Bootstrap distributions

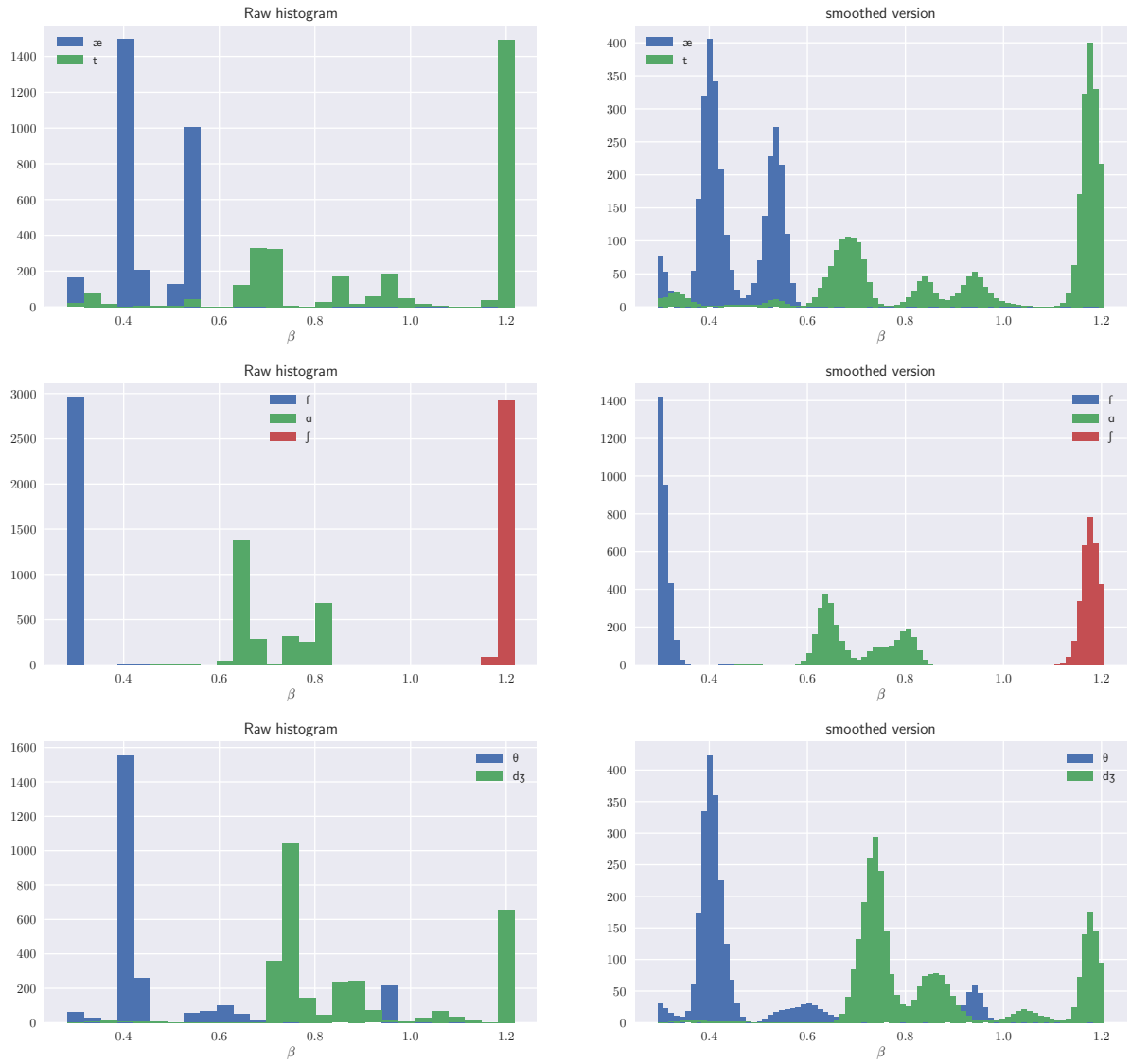


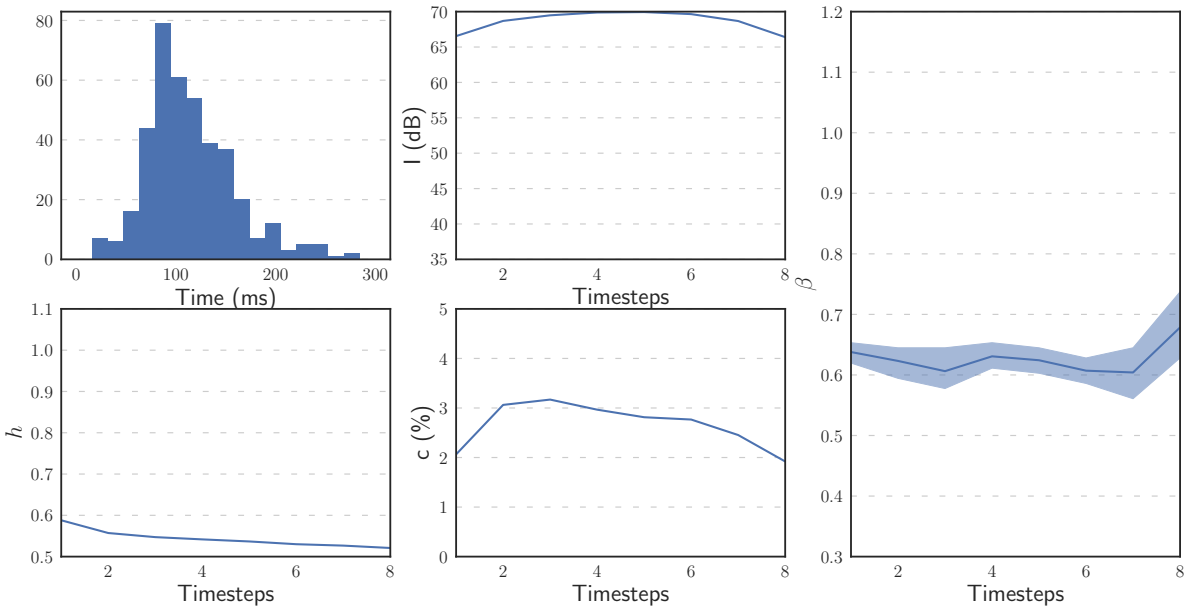
Figure: Bootstrap distributions for phonemes æ,t (up), f, ɑ, ʃ (middle), θ, and dʒ (bottom). The procedure is the so-called smoothed percentile bootstrap. $\beta^\star$ is computed after resampling with repetition of 400 slices from the TIMIT database (this is done 3000 times). This results in a first histogram (left) that is smoothed with a Gaussian filter ($\sigma = 0.015$) to obtain the final bootstrap distribution (right). The mean value and the bootstrap confidence intervals are then extracted from the distribution (ex: to get the 70% CI, 15% of the samples on the left and right are excluded).

Temporal evolutions of β

Top left: Duration histograms for the occurrences retrieved from the TIMIT database (absolute time), **top center:** intensity I (dB) vs time, **bottom left:** entropy score h vs time, **bottom center:** contrast c vs time, **right:** exponent β vs time.

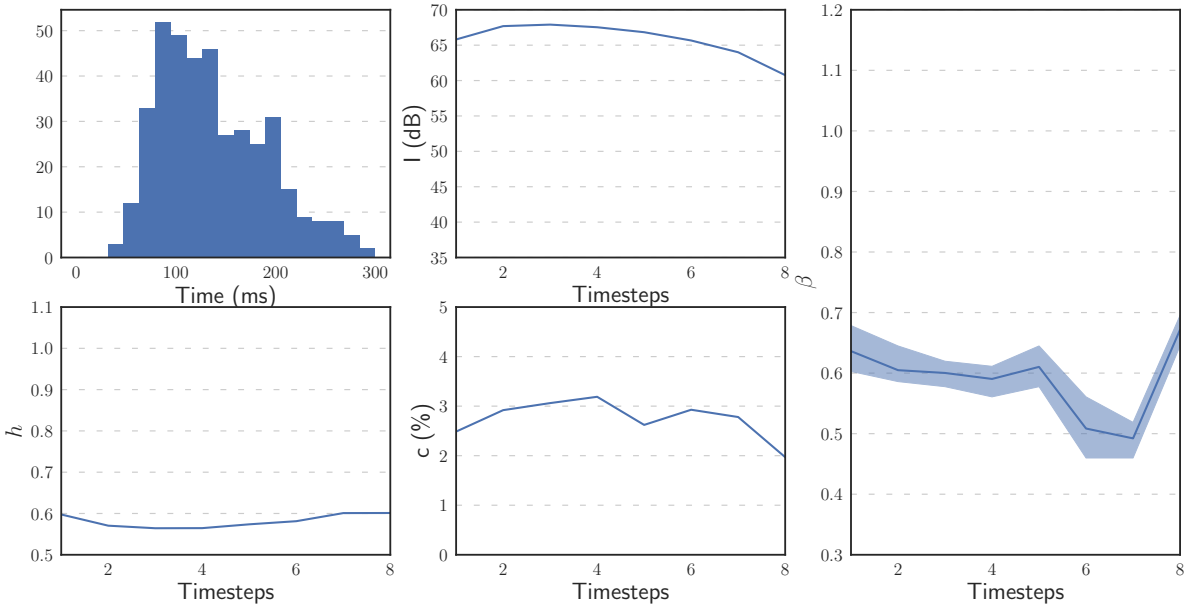
Vowels and Diphtongs

ɑ



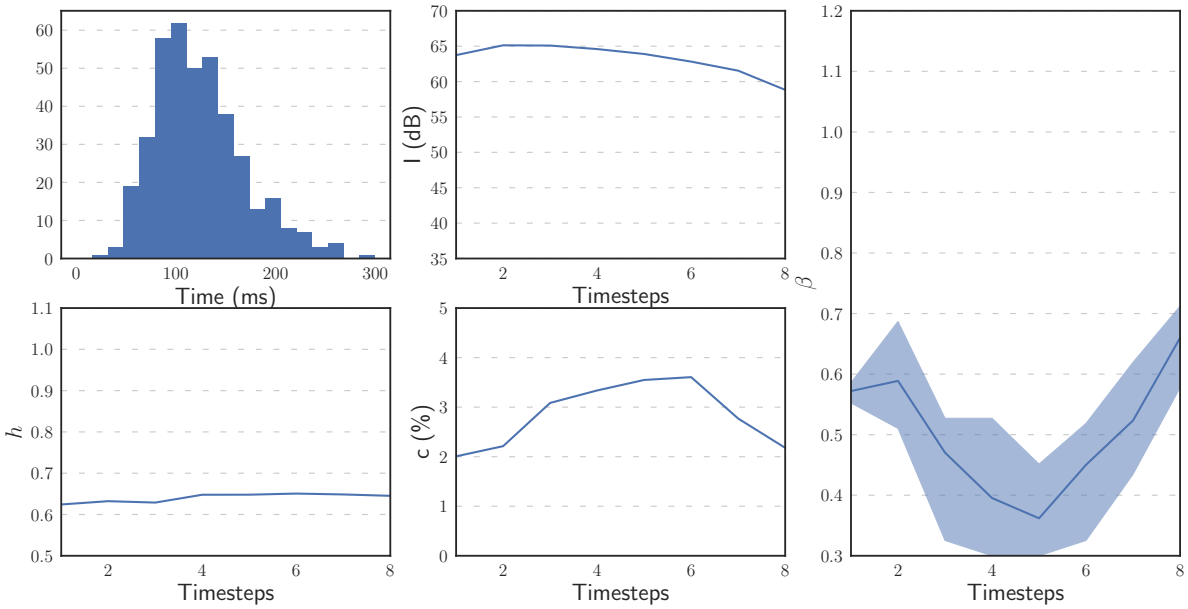
mean duration : 116.1 ms (std : 49.9 ms)

aɪ



mean duration : 139.1 ms (std : 56.6 ms)

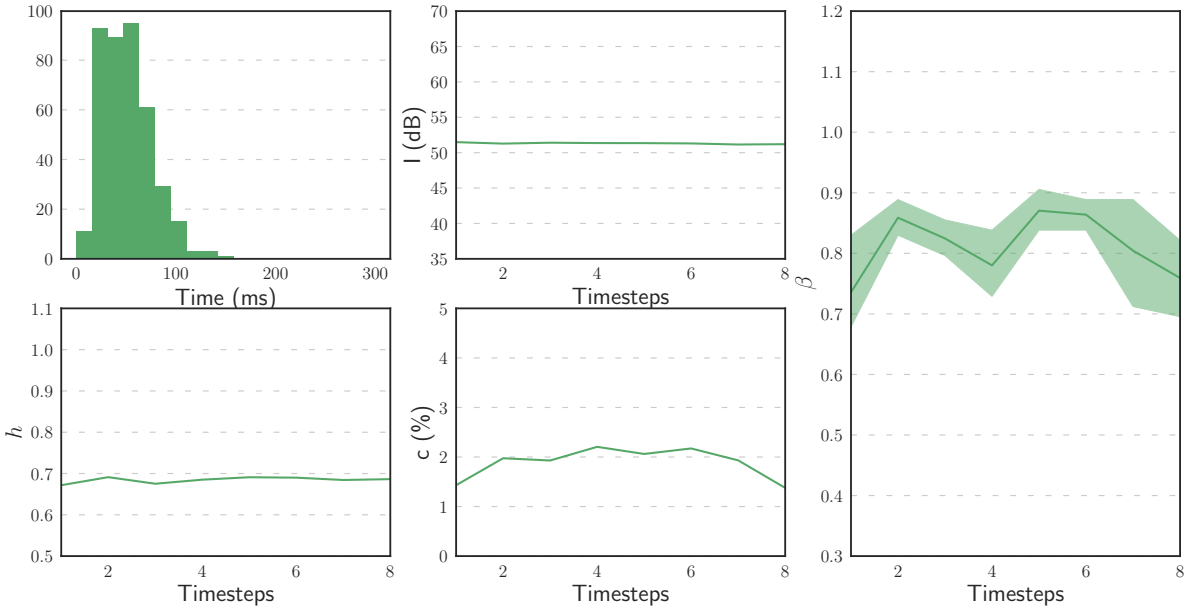
eɪ



mean duration : 126.7 ms (std : 50.8 ms)

