A review of P-Centre models

Rudi Villing

Tomas Ward

Joseph Timoney

Contents

- Introduction and motivation
- About the models
- Test corpus
- Comparison of predicted P-centres
- Subjective evaluation of predicted P-centres
- Conclusion

Introduction and motivation

- The P-centre (or Perceptual Attack Time) is hypothesised to be…
  - the perceptual "moment of occurrence" of a sound
  - that which is regular in a perceptually regular sequence of sounds
- P-centres and rhythm (by definition)
  - The rhythm of a sequence of sounds is given by the interval between P-centres
  - Applies to perception and production
- Typical P-centre assumptions
  - The P-centre is a single unique location (not a region) in a sound
  - The p-centre is context independent (e.g. doesn't depend on neighbouring sounds in a sequence)
  - All the models being reviewed make these assumptions

    (though Pompino-Marschall does suggest two conflicting features exist)
- Many unresolved questions (not addressed in this work)
  - E.g. Are P-centres a feature of all sounds? Is speech special?

- Good P-centre model(s) required…
  - to accurately analyse rhythm in the natural performance of music or production of speech
  - to accurately construct/edit/synthesise speech or music with a specific rhythm
- Several existing models
  - Are their predictions similar or different?
  - Do their predictions match subjective perception?
- No published comparison or evaluation of all models exists (to our knowledge)
  - Most recent model published in 1997
  - This work is in progress

About the models

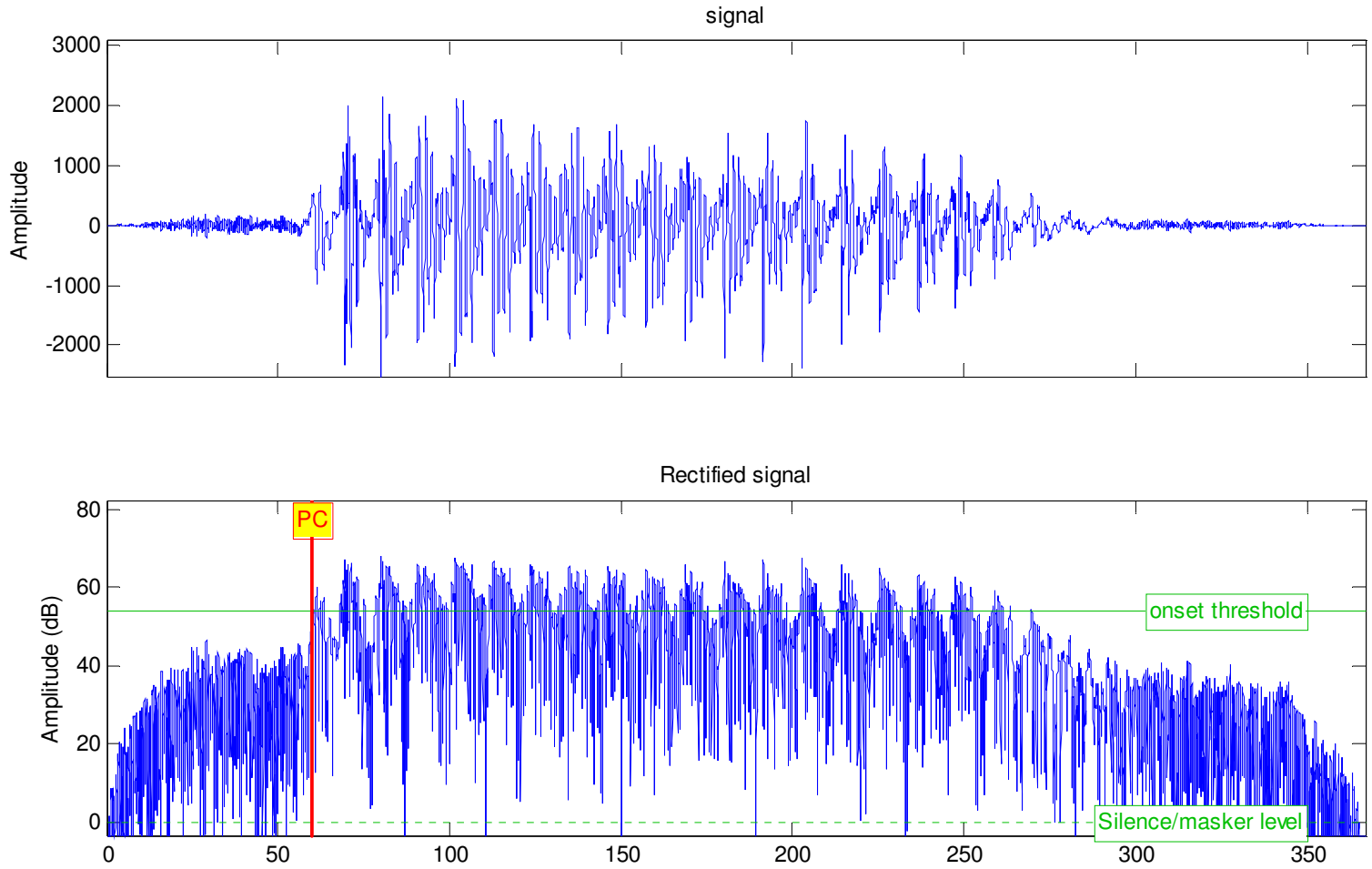
- 2 broad categories
- Using local onset features
  - Rapp-Holmgren (1971)
  - Vos & Rasch (1981)
  - Gordon (1987)
  - Scott (1993)
- Using weighted sum of "global" features
  - Marcus (1981)
  - Howell (1984/1988) [not implemented in this work]
  - Pompino-Marschall (1989)
  - Harsin (1997)

- Models using local onset features
  - Insensitive to post-onset differences
  - Threshold approaches insensitive to supra-threshold differences
  - Simple AM approaches fairly insensitive to timbre/pitch changes
  - P-centre "decision" available before sound has ended
- Models using weighted sum of "global" features
  - Sensitive to differences throughout the sound, but usually weighted to favour onset
  - Speech segment approaches may not be applicable to non-speech sounds
  - "Events" identified in event based approaches may not be perceptually relevant
  - P-centre "decision" only after end of sound
- All current models
  - Require isolated sounds with single P-centres, so no continuous sequences

