The Nijmegen Lectures: Lecture 1

On how speech is pretty special

David Poeppel
Max-Planck-Institute, Frankfurt
Department of Psychology, NYU
Recurring themes through the three lectures:

• *the notion of specialization* (classical cognitive science)
• *cortical oscillations* (contemporary neuroscience)
• *the mapping problem* (aligning cognition and neuroscience)
Outline

• Warm-up for specialization in the auditory domain

• A tantalizing hint from neuropsychology

• An fMRI experiment on specialization for speech, in detail

• The modulation of speech
Current Biology

Human

Communities

Report

Arnal et al. 2015, Current Biology
Figure A: 1 kHz tone, 25 Hz AM
- Amplitude
- Frequency (kHz)
- Time (s)

Figure B: Sentence
- Amplitude
- Frequency (kHz)
- Time (s)

Figure C: Modulation Power Spectrum (MPS)
- Spectral Modulation (cyc/oct)
- Temporal Modulation (Hz)

Legend:
- Fundamental frequency (gender)
- Slow fluctuations (meaning)
- Roughness (zona incognita)
A

SCREAMED vs. NEUTRAL

B

[a] vs. SENTENCE

C

SCREAM effect

SENTENCE effect
Hypothesis: What makes a scream so effective: **roughness**
• reserved “acoustic niche” for alarm screams (roughness)
• speculation: sequential ‘colonization’ of the acoustic biotope (scream, identity, meaning)
Outline

- Warm-up for specialization in the auditory domain
- A tantalizing hint from neuropsychology
- An fMRI experiment on specialization for speech, in detail
- The modulation of speech
Zooming in on the problem: from vibrations in the ear to abstractions in the head

(a) Ear vibrations

(b) Spectrogram

(c) Phonological primal sketch

(d) Phonetic features:

- [+ cons, -son]
- [-con]
- LAR/PHAR
- PLACE
- GLOT
- DORSAL
- [-voice]

- [+ cons, -son]
- [-cont]
- LAR/PHAR
- PLACE
- GLOT
- CORONAL
- [-voice]
- [-ant]

Ear -> Brain

Poeppel, v. Wassenhove, Idsardi 2008
Possible relations between speech and non-speech recognition

(a)
Possible relations between speech and non-speech recognition

Poeppel 2001, Cognitive Science
The predictions of these models differ, for example in the case of neuropsychological patients

If (a) is correct, then a lesion leading to problems with auditory pattern recognition (agnosia) must also be associated with speech perception deficits. If (b) is correct, there is a double dissociation.

Poeppel 2001, Cognitive Science
**Test case: pure word deafness**

**Table 1** *Auditory disorders following cortical and/or subcortical lesions*

<table>
<thead>
<tr>
<th></th>
<th>Pure word deafness</th>
<th>Auditory agnosia</th>
<th>Cortical deafness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech comprehension</td>
<td>impaired</td>
<td>+</td>
<td>impaired</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(or mildly impaired)</td>
<td></td>
</tr>
<tr>
<td>Speech repetition</td>
<td>impaired</td>
<td>+</td>
<td>impaired</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(or mildly impaired)</td>
<td></td>
</tr>
<tr>
<td>Recognition of familiar non-speech sounds</td>
<td>+</td>
<td>impaired</td>
<td>impaired</td>
</tr>
<tr>
<td>Recognition of music</td>
<td>+</td>
<td>+/-</td>
<td>impaired</td>
</tr>
<tr>
<td>Hearing sensitivity (audiometry)</td>
<td>+</td>
<td>+</td>
<td>impaired</td>
</tr>
<tr>
<td>Language I: Spontaneous speech</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Language II: Reading comprehension</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Language III: Writing</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

+-sign indicates adequate performance in a given domain.

*Poeppel 2001, Cognitive Science*
Lesion pattern in PWD patients:

2/3 bilateral STG/STS (often sparing putative primary auditory ctx)

1/3 deep white matter lesions, most typically left WM, arguably disconnecting left and Right STG/STS.

Poeppel 2001, Cognitive Science

Fig. 2 – Schematic of the approximate lesion extent in the temporal lobe in the lateral view (top) and in top view (lower, superior temporal plane). Abbreviations: STS: superior temporal sulcus; STG: superior temporal gyrus; MTG: middle temporal gyrus; ITG: inferior temporal gyrus; HG: Heschl's gyrus; PT: Planum temporale. Note that the sketch of temporal lobe anatomy is based on a standard brain and not on the individual patient anatomy.

Gutschalk et al. 2015, Cortex
Functional anatomy of speech sound processing

Articulatory network
- pIFG, PM, anterior insula (left dominant)

Sensorimotor interface
- Parietal–temporal Spt (left dominant)

Input from other sensory modalities

Via higher-order frontal networks

Spectrotemporal analysis
- Dorsal STG (bilateral)

Phonological network
- Mid-post STS (bilateral)

Conceptual network
- Widely distributed

Combinatorial network
- aMTG, aITs (left dominant?)

Lexical interface
- pMTG, pITs (weak left-hemisphere bias)

Dorsal stream

Ventral stream

Hickok & Poeppel, 2007, Nat Rev Neurosci
Outline

• Warm-up for specialization in the auditory domain

• A tantalizing hint from neuropsychology

• An fMRI experiment on specialization for speech, in detail

• The modulation of speech
The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts

Tobias Overath, Josh H. McDermott, Jean Mary Zarate, David Poeppel

Duke  MIT  NPG

Department of Psychology, Center for Neural Science, NYU
Functional anatomy of speech sound processing

Articulatory network
pIFG, PM, anterior insula (left dominant)

Sensorimotor interface
Parietal–temporal Spt (left dominant)

Input from other sensory modalities

Via higher-order frontal networks

Spectrotemporal analysis
Dorsal STG (bilateral)

Phonological network
Mid-post STS (bilateral)

Conceptual network
Widely distributed

Combinatorial network
aMTG, aITS (left dominant?)

Lexical interface
pMTG, pITS (weak left-hemisphere bias)

Dorsal stream

Ventral stream

Hickok & Poeppel, 2