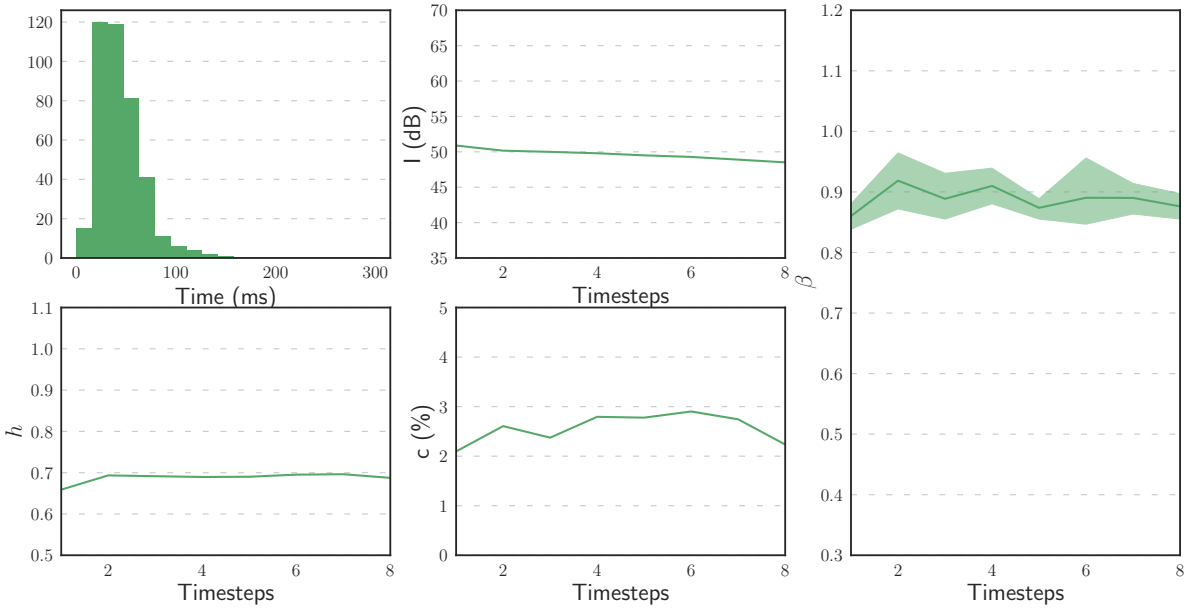
Nasals

m



mean duration : 50.9 ms (std : 24.6 ms)

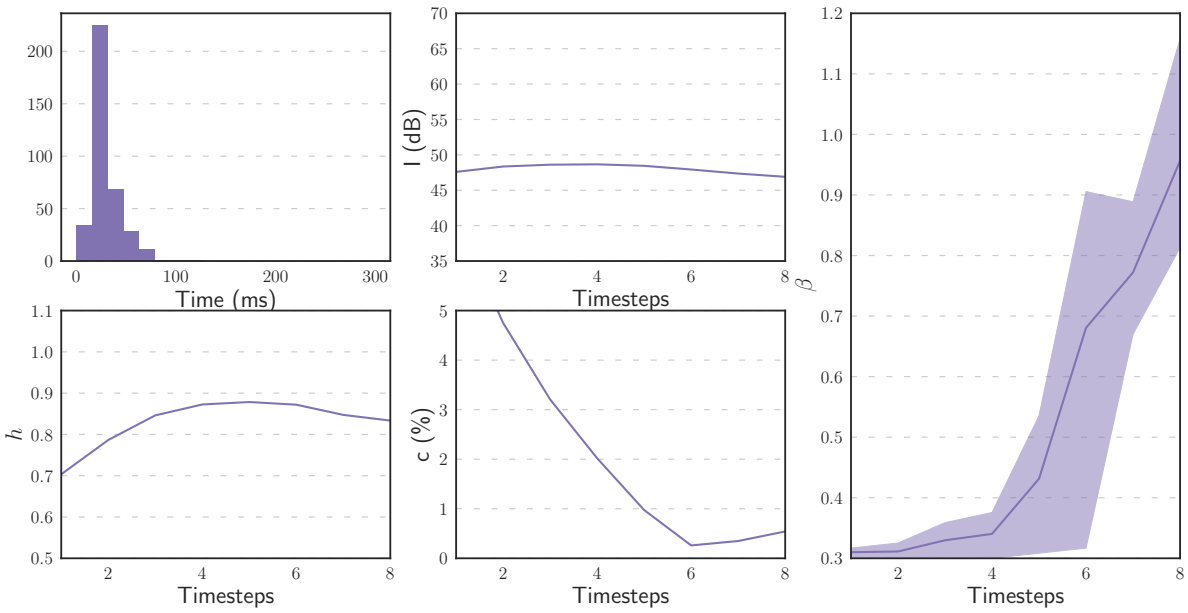
n



mean duration : 44.0 ms (std : 22.2 ms)

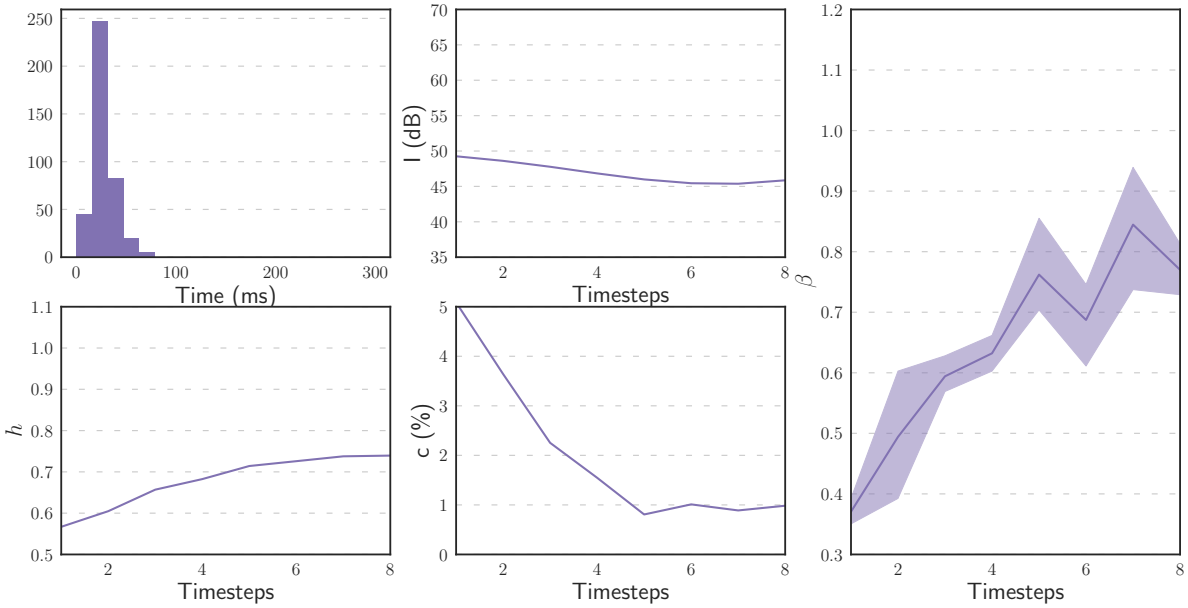
Stops

d



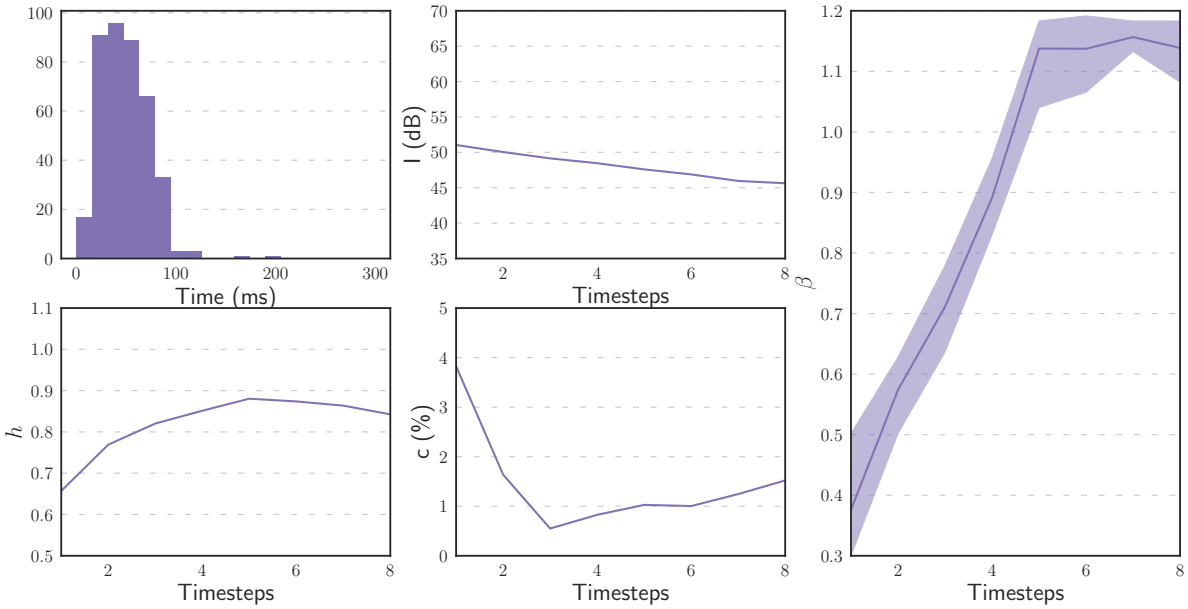
mean duration : 28.4 ms (std : 14.9 ms)

g



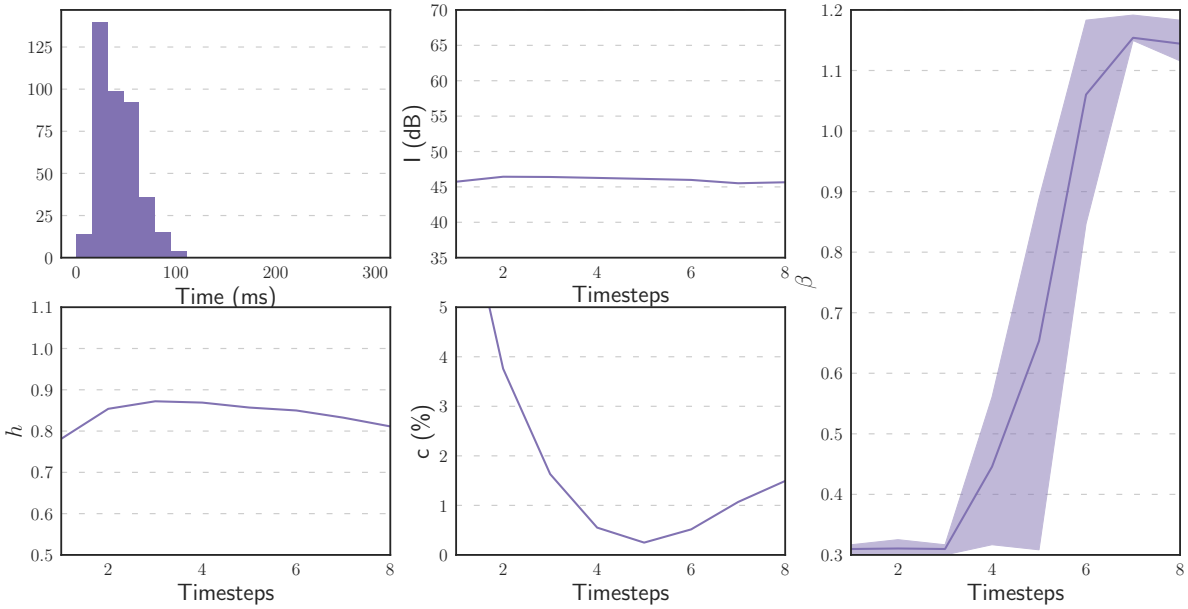
mean duration : 26.5 ms (std : 11.1 ms)

k



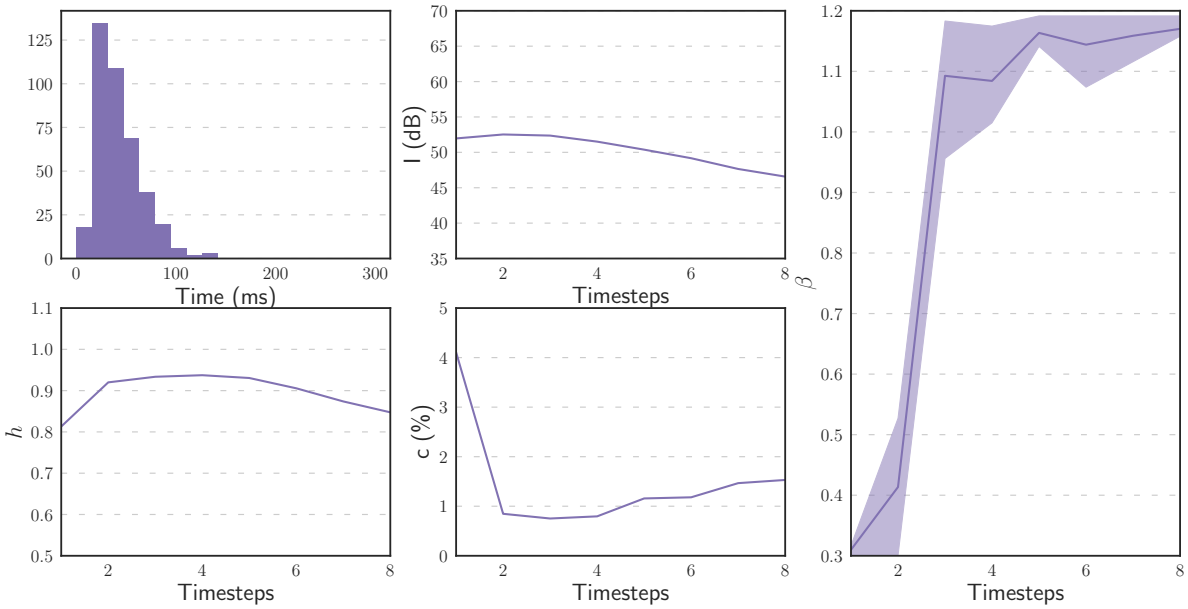
mean duration : 48.9 ms (std : 23.8 ms)

p



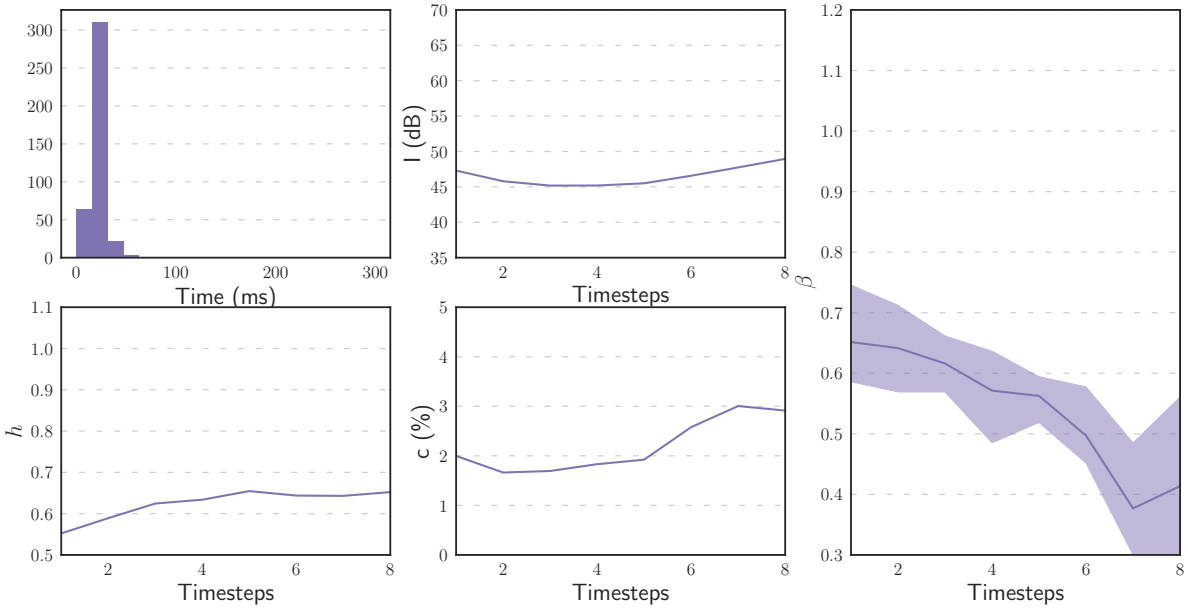
mean duration : 41.9 ms (std : 19.4 ms)

t



mean duration : 42.9 ms (std : 22.3 ms)

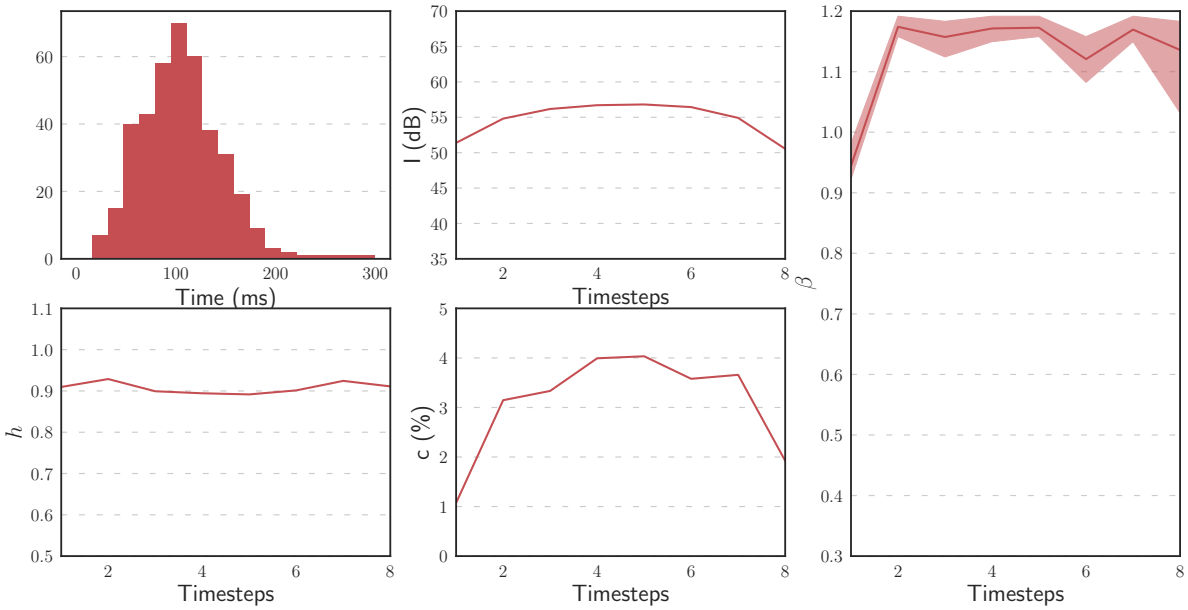
r



mean duration : 21.1 ms (std : 6.6 ms)

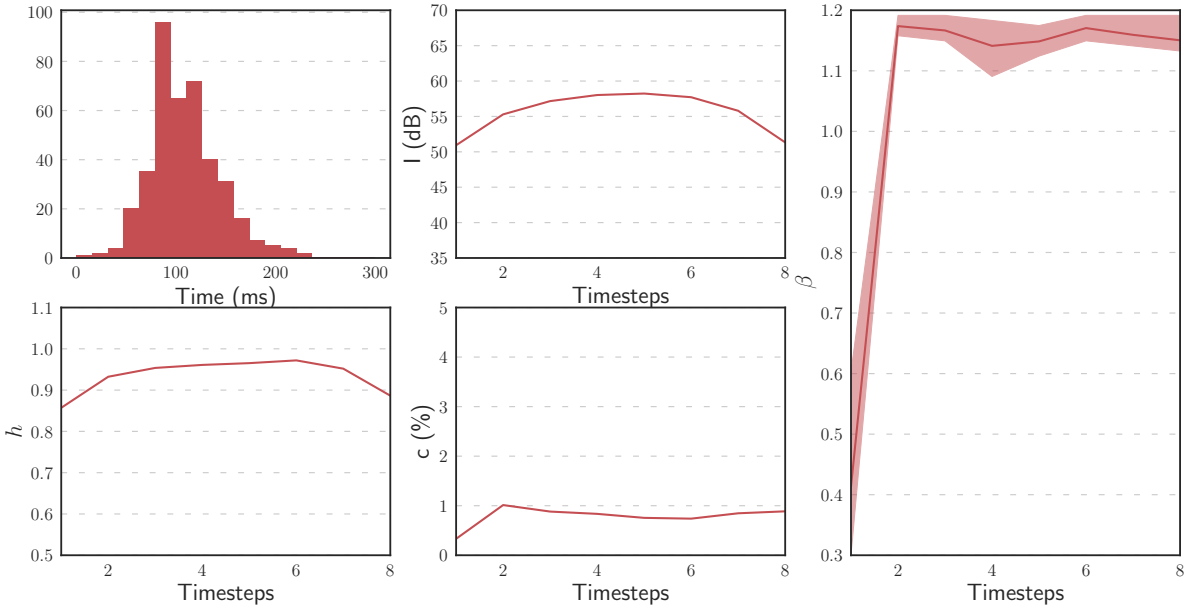
Fricatives

s



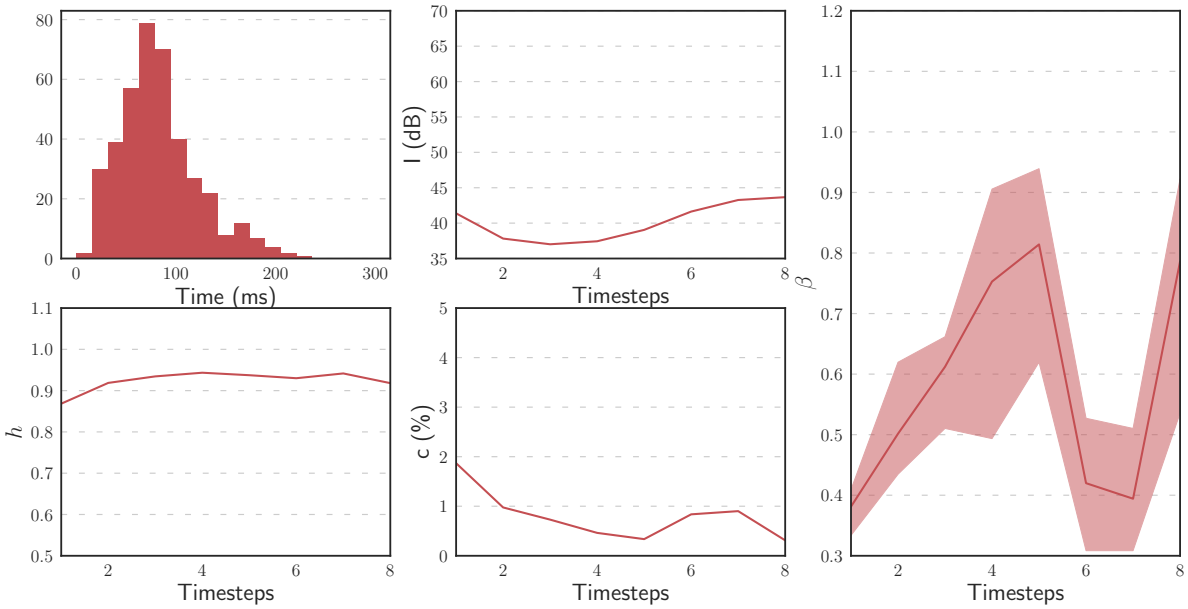
mean duration : 105.8 ms (std : 41.2 ms)

ʃ



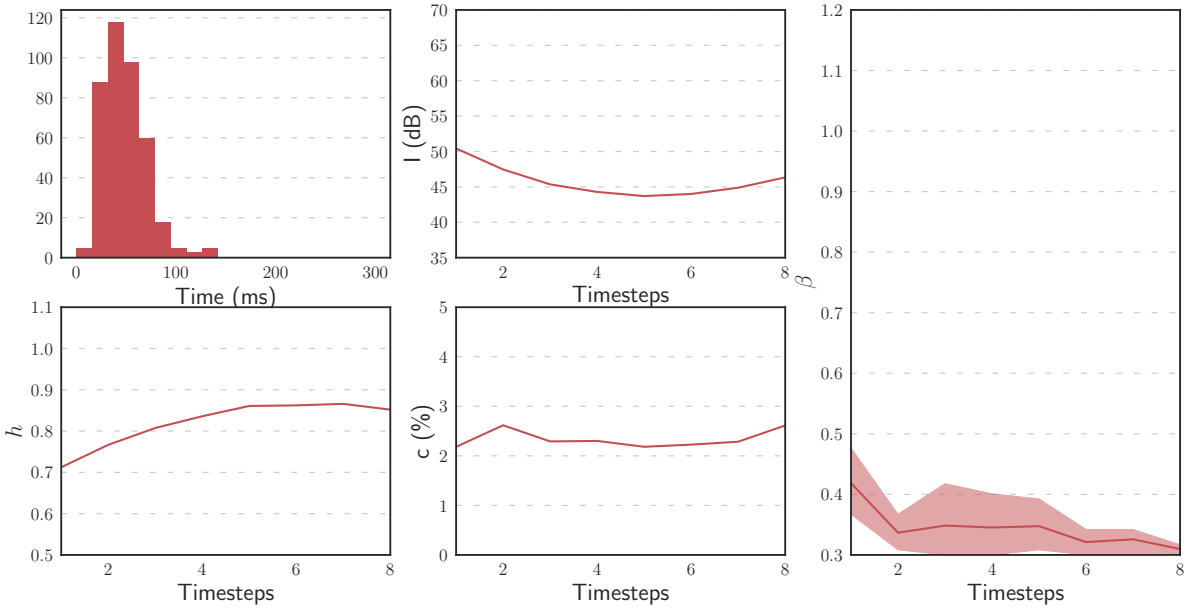
mean duration : 108.8 ms (std : 34.1 ms)

θ



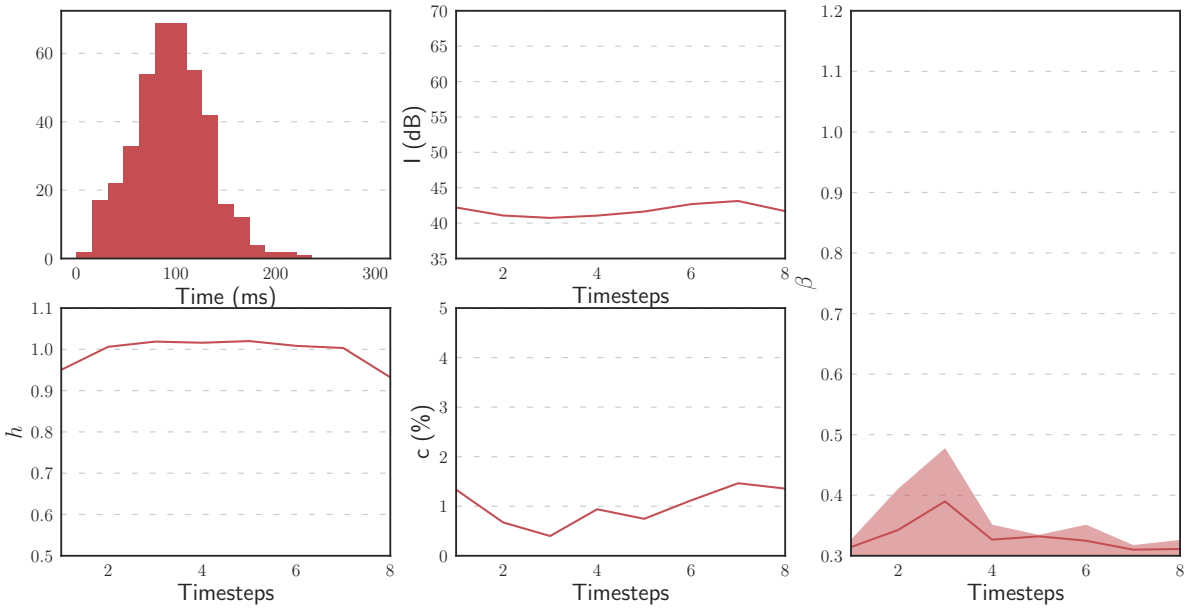
mean duration : 83.4 ms (std : 39.7 ms)

v



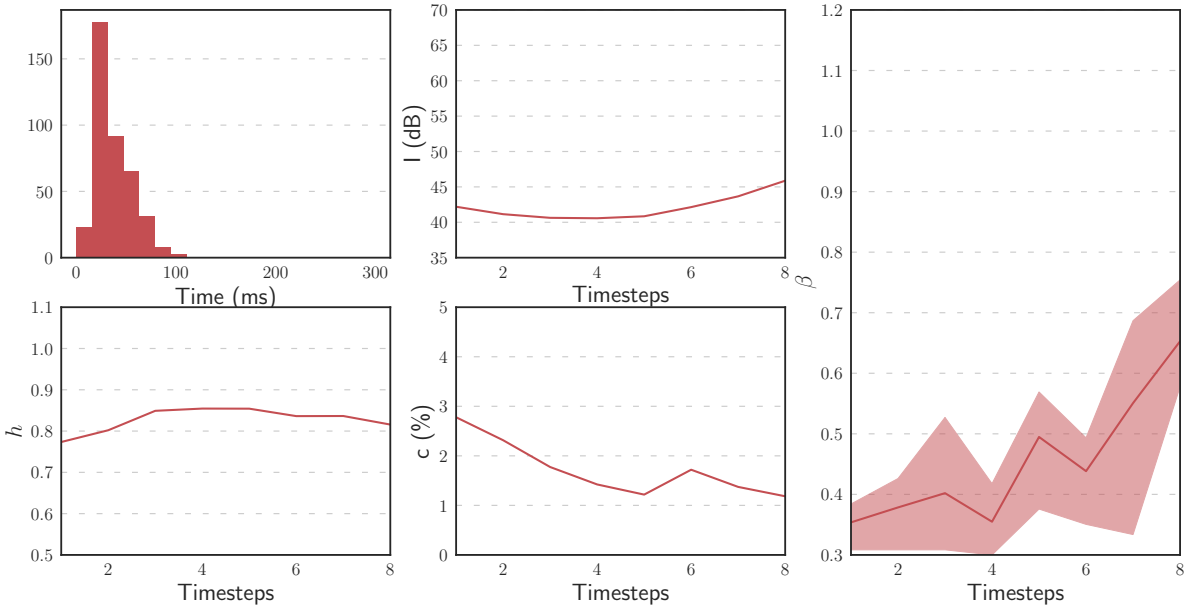
mean duration : 49.1 ms (std : 21.9 ms)

f



mean duration : 95.8 ms (std : 37.3 ms)

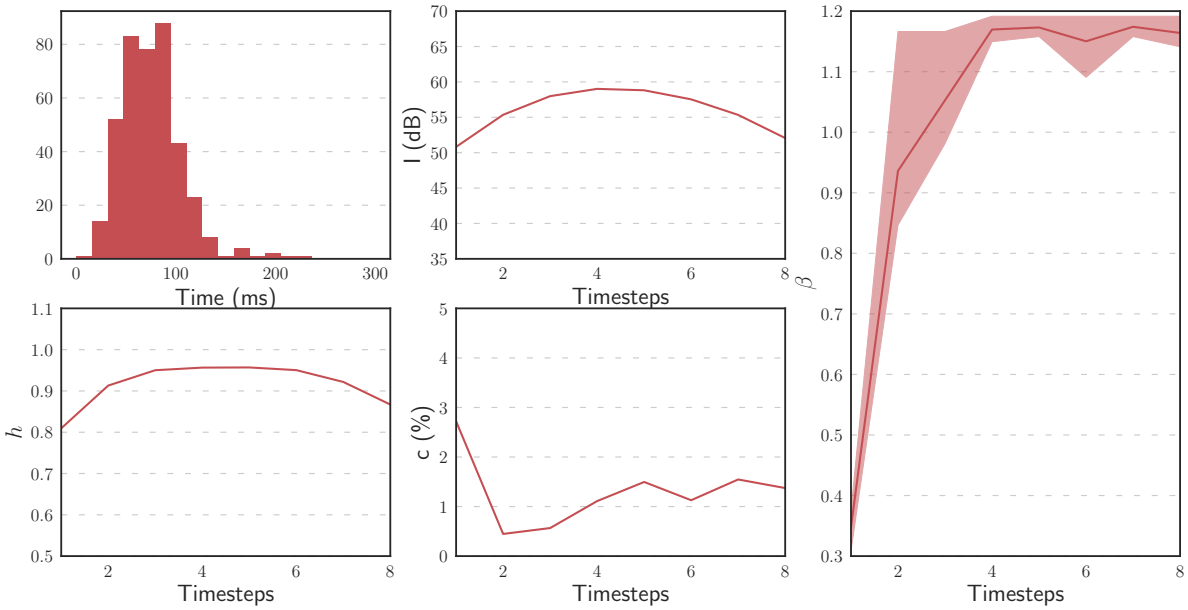
ð



mean duration : 36.4 ms (std : 18.4 ms)

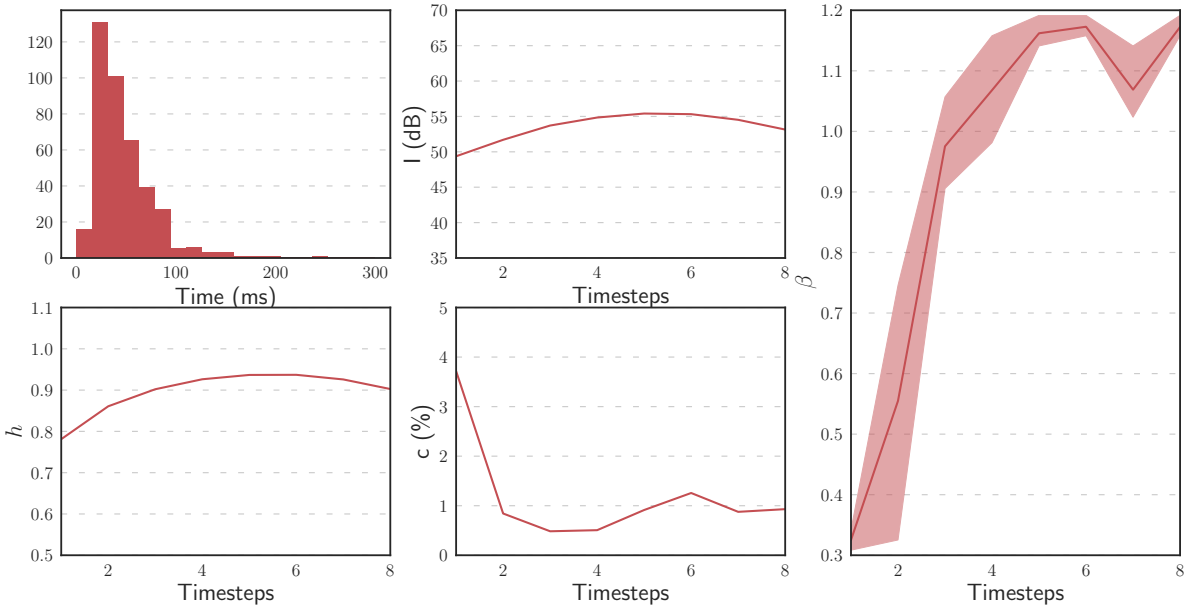
Affricates

tʃ



mean duration : 75.1 ms (std : 30.4 ms)

dʒ



mean duration : 47.0 ms (std : 29.7 ms)

β as a function of intensity level for synthetic vowel-like sounds

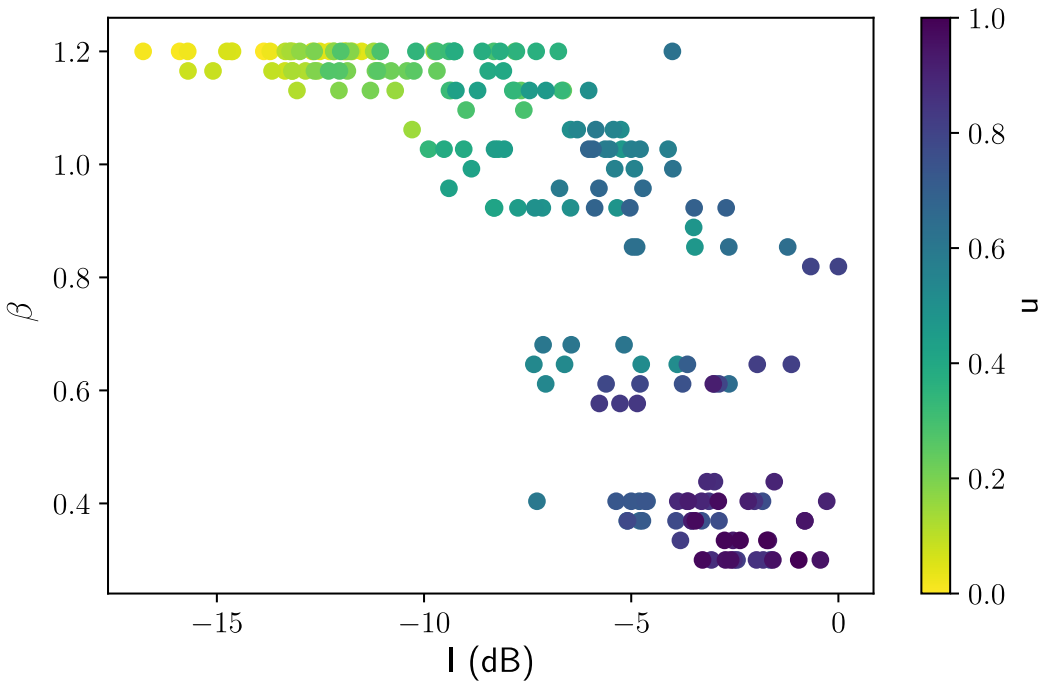


Figure: Scatter plot of simulated samples on the (I, β) plane: exponent β against intensity I in dB (ref:max). Each point is a sample of Simulation 2 on synthesized vowels. The parameter u controls linearly the aperture of the cylindrical waveguide, from $r = 0.2cm$ $(u = 0)$ to $r = 1.3cm$ $(u = 1)$.

Distribution of American English phonemes in the (β, h) plane for the different weighting strategies

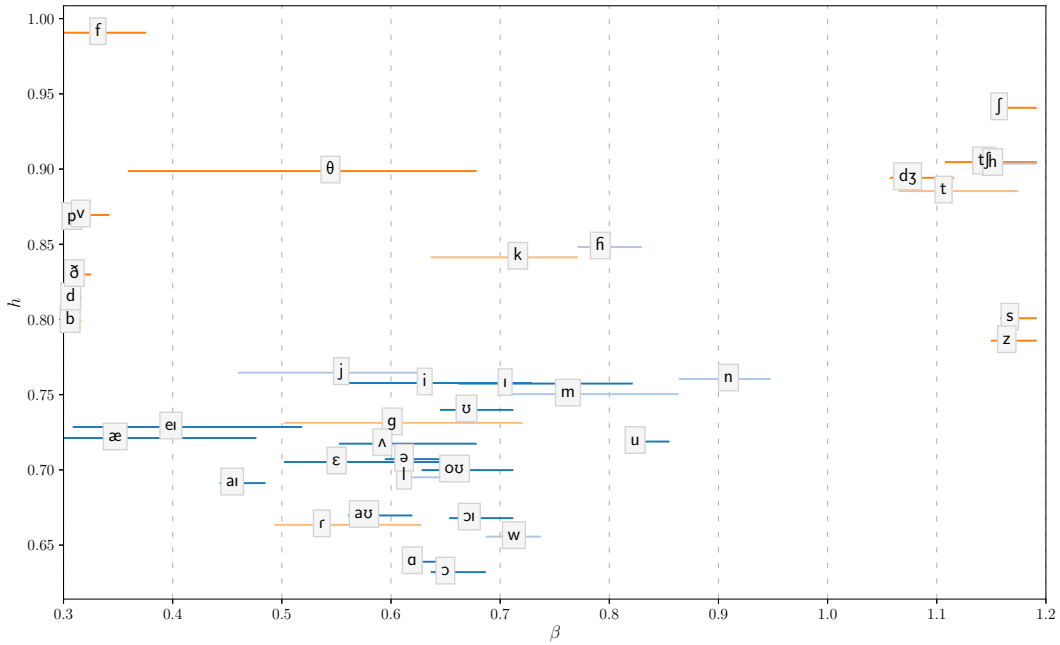


Figure 1: Distribution of American English phonemes in the (β, h) plane with *Strategy A*: raw scores (no spectral whitening). Labels are positioned on bootstrap distribution averages, lines represent 70% bootstrap confidence intervals. Bootstrap distributions are based on 400 occurrences for each phoneme and 3000 repetitions. Not represented: r, ɝ, ɚ ($\beta = 0.93, h = 0.58$).

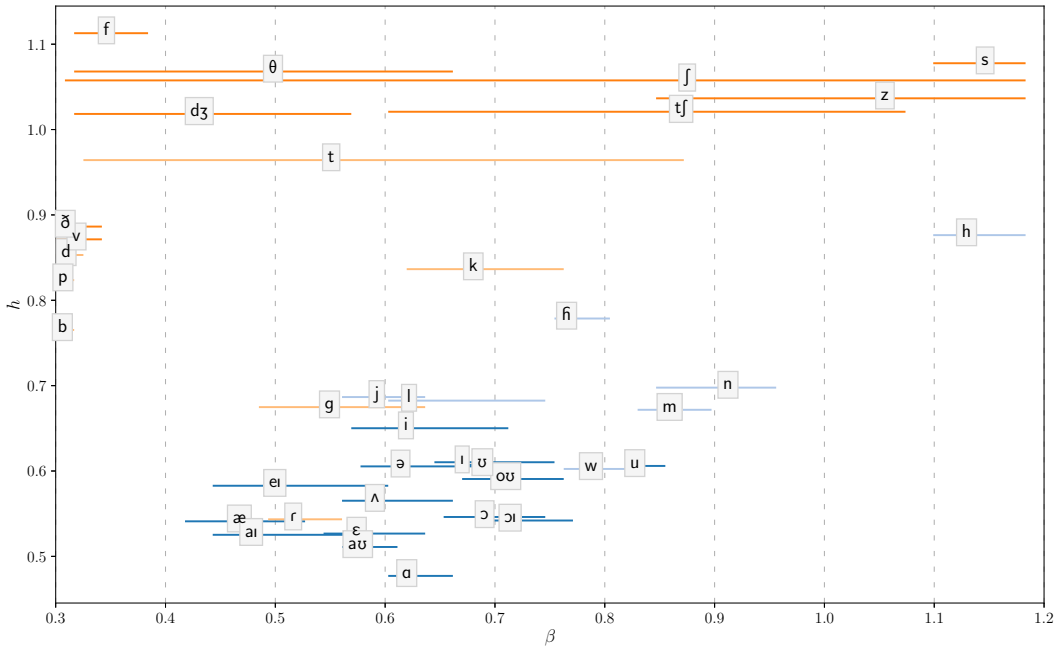


Figure 2: Distribution of American English phonemes in the (β, h) plane with *Strategy B*: spectral whitening +5dB/octave on weights. Labels are positioned on bootstrap distribution averages, lines represent 70% bootstrap confidence intervals. Bootstrap distributions are based on 400 occurrences for each phoneme and 3000 repetitions.
Not represented: r ($\beta = 0.86, h = 0.37$), ɝ($\beta = 0.91, h = 0.36$), ɚ ($\beta = 0.84, h = 0.37$).

Figure with *Strategy C* (slighter gain +2.5dB/octave): in article.

Analyses of vowels and nasals using Independent Component Analysis

Objective

This complementary analysis aims to clarify the behavior of Q_{10} as a function of f_c for single phonemes (vowels and nasals). This analysis is based on Independent Component Analysis (ICA). A simultaneous objective is to verify the consistency of the results obtained with the parametric method, described in the main text.

Methods

The experiments repeated the procedure described in Stilp and Lewicki, 2013, for the study of broad phonetic categories. The algorithm and the estimation of the quality factors Q_{10} follow the same path, and are not described in details. The main difference between the two experiments are in the estimation of the Q_{10}/f_c slope, as the lines were constrained to cross a fixed point.

- Inputs: Inputs were concatenated occurrences of the same phoneme, retrieved from the TIMIT database by selecting sentences at random. Each occurrence was windowed by a Tukey window ($\alpha = 0.5$). The files made up of the concatenated samples lasted 1 minute each (1 file corresponding to 1 phoneme).
- Experiments: ICA was conducted 7 times on each file with different initializations. The input vectors were 8-ms slices (dimension: 128 at 16 kHz) extracted at random from the input files. At the beginning of each optimization process, the 128x128 matrix W was a random mixture of time/frequency patterns (impulsions, sinusoids), and noise. The gradient descent was done with approximately 10 000 steps.
- **Estimation of β :** As in the previous work, the β parameter is the regression slope of Q_{10} against center frequency f_c for the filters learned (on a *log-log* scale). The estimation, however, differed here from the previous work as the regression line was constrained to cross the point ($f_0 = 1.0\,\text{kHz}, Q_0 = 2.0$). This choice has been made to be consistent with the parametric representation described in the main text. Therefore, the β parameter was estimated using the following formula :

$$\beta_{ICA} = \frac{\mathbb{E}\left(\log(Q_{10}/Q_0)\log(f_c/f_0)\right)}{\mathbb{E}(\log(f_c/f_0)^2)} \ .$$

  The values that are presented in the Results sections are averaged over the 7 experiments.

Results

Figure 1 shows the comparison between the estimates of β using the parametric method (described in the main text) or ICA, on vowels and nasals. Figure 2 and 3 shows the comparison between the two estimates row by row, for vowels and nasals taken apart. These figures also display the bootstrap confidence intervals (parametric method) and the standard deviations on the 7 simulations (ICA). We added the plots of Q_{10} as a function of f_c for most phonemes at the end of the document.

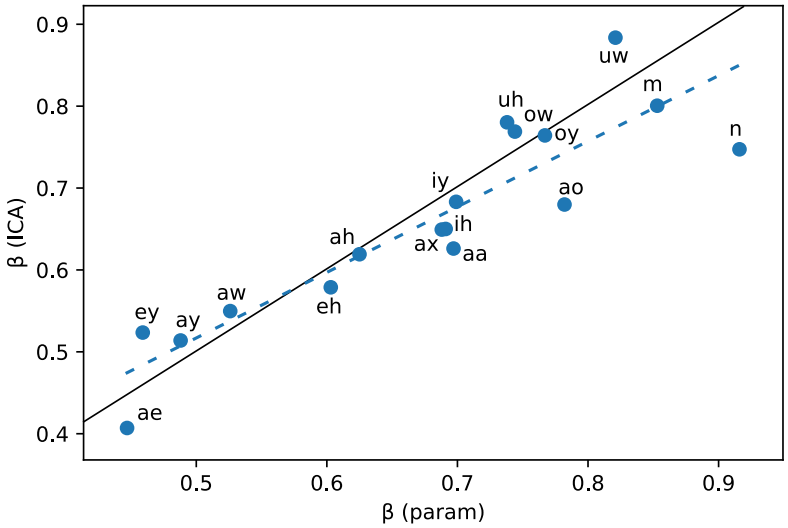


Figure 1: Estimates of β using the parametric method (on the x-axis) and ICA (on the y-axis). The phonemes are indicated with the ARPABET notation. The dark line is the diagonal and the blue line is the linear regression connecting the points (slope:0.80, intercept:0.12). Standard deviation from the diagonal : 0.058, from the linear regression: 0.052. The correlation coefficient is r=0.90.

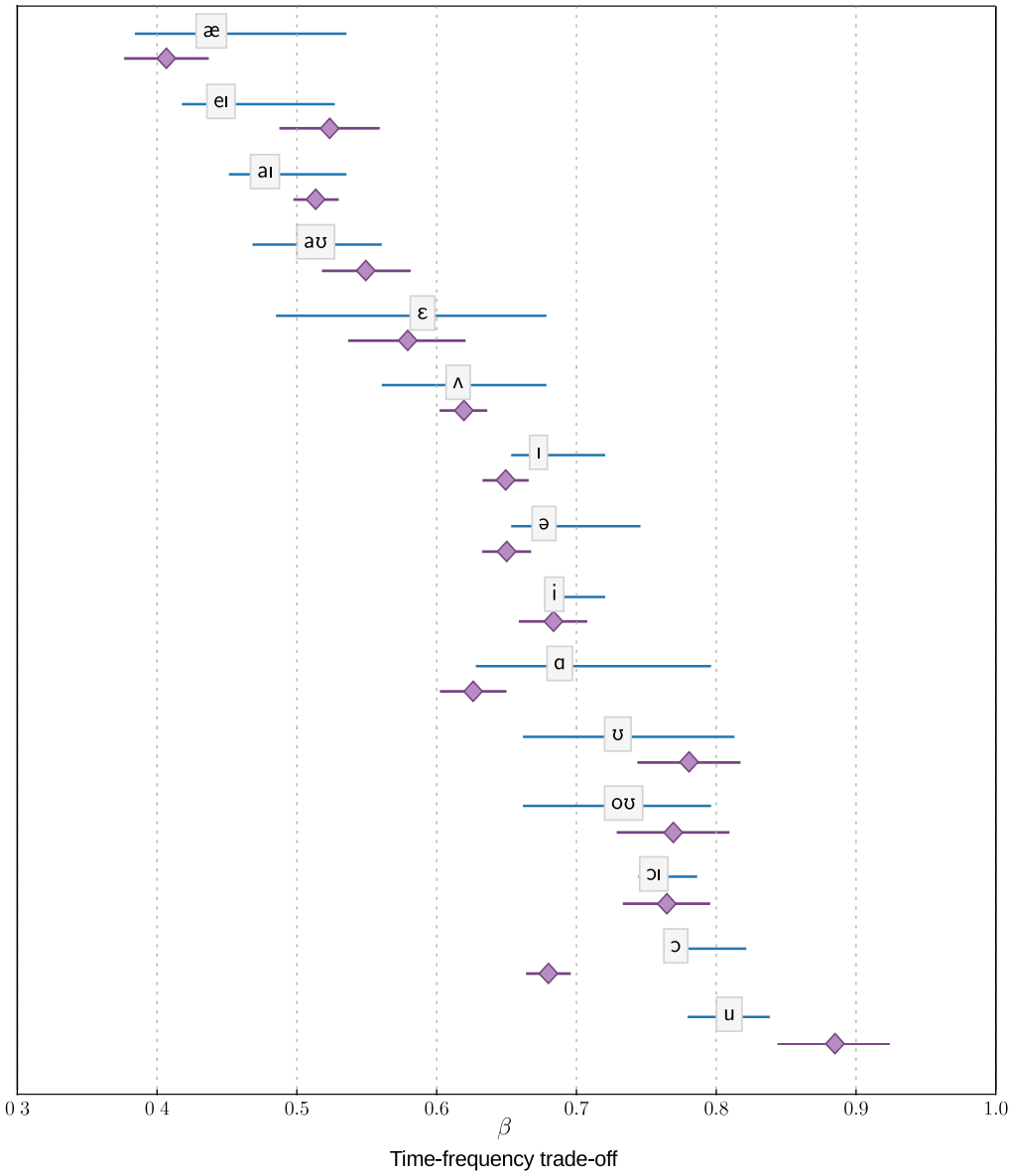


Figure 2: Estimates of β using the parametric method (labels) and ICA (diamonds) for vowels. The diagram shows the 70% bootstrap confidence intervals (parametric method) and +/- standard deviations computed on 7 simulations (ICA).

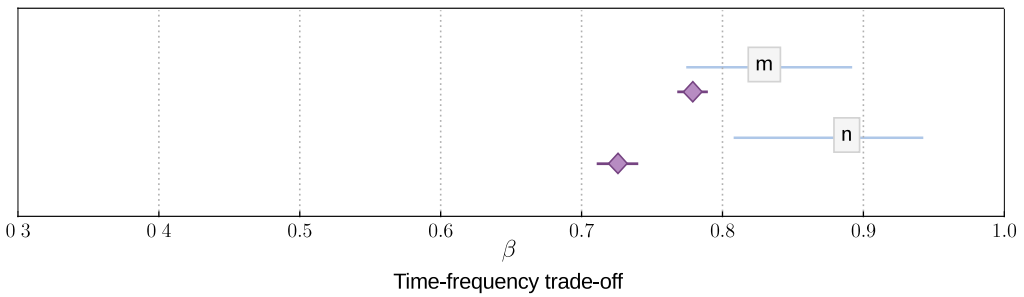


Figure 3: Estimates of β using the parametric method (labels) and ICA (diamonds) for nasals.

Discussion

The comparison between ICA and the parametric method shows the overall consistency of the β estimates and confirm the main conclusions on the statistical structure of vowels and nasals, detailed in the main text. The largest differences are found for the phonemes [n], [ɔ] (gap greater than 0.1), followed by [ɑ], [ei], [u][1]. All the other phonemes exhibit differences of less than 0.6.

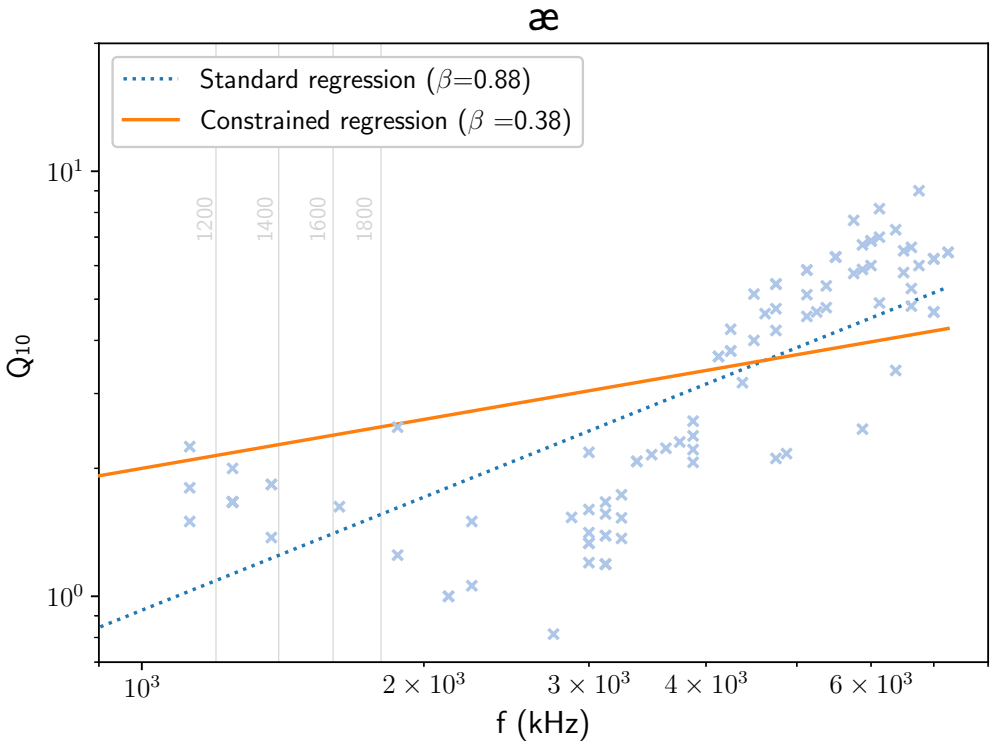
The figures of the quality factors as a function of frequency, provided in the section below, offer additional insights. The main differences in the quality factors between vowels are shown to be in the region 1–5 kHz. In particular, mid-open vowels ([ɛ], [ʌ]), diphthongs (e.g. [ai]), but especially the sound [æ], can produce very low quality factors below 3 kHz. This fact accounts for a significant difference in behavior between this analysis and Stilp and Lewicki's analysis, potentially leading to opposing observations. The low quality factors at medium frequencies make the regression slope steeper in the unconstrained setting – as it is the case in Stilp and Lewicki's work. For instance, the diphthongs [ai] or [ei] are associated with high β values if the regression is unconstrained (~ 1), but with much lower values if constrained (~ 0.5). This fact also explains why the β values found in this analysis are generally lower than those previously reported for vowels and nasals. The figures reveal several local behaviors. The Q_{10} factors are higher in regions where formants are found, especially between 1 and 2 kHz. For example, the *schwa* [ə] and the related sound [ʌ] both present a "bump" at $f_c = 1.5$ kHz. This effect is also visible when comparing the plots between back vowels (e.g. [ɔ]), which have a low second formant (~ 1 kHz), and front vowels (e.g [i], [ɪ]), which have higher second formant (> 2 kHz). Another local behavior is the rise of Q_{10} near an antiresonance, particularly visible for the sound [m], close to $f_c = 3$ kHz (second antiresonance of the mouth cavity). This behavior, however, is not systematic (not apparent on [n]).

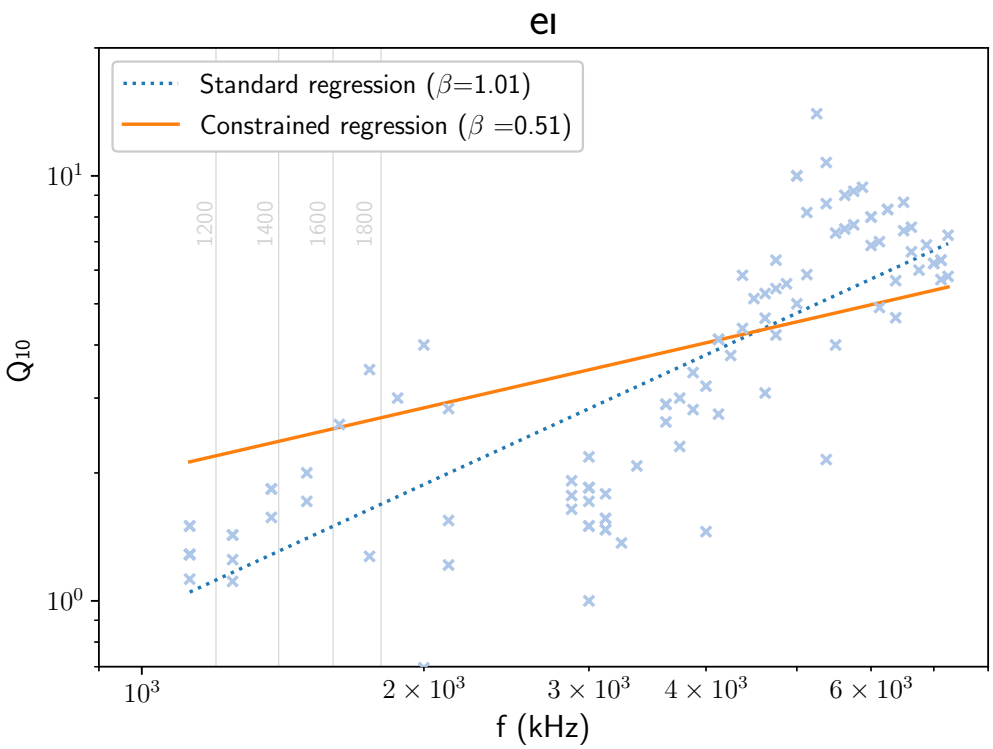
Additional figures

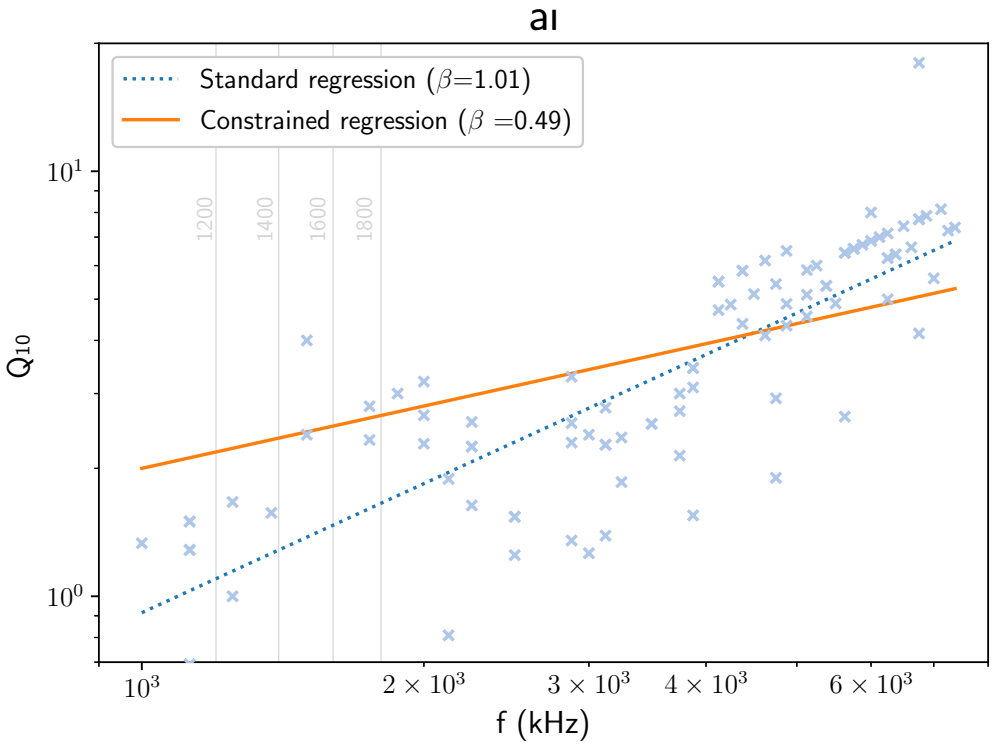
In this section we provide examples of plots of Q_{10} as a function of f_c. The selected examples were the ones that provided the β estimate closest to the average value (for the same phoneme).

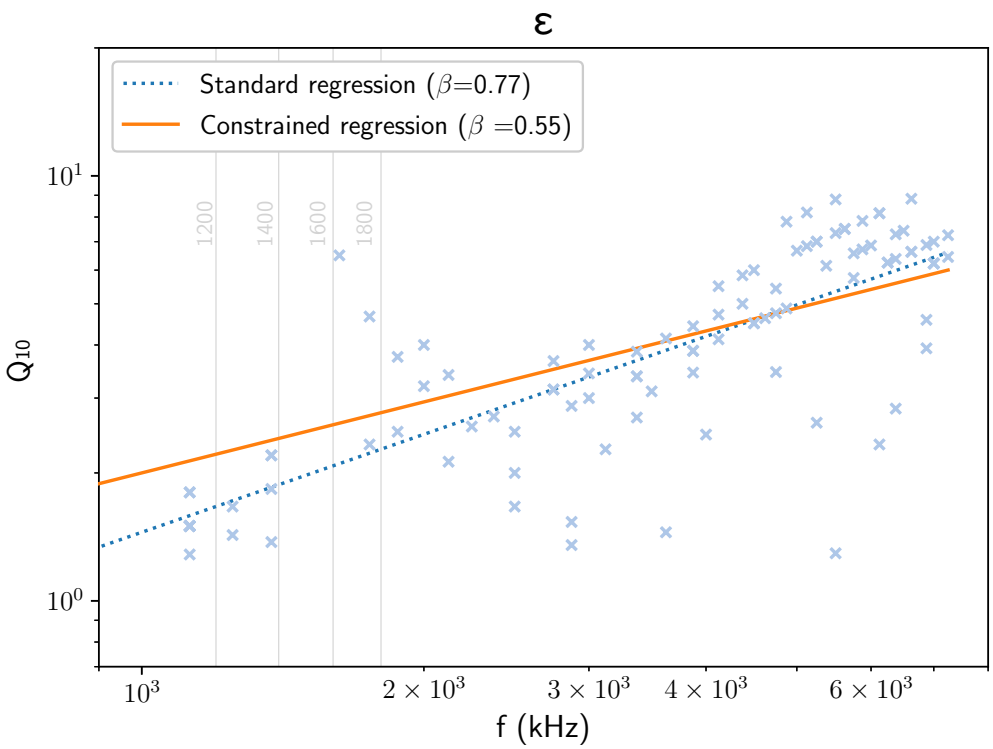
The figures show the regression lines both in the constrained and the unconstrained (standard regression) cases.

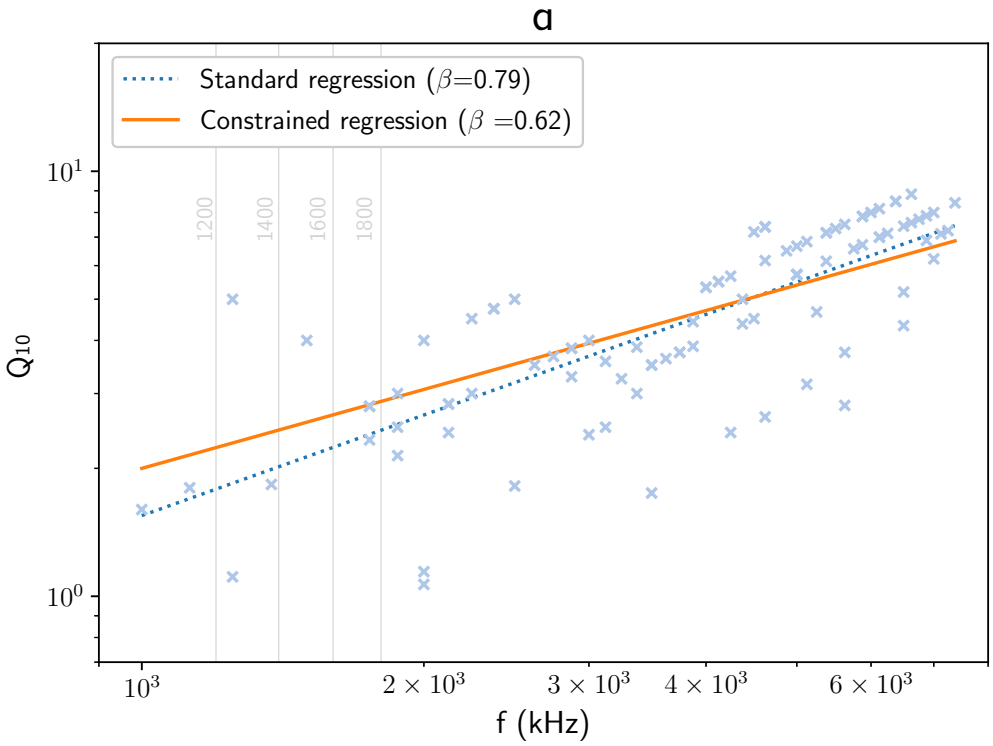
[1] In the other experiments based on different weightings strategies (appendix S1), the values for [ɔ], [ɑ] were consistent with ICA. It is possible that the chosen settings for the main figure (Strategy C) is not the most appropriate for the low back vowels.

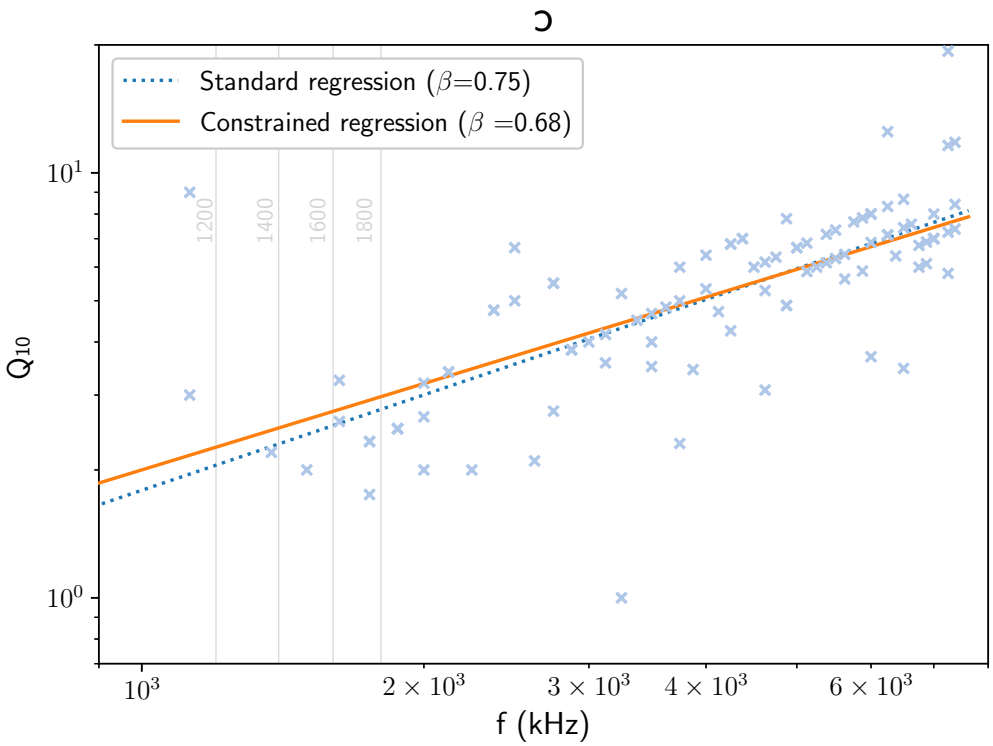

æ
Standard regression (β=0.88)
Constrained regression (β =0.38)
1200
1400
1600
1800
10^1
10^0
Q_{10}
10^3
2×10^3
3×10^3
4×10^3
6×10^3
f (kHz)

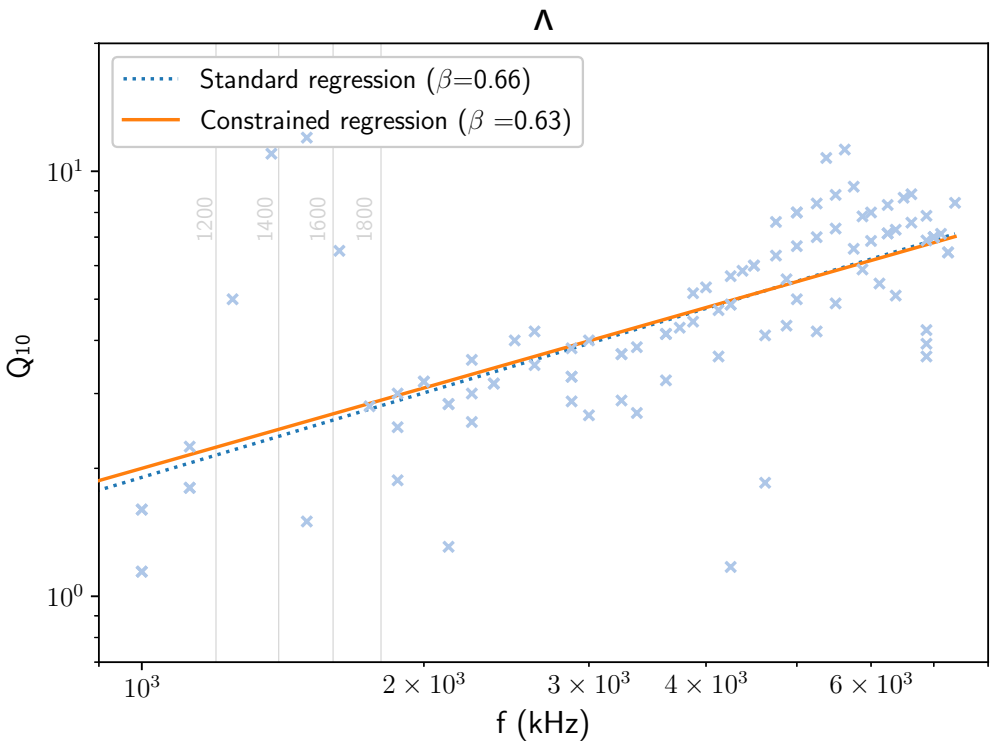

eɪ
Standard regression (β=1.01)
Constrained regression (β =0.51)
1200
1400
1600
1800
10^1
10^0
Q_{10}
10^3
2×10^3
3×10^3
4×10^3
6×10^3
f (kHz)

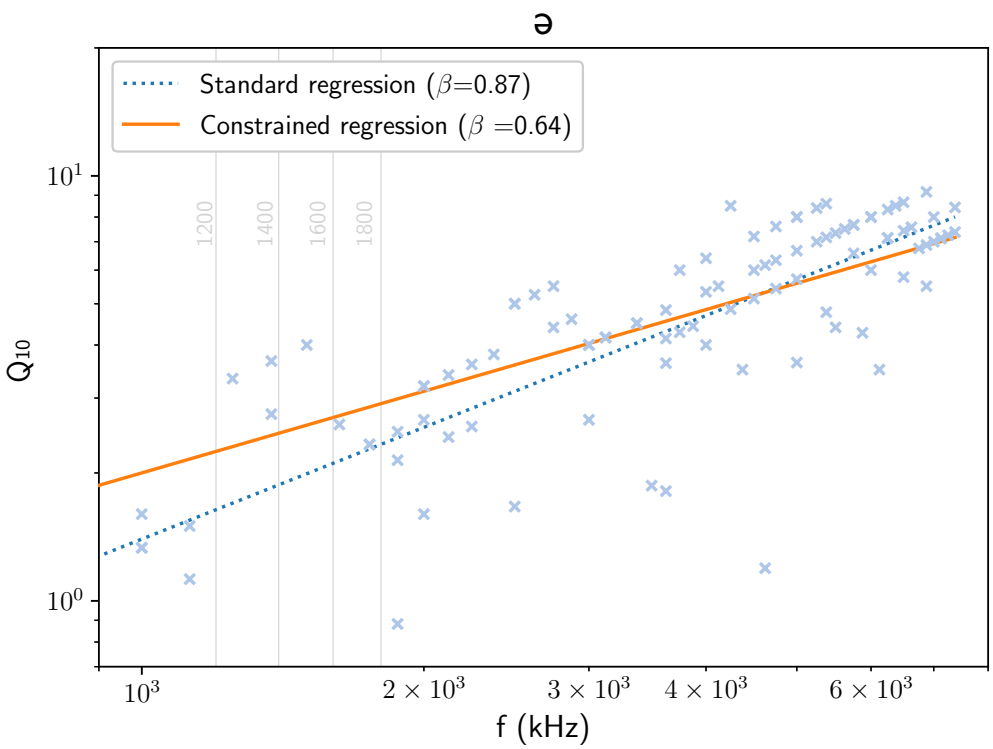

aI
Standard regression (β=1.01)
Constrained regression (β =0.49)
10^1
10^0
Q_{10}
1200
1400
1600
1800
10^3
2×10^3
3×10^3
4×10^3
6×10^3
f (kHz)

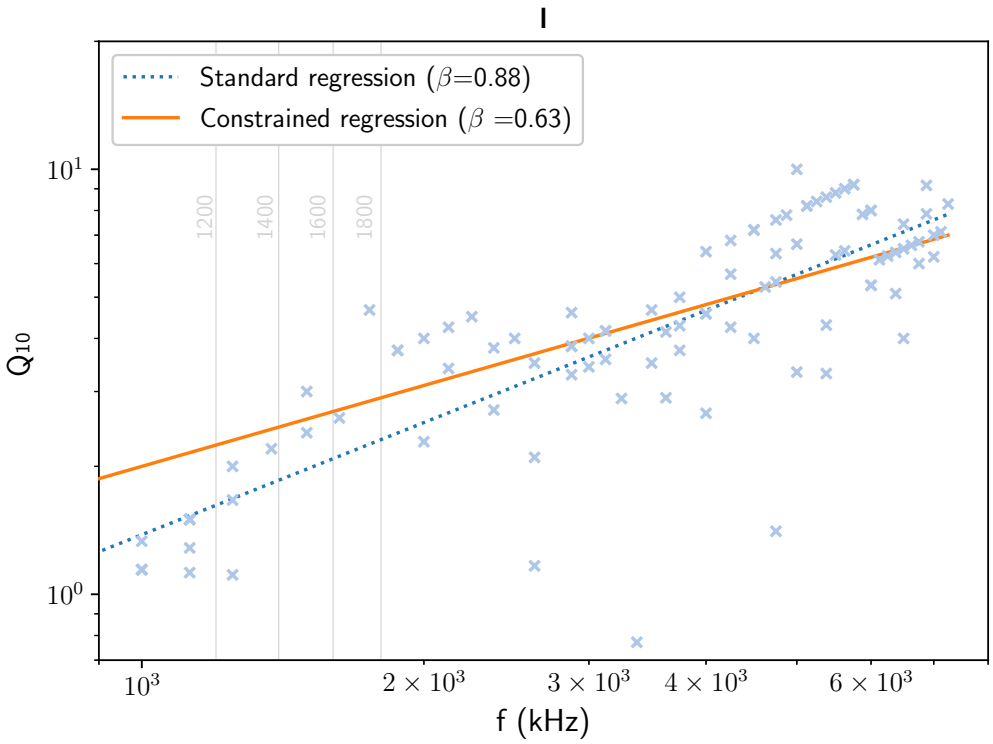

ε
Standard regression (β=0.77)
Constrained regression (β =0.55)
10^1
10^0
Q_{10}
1200
1400
1600
1800
10^3
2×10^3
3×10^3
4×10^3
6×10^3
f (kHz)

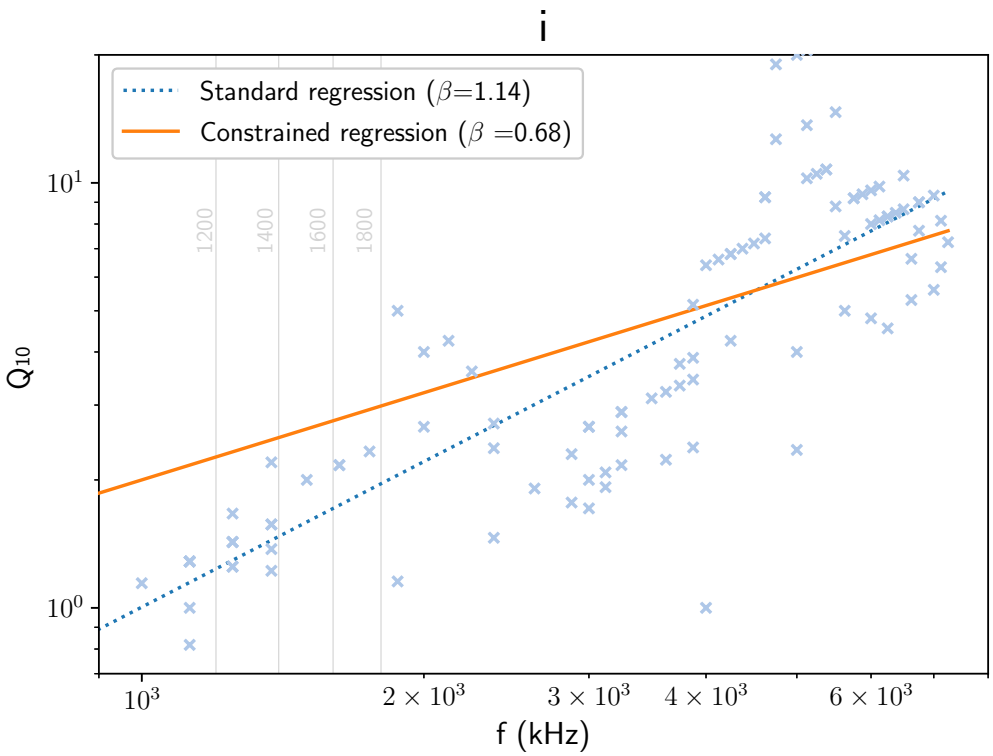

a
Standard regression (β=0.79)
Constrained regression (β =0.62)
10^1
10^0
1200
1400
1600
1800
Q_{10}
10^3
2×10^3
3×10^3
4×10^3
6×10^3
f (kHz)

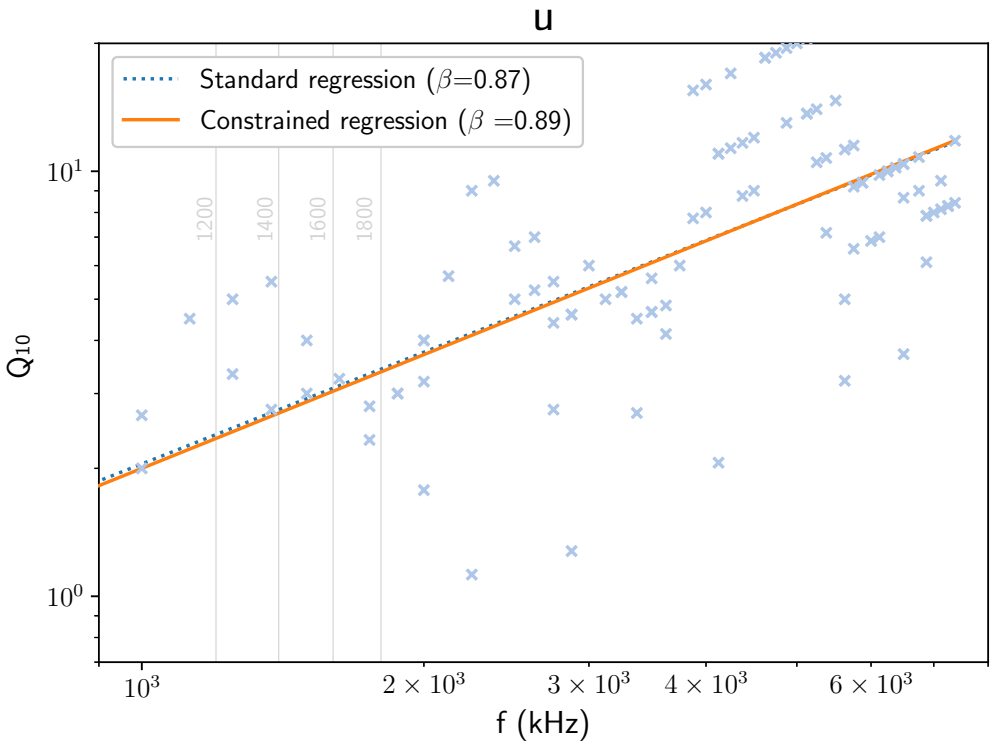

c
Standard regression (β=0.75)
Constrained regression (β =0.68)
10^1
10^0
1200
1400
1600
1800
Q_{10}
10^3
2×10^3
3×10^3
4×10^3
6×10^3
f (kHz)

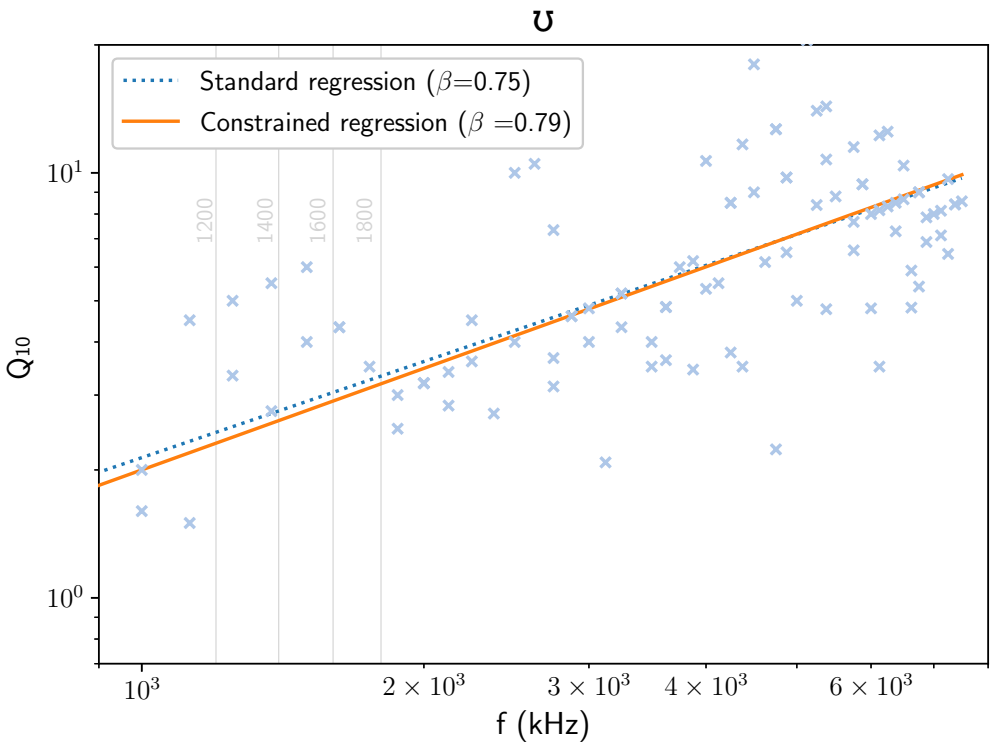

ʌ
Standard regression (β=0.66)
Constrained regression (β =0.63)
1200
1400
1600
1800
10^1
10^0
Q_{10}
10^3
2×10^3
3×10^3
4×10^3
6×10^3
f (kHz)

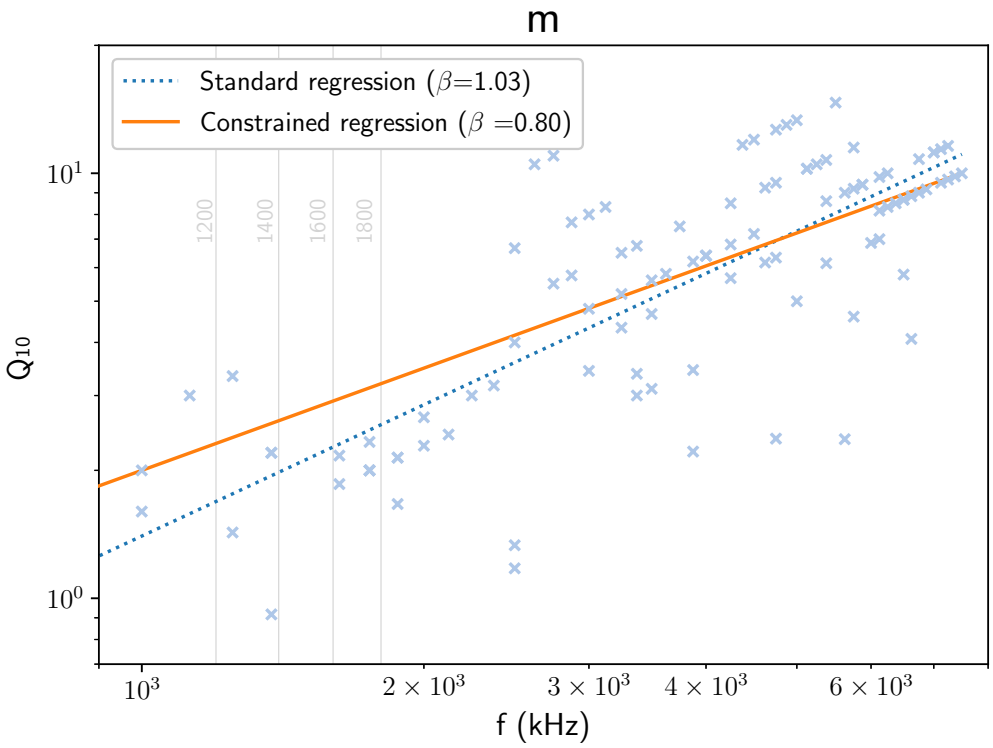

ə
Standard regression (β=0.87)
Constrained regression (β =0.64)
1200
1400
1600
1800
10^1
10^0
Q_{10}
10^3
2×10^3
3×10^3
4×10^3
6×10^3
f (kHz)

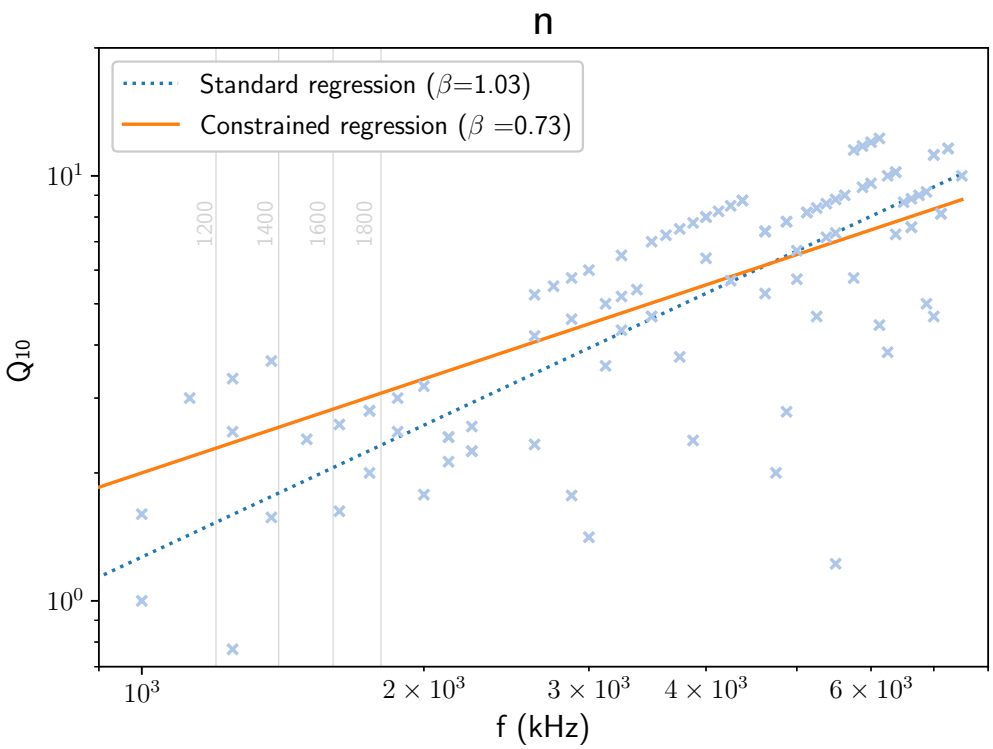

l
Standard regression (β=0.88)
Constrained regression (β =0.63)
10^1
10^0
Q_{10}
1200
1400
1600
1800
10^3
2×10^3
3×10^3
4×10^3
6×10^3
f (kHz)


i
Standard regression (β=1.14)
Constrained regression (β =0.68)
10^1
10^0
Q_{10}
1200
1400
1600
1800
10^3
2×10^3
3×10^3
4×10^3
6×10^3
f (kHz)


u
Standard regression (β=0.87)
Constrained regression (β =0.89)
1200
1400
1600
1800
10^1
10^0
Q_{10}
10^3
2×10^3
3×10^3
4×10^3
6×10^3
f (kHz)


ʊ
Standard regression (β=0.75)
Constrained regression (β =0.79)
1200
1400
1600
1800
10^1
10^0
Q_{10}
10^3
2×10^3
3×10^3
4×10^3
6×10^3
f (kHz)


m
Standard regression (β=1.03)
Constrained regression (β =0.80)
Q_{10}
10^1
10^0
1200
1400
1600
1800
10^3
2×10^3
3×10^3
4×10^3
6×10^3
f (kHz)


n
Standard regression (β=1.03)
Constrained regression (β =0.73)
Q_{10}
10^1
10^0
1200
1400
1600
1800
10^3
2×10^3
3×10^3
4×10^3
6×10^3
f (kHz)