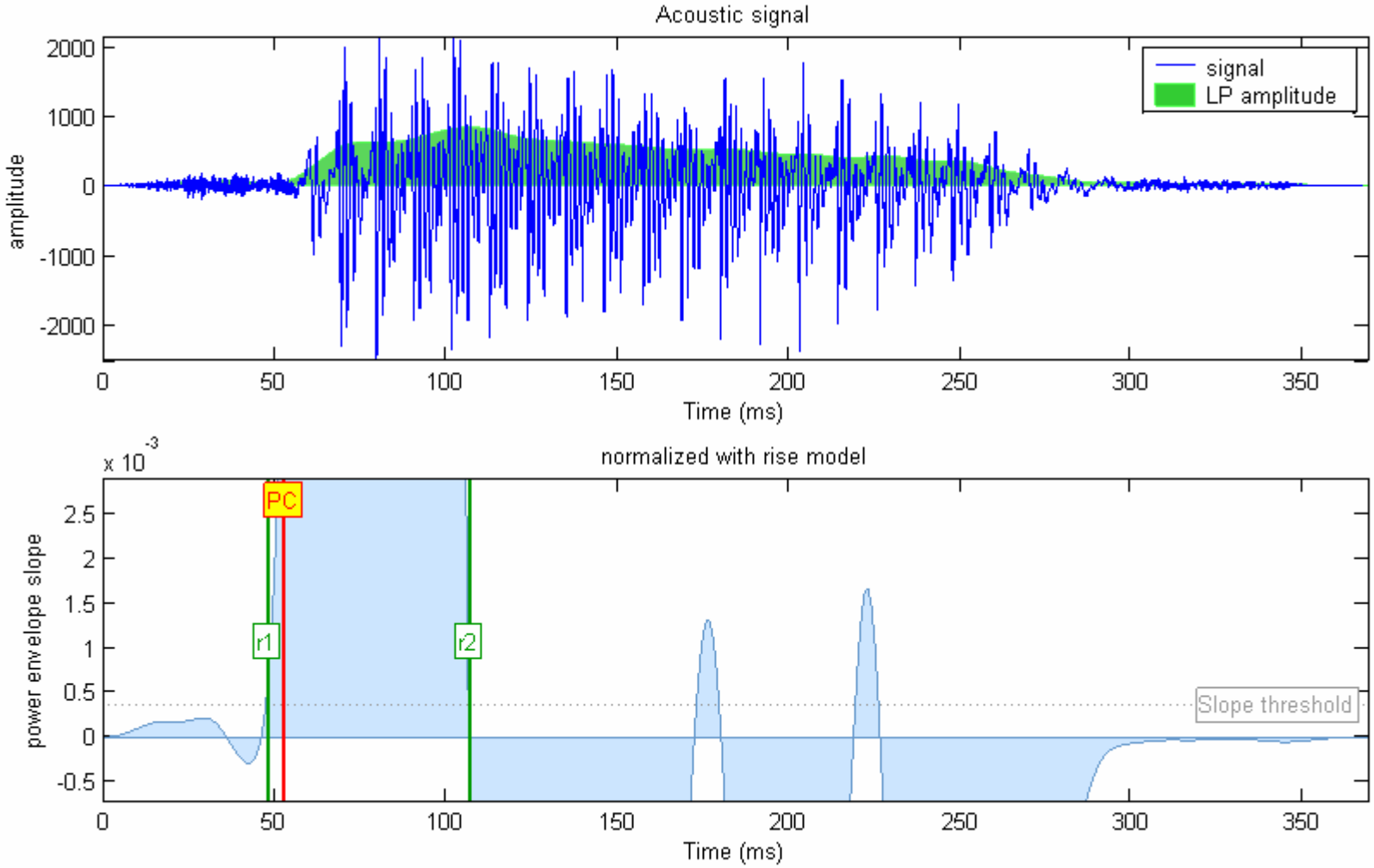
Data used for modelling

- Due to difficulty measuring P-centres, most models have been trained/fitted with a sparse corpus
  - Marcus: one … nine (natural); ba, da, ga, ta, ka (edited)
  - Vos and Rasch: Synthetic sawtooth (various onset ramps)
  - Gordon: 16 instrument tones (re-synthesised natural)
  - Pompino-Marschall: ma, am, shi (synthesised, various durations)
  - Scott: one, two (several speakers); la, ra, wa, ya (peak clipped), eight (edited), cha-sha, wa, ae (various onset ramps)
  - Harsin: Sha, na, ra; ta da ka ga; ash, an, ar (edited)
- Modelling data also differs by
  - Frequency content of sounds (nyquist limited)
  - Loudness of sounds (e.g. Gordon used 90dB)
  - Presentation method and subjective listening paradigm
- Can the models be successfully applied to…
  - sounds not in the training corpus?
  - sounds presented in a different environment?

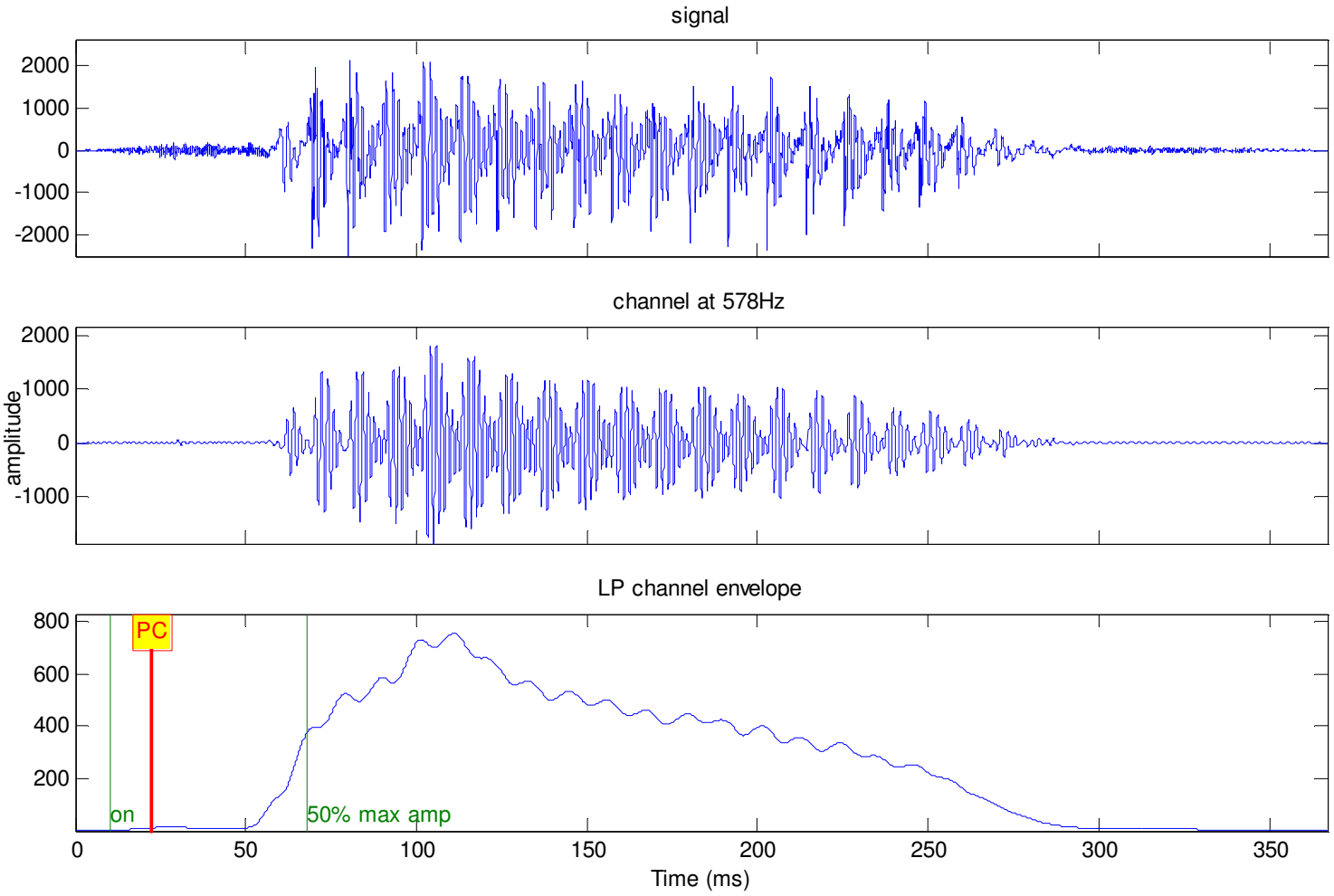
Vos & Rasch model



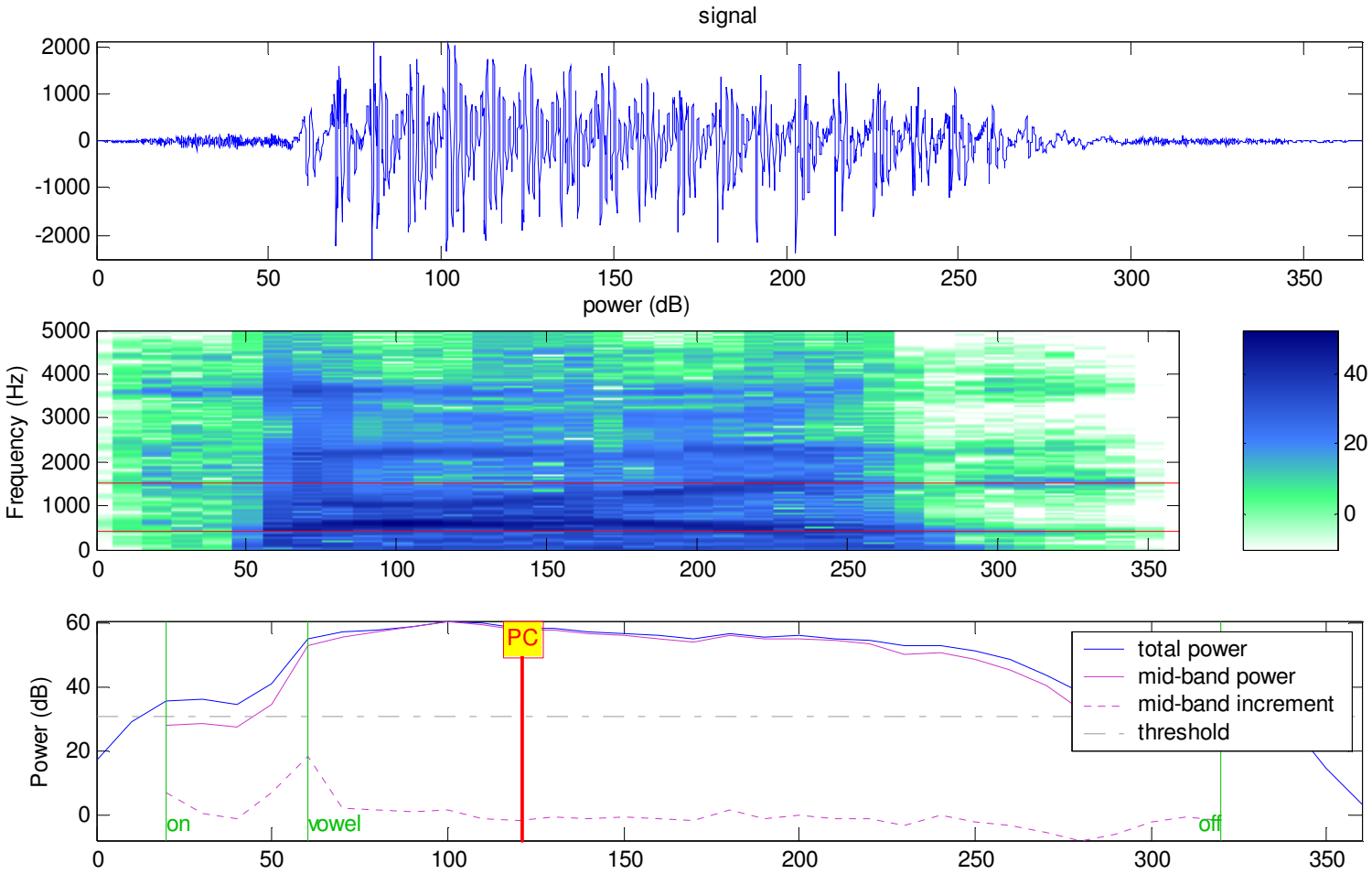
Gordon's model



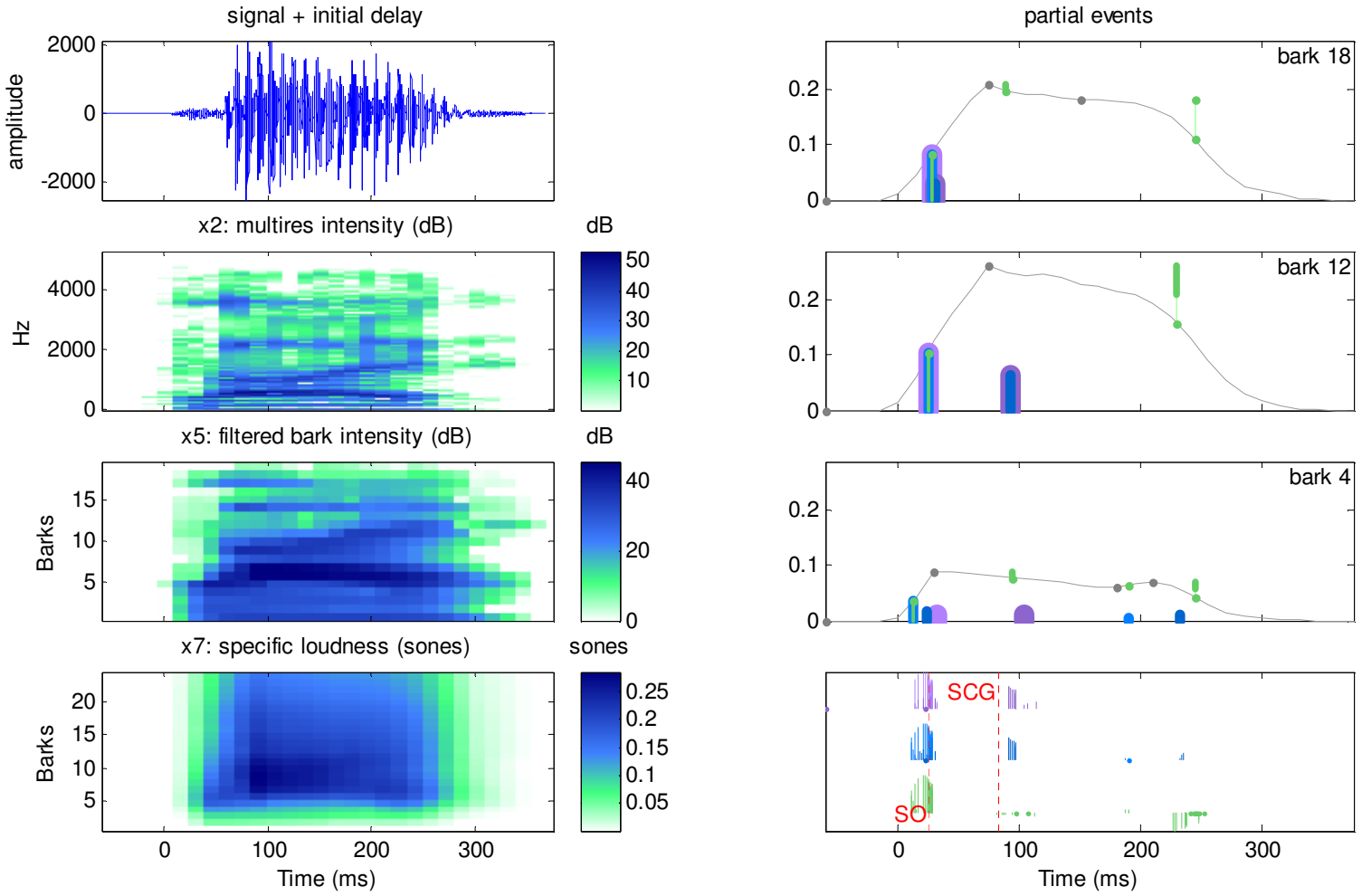
Scott's model



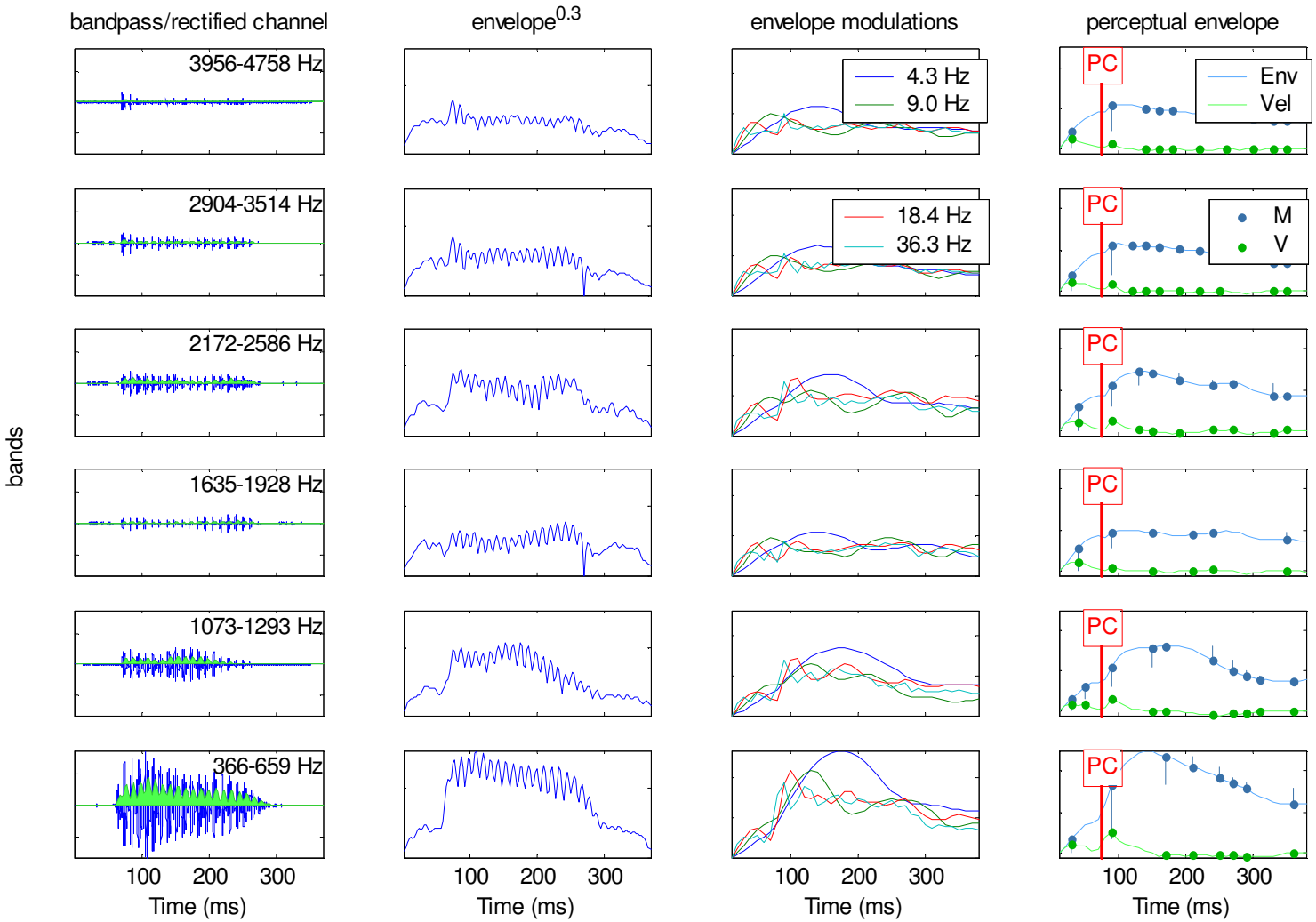
Marcus' (and Rapp's) model



Pompino-Marschall's model



Harsin's model



Test corpus

- The need for a standard corpus
  - Speech recognition/synthesis has benefited enormously from the existence of phonetically labelled speech corpora
  - Enables researchers to concentrate on modelling or P-centre measurement as appropriate
    - P-centre research is slowed by need for each researcher to label similar data using subjective listening experiments
  - There is currently no corpus of P-centre labelled sounds
  - Database published by Patel, Lofqvist et al (1999) has some limitations and is not labelled
- Would a P-centre labelled corpus be of use to other rhythm researchers?
  - Is there interest in participating/collaborating on such a database?

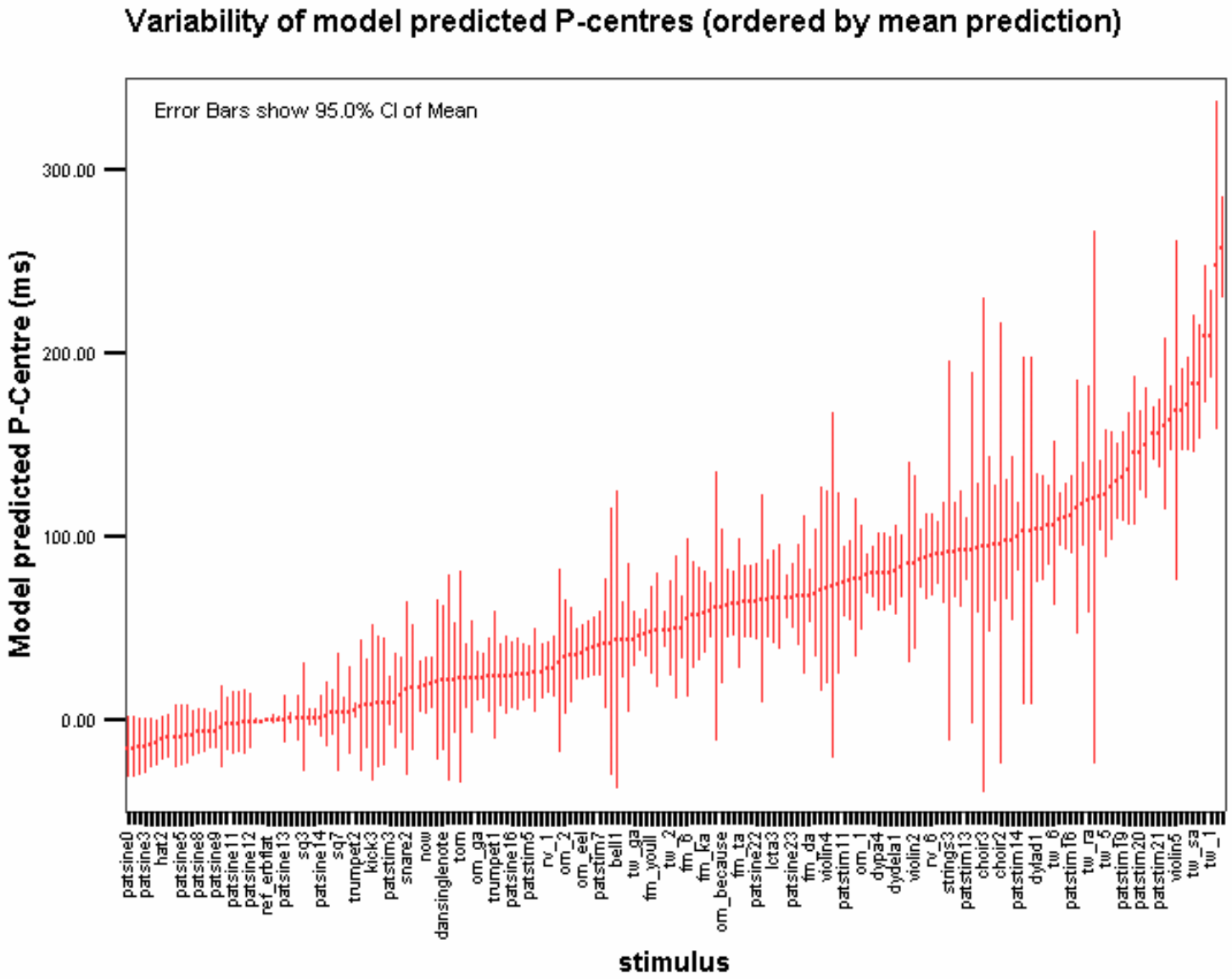
Our test corpus

- 189 sounds in three categories
  - Natural speech
    - Almost exclusively isolated monosyllables
    - Some sounds recorded at NUIM
    - A subset of the database published by Patel also included
  - Synthetic sounds
    - Amplitude modulated noise, pure and complex tones
    - Some from NUIM, and all those from the database of Collins
  - Musical Instruments
    - A subset of the database of sounds used by Collins

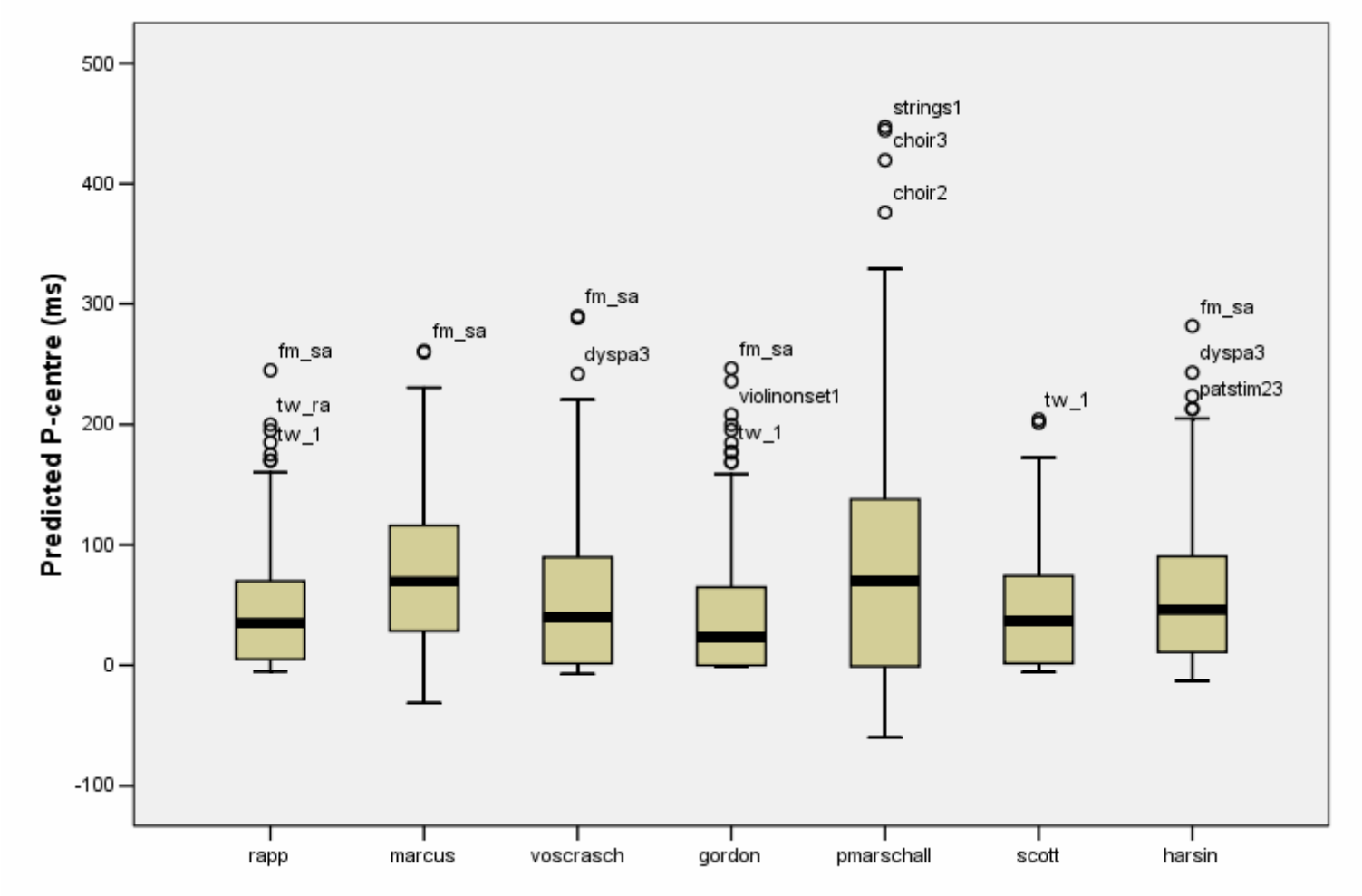
- All sounds approximately equalised for loudness
  - All sounds equalised to same loudness as a (nominally) 60dB SPL tone
  - BS 1770 weighting curve with short time scale (125ms) exponential averaging
- Reference sound
  - A 0dB noise to tone mix of pink noise and a harmonic tone with the same spectral shape
    - Intended to mitigate streaming effects when repetitively alternated with a variety of target stimuli
  - 200ms duration with cosine shaped onset (20ms) and offset (180ms)
  - "Absolute" P-Centre of reference sound assumed to be zero

Objective evaluation method

- Use models to predict P-centres for all sounds
- For each model, normalise predicted P-centres relative to reference sound
- Analysis
  - Which sound's predicted P-centres vary most/least?
  - Which model's predictions vary most/least?
  - Which models are in closest agreement?


Variability of model predicted P-centres (ordered by mean prediction)
Error Bars show 95.0% CI of Mean
Model predicted P-Centre (ms)
300.00
200.00
100.00
0.00
stimulus
patsine0
patsine3
hat2
patsine5
patsine8
patsine9
patsine11
patsine12
ref_erbflat
patsine13
sq3
patsine14
sq7
trumpet2
kick3
patstim3
snare2
now
dansinglenote
tom
om_ga
trumpet1
patsine16
patstim5
rv_1
om_2
om_eel
patstim7
bell1
tw_ga
fm_youll
tw_2
fm_6
fm_ka
om_because
fm_ta
patsine22
lota3
patsine23
fm_da
violin4
patstim11
om_1
dyga4
dydela1
violin2
rv_6
strings3
patstim13
choir3
choir2
patstim14
dylad1
tw_6
patstim16
tw_ra
tw_5
patstim19
patstim20
patstim21
violin5
tw_sa
tw_1

P-centre prediction variability by model



Agreement between models

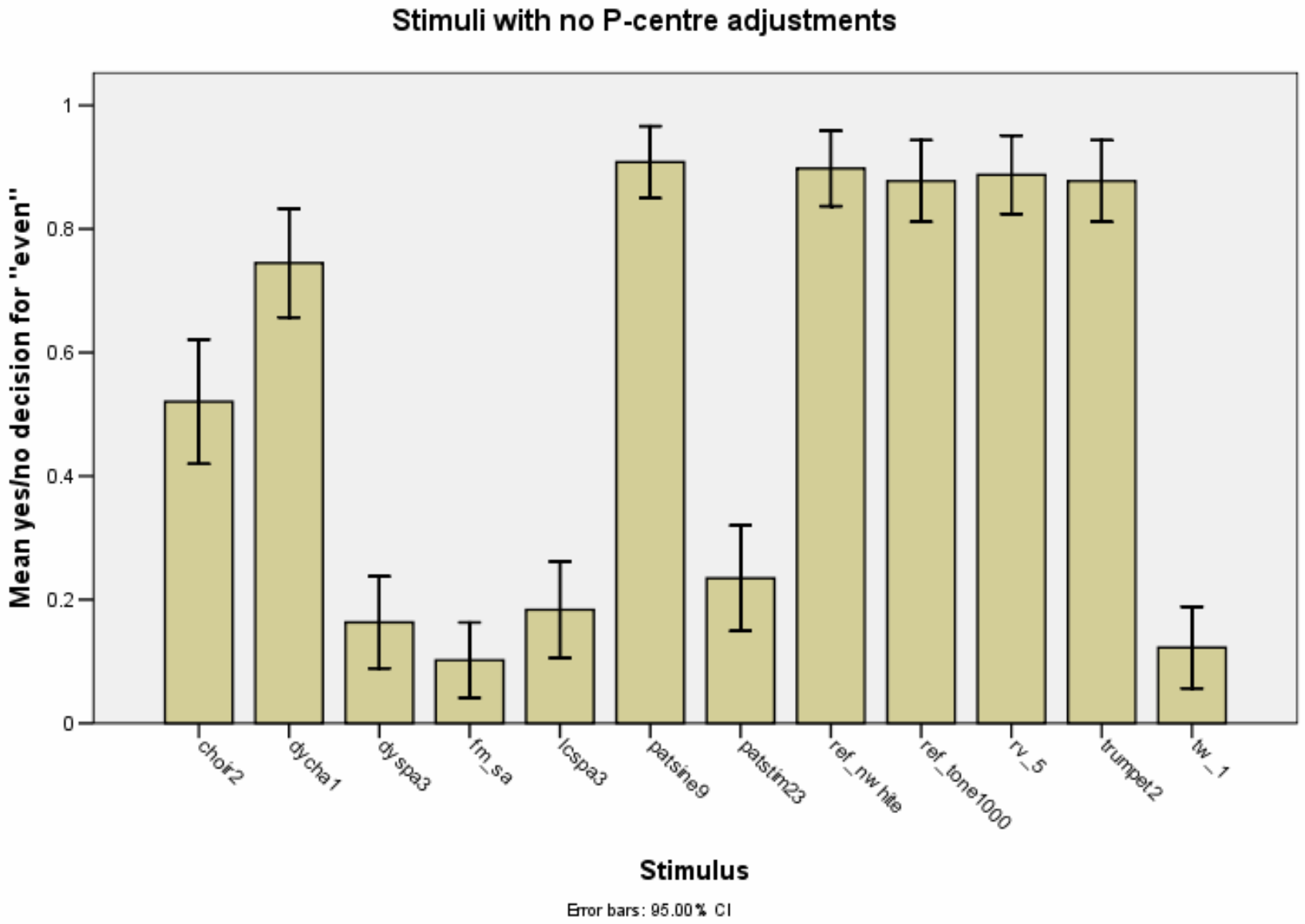
Proximity Matrix

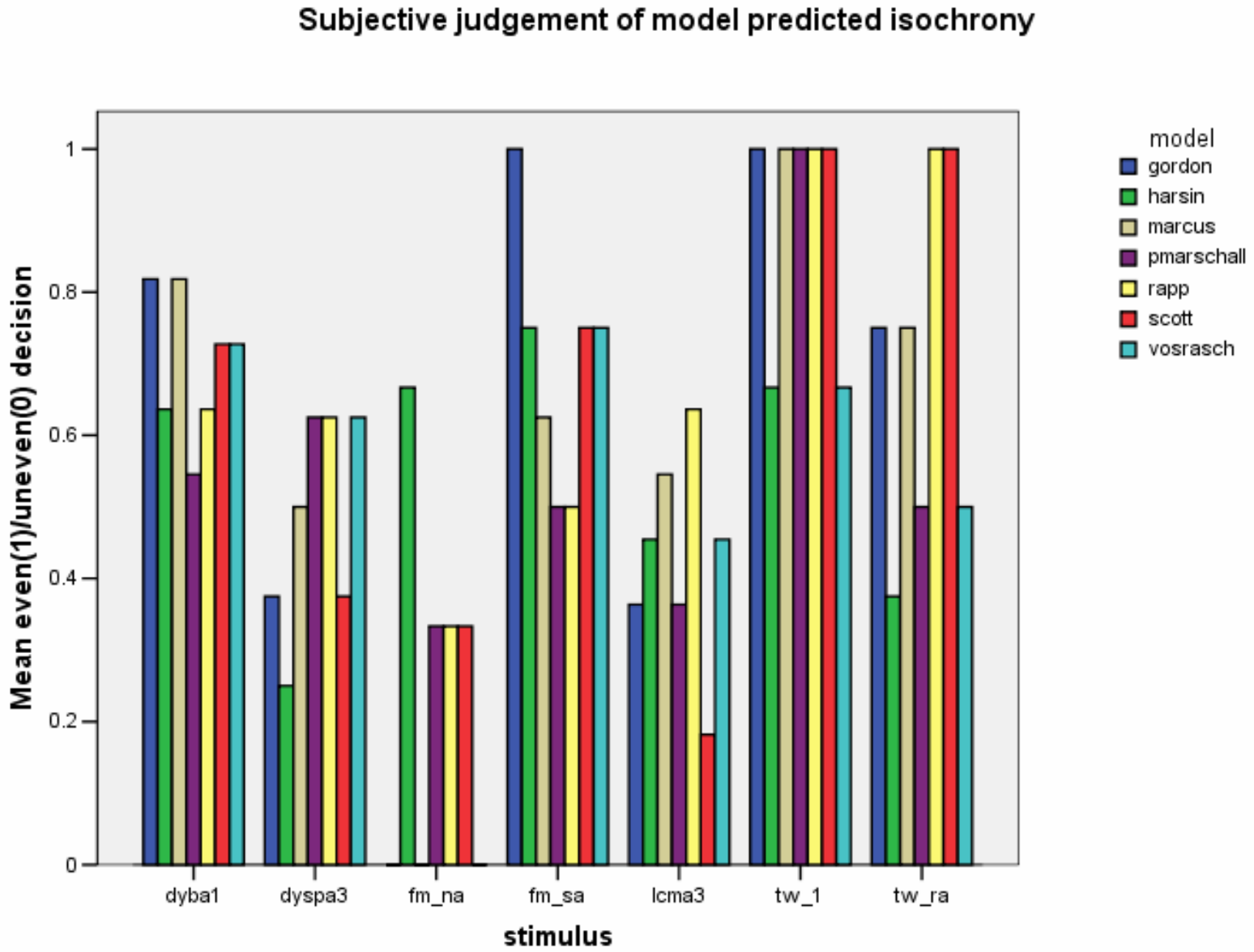
	Correlation between Vectors of Values						
	rapp	marcus	voscrasch	gordon	pmarschall	scott	harsin
rapp	1.000	.806	.795	.749	.487	.740	.770
marcus	.806	1.000	.692	.596	.741	.694	.621
voscrasch	.795	.692	1.000	**.913**	.614	**.881**	**.893**
gordon	.749	.596	**.913**	1.000	.482	.834	**.906**
pmarschall	.487	.741	.614	.482	1.000	.726	.532
scott	.740	.694	**.881**	.834	.726	1.000	**.878**
harsin	.770	.621	**.893**	**.906**	.532	**.878**	1.000

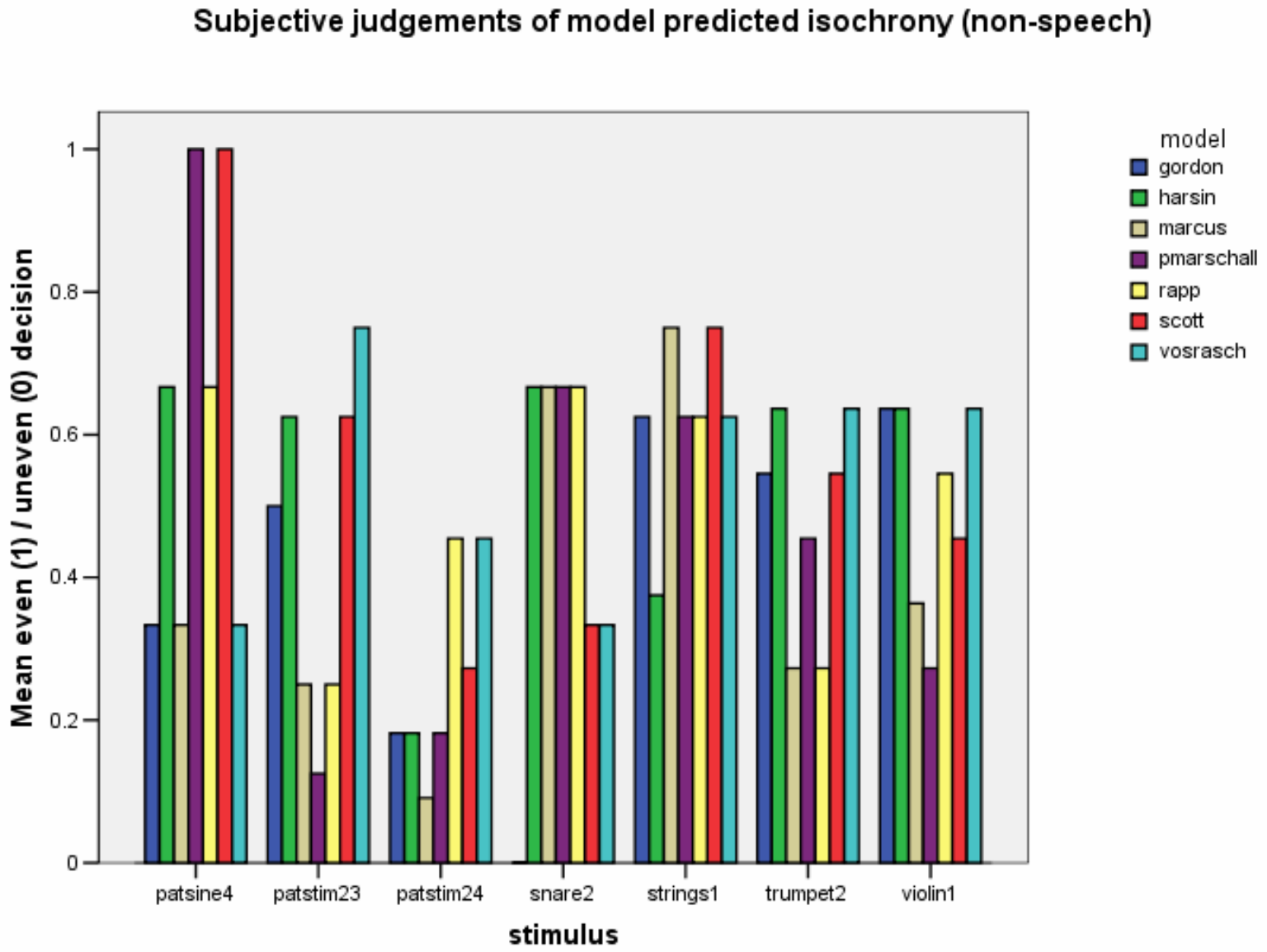
This is a similarity matrix

Simple subjective evaluation

- Quick test
  - Identify sounds with lowest/highest variance in speech, synthetic and instrumental categories
- For each model
  - construct cyclic sequence of alternating sounds predicted to be perceptually isochronous
  - Nominal ISI is 600ms
  - Sequence pattern: A-B-A-B-A-B-A
- Participant task
  - Sequence presented once over headphones
  - Is sequence isochronous? Yes/no forced choice.
- Analysis
  - Are there sequences which are not judged perceptually regular for any/all models?


Stimuli with no P-centre adjustments
Mean yes/no decision for "even"
1
0.8
0.6
0.4
0.2
0
choir2
dychat1
dyspa3
fm_sa
lcspa3
patsine9
patstim23
ref_nwhite
ref_tone1000
rv_5
trumpet2
tw_1
Stimulus
Error bars: 95.00% CI


Subjective judgement of model predicted isochrony
Mean even(1)/uneven(0) decision
1
0.8
0.6
0.4
0.2
0
dyba1
dyspa3
fm_na
fm_sa
lcma3
tw_1
tw_ra
stimulus
model
gordon
harsin
marcus
pmarschall
rapp
scott
vosrasch


Subjective judgements of model predicted isochrony (non-speech)
Mean even (1) / uneven (0) decision
1
0.8
0.6
0.4
0.2
0
patsine4
patstim23
patstim24
snare2
strings1
trumpet2
violin1
stimulus
model
gordon
harsin
marcus
pmarschall
rapp
scott
vosrasch

Conclusions

- More data required, but initial data is suggestive
  - All models have problems with some sounds
  - Subjective isochrony appears to depend more on stimulus than model
    - Is P-centre of such stimuli ill-defined?
    - Do models share common assumptions or elements that result in common failures?
- Any future P-centre modelling exercise must consider a broader corpus of sounds

Thank you.

Questions?