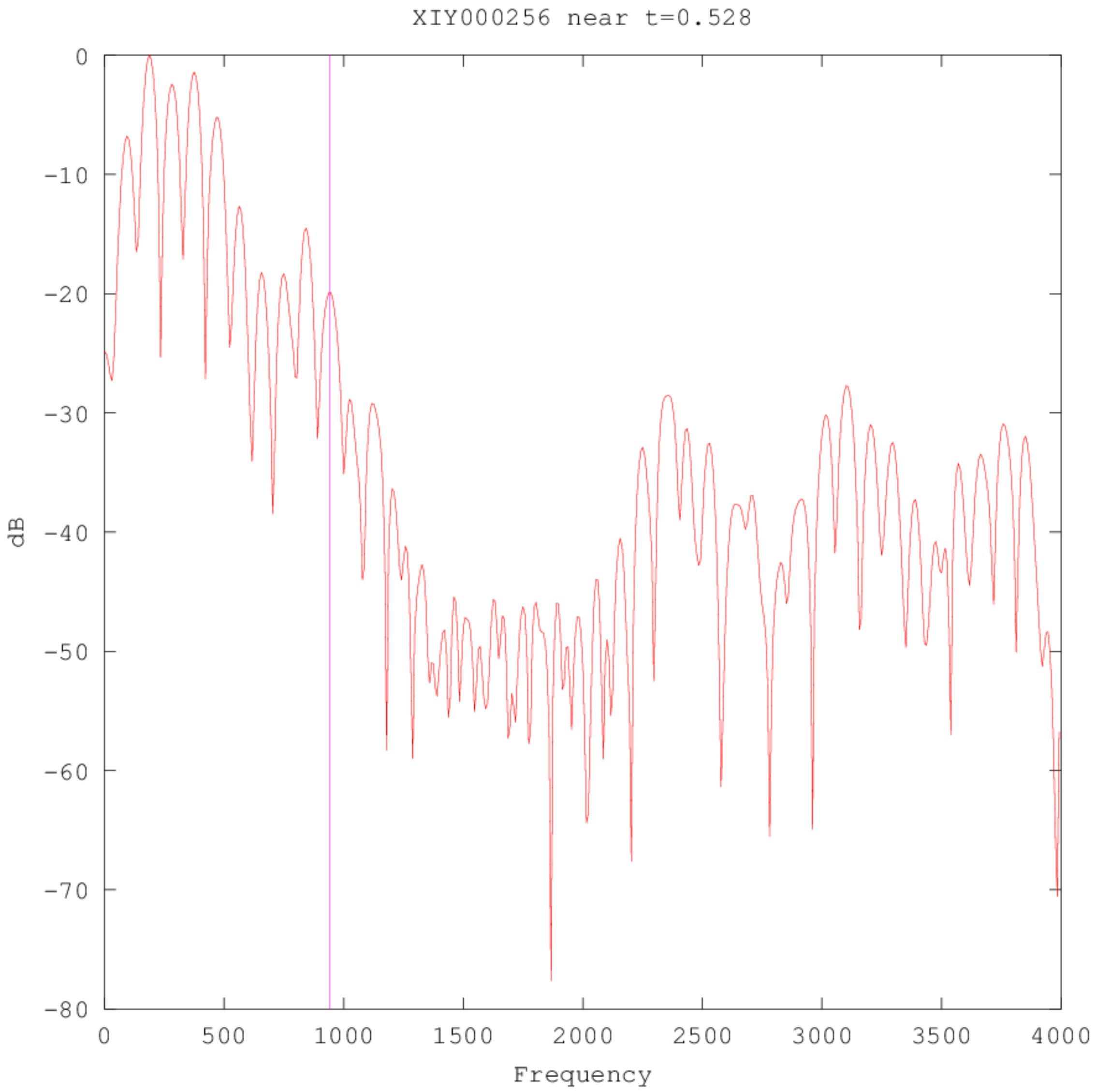
Dennis Klatt, "Speech Perception: A model of acoustic-phonetic analysis and lexical access", *Journal of Phonetics* 1979:

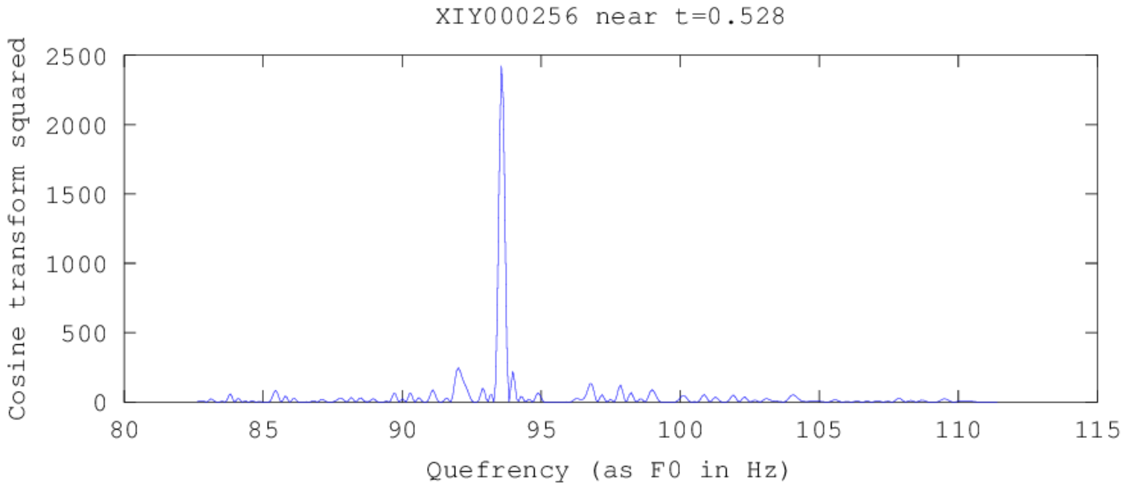
The role of formant frequencies No mention has been made of formant frequencies and formant motions as possible cues for phonetic perception. In the acoustic theory of speech production, formant frequencies play a central role, characterizing the natural resonant modes of the vocal tract for a given articulatory configuration (Fant, 1960). However, automatic extraction of formant frequency information from the speech waveform is a difficult engineering task. It is still tacitly assumed by many that formant frequencies are psychologically real dimensions employed in perceptual decoding strategies (Delattre *et al*., 1955; Carlson, Fant & Granstrom, 1975). We have no perceptual data that would refute this assumption, but there are several reasons to question its plausibility. For example, occasional formant tracking errors should result in dramatic errors in phonetic perception, whereas observed phonetic errors demonstrate a strong tendency to be acoustically similar to the intended vowels and consonants (see, e.g. Miller & Nicely, 1955). As we have argued, absolute decisions at any level below the word (parametric representation, phonetic feature representation, or segmental representation) should be avoided if at all possible for optimal lexical decoding.


kHz
7
6
5
4
3
2
1
Hz
170
160
150
140
130
120
110
100
90
80
70
19901
-30424


XIY000256 near t=0.528
Amplitude
Time
0.2
0.1
0
-0.1
-0.2
-0.3
0.5
0.51
0.52
0.53
0.54
0.55


XIY000256 near t=0.528
Max 0.991 at lag of 172 samples = 10.8 ms = 93.0 Hz
Max 0.993 at lag of 343 samples = 21.4 ms = 46.6 Hz
Correlation (15 ms window)
Lag in Samples
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
150
200
250
300
350


XIY000256 near t=0.528
0
-10
-20
-30
-40
-50
-60
-70
-80
dB
0
500
1000
1500
2000
2500
3000
3500
4000
Frequency


XIY000256 near t=0.528
Cosine transform squared
2500
2000
1500
1000
500
0
80
85
90
95
100
105
110
115
Quefrency (as F0 in Hz)