Grid: A corpus for perceptual and computational modelling studies of speech identification in noise

Martin Cooke, Jon Barker & Stuart Cunningham

Speech and Hearing Research
Department of Computer Science
University of Sheffield
http://www.dcs.shef.ac.uk/~martin

Caveat

- This is not a corpus for general-purpose auditory scene analysis
- Instead, it is targetted at the problem of speech identification in multitalker conditions [N-babble continuum is a useful source of backgrounds in which to study the effects of energetic and informational masking, stationarity, background/foreground grouping cues, background/foreground speech models, etc]

Existing models of speech intelligibility

Plenty of *macroscopic* models of speech perception (energetic masking)

Mainstream

- Articulation index (French & Steinberg, 1947)
- Speech Intelligibility Index (ANSI S3.5, 1997)
- Speech Transmission Index (Steeneken & Houtgast, 1980; 1999)

Recent models

- Speech Recognition Sensitivity (Musch & Buus, 2001a,b)
- Spectro-Temporal Modulation Index (Elhilali, Chi & Shamma, 2003)
- Multiple-looks for speech in noise (Hant & Alwan, 2003)

Macroscopic models

Pros

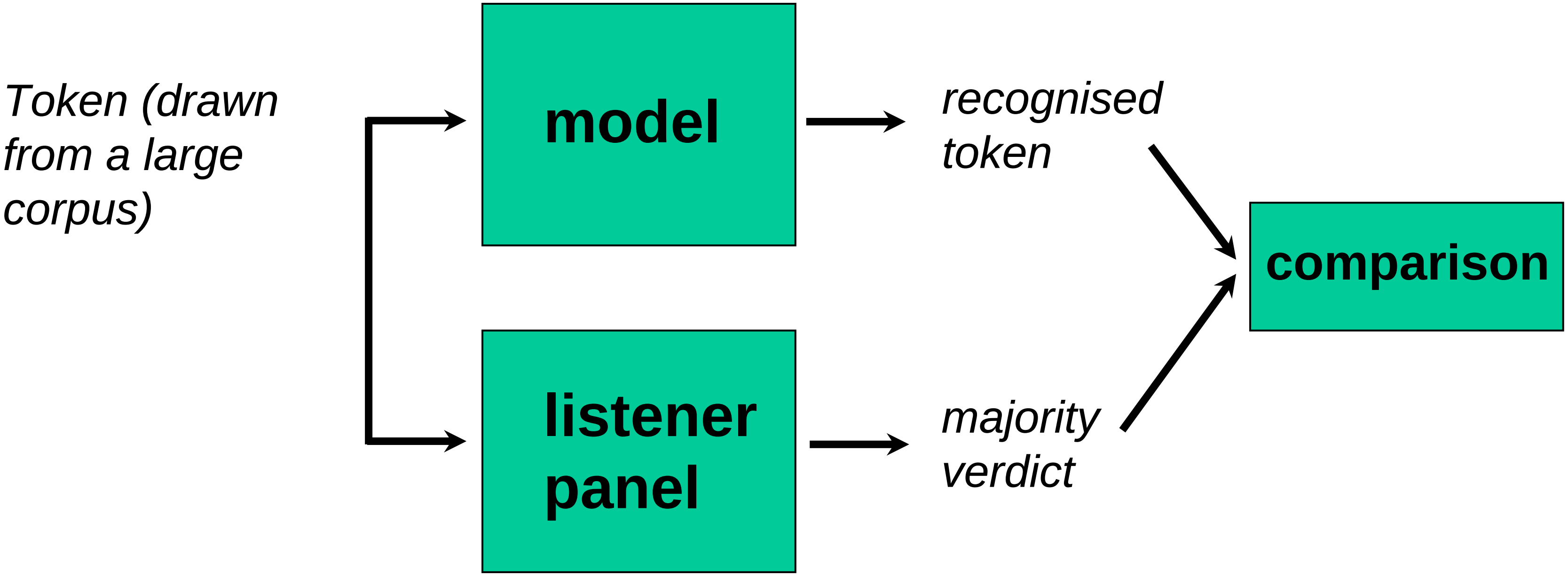
- Can obtain a rapid intelligibility estimate
- Quite accurate for a range of transmission conditions involving filtering, slowly-varying noise, level differences and reverberation

Cons

- Not designed for many common listening situations eg competing talkers
- May be easy to predict mean intelligibility with an incorrect model
- Usually cannot predict response to individual tokens or patterns of confusions

Not likely to lead to detailed insights about speech perception in everyday conditions?

Wouldn’t it be nice if … ?



Issues

- Slow and potentially laborious to collect panel responses for large corpus
- Is current state of knowledge about the how of speech perception sufficiently advanced?
- Few (any?) suitable corpora for human/model comparison

Existing corpora (1)

From speech perception

eg DRT, MRT, HINT, Shannon et al VCV, CRM, …

- Too small
- Too little variation
- Too controlled (synthetic, slow/clear speech, …)
- Usually contain tokens which are too short eg vowels, diphones, VCV syllables

From ASR

eg TIMIT, TIDigits, WSJ, Broadcast News, Switchboard,…

- Uncontrolled: contain too much unwanted variation
- Frequently unbalanced (phonetically/linguistically)
- Contain tokens which are too long for psychoacoustic work

Existing corpora (2)

Some corpora have been used for SP & modelling ….

- **Double-vowels** (used by Scheffers, Assmann & Summerfield, Meddis & Hewitt, Culling & Darwin, …)
- **Digit sequence**s eg TIDigits (used by Palomaki)
- **Syllables**
  - DRT (used by Ghitza)
  - Shannon et al VCV (used by Cooke, Meyer, …)
- **Low-perplexity sentences**
  - CRM (used by Barker & Cooke)

… but all have problems

VCVs (Shannon et al, 1999)

Design

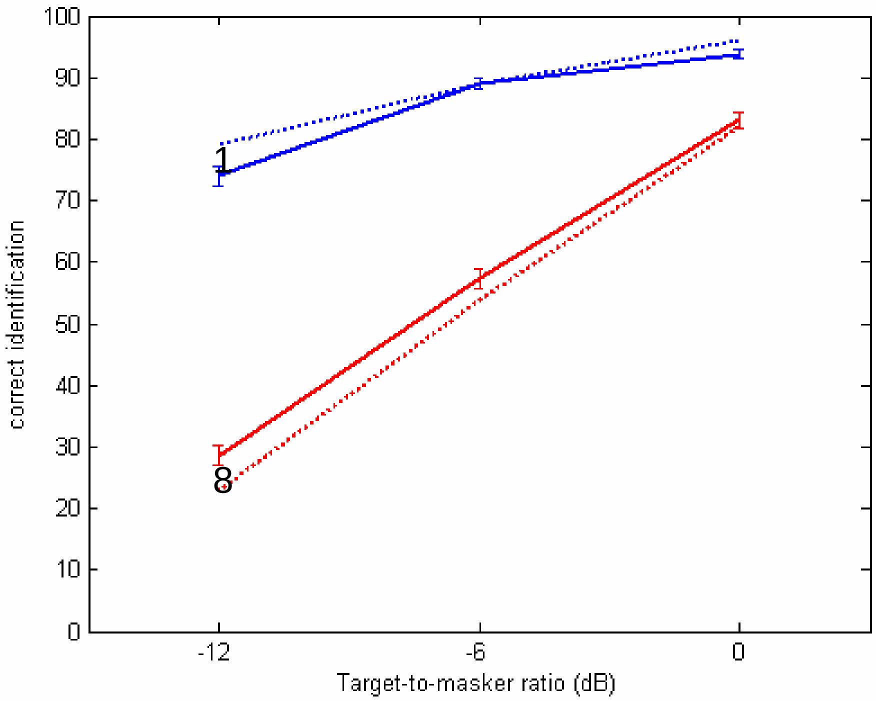
25 English consonants in CV and VCV settings for 3 vowels /i, a, u/, spoken by 5 males and 5 females

Pros

- Reasonable for measuring energetic masking
- Fast to train ASR component

Cons

- Unnaturally slow utterances
- Difficult to produce informational masking
- 10 repeats of each sufficient to train ASR, but lack of variability



Example use of VCV corpus in Cooke (2003) showing listeners (solid) vs glimpsing model (dotted)

Coordinate Response Measure (CRM, Bolia et al, 2000)

Design

READY <callsign:8> GO TO <color:4> <number:8> NOW

“Ready baron go to green three now”

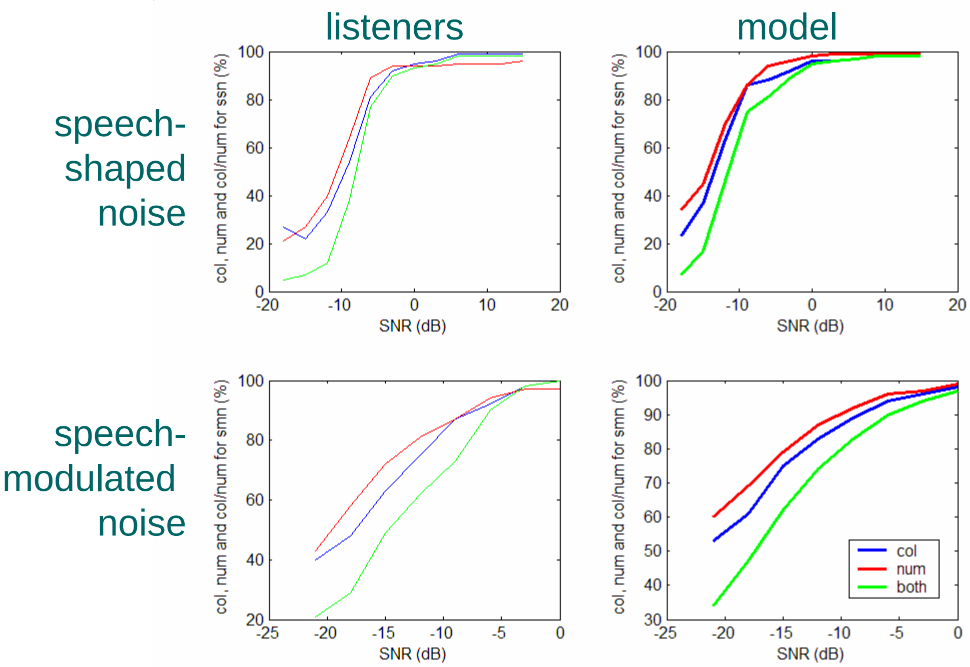
8 talkers, all combinations of callsigns (8), colours (4) and numbers (8) = 256 sentences each

Pros

- Many listening studies involving speech identification in multitalker environments (Brungart et al)
- Good for info masking
- Fast to train ASR component

Cons

- Small vocabulary effects/lack of phonetic balance
- Artefacts in multitalker stimuli due to identical fillers eg N-talker CRM babble
- Low variability across tokens



Example use of CRM corpus in Barker & Cooke (2004)

The Grid corpus

Design aims

- Designed explicitly for joint **modelling and perceptual** studies
- Build on **CRM** experience
- **Not too large a step** from state-of-the-art in robust ASR (cf ShATR corpus)
- **Easy to build** ASR without need for a large infrastructure (no high-level linguistic component)
- **Useful** robust ASR task
- Make up for lack of a large, free ***audiovisual*** corpus

Design

Format

<action:4> <colour:4><preposition:4><alpha:26><digit:10><endfiller:4>

“*Put green at A4 now*”

“*Place red in Q9 please*”

Extends CRM:

- improved phonetic balance (alphadigits rather than colours)
- reduced artefacts due to constant fillers (use of variable fillers)
- increased variability (64 speakers rather than 10)
- increased size (64000 sentences vs 256)
- incorporates important ASR problem domains: alphas/digits
- adds visual component for AV studies
- allows variable ‘callsign’ to target distance (including backward)
- … removes militaristic connotation

Timescale

Collection

Nov-Dec 2004

Audio, visual and audiovisual intelligibility assessment

Q1 2005

Annotation and release (free on web, DVDs at cost)

Q2-3 2005

Issues

- Not really representative of everyday spoken language communication
- Anechoic and monaural (but easy to synthesise reverberant and binaural tokens)
- No non-speech sources (but can be added)
- No high-level linguistic component
- More a corpus for next-generation detailed models of early speech perception than for general-purpose models of speech/source separation
- British English only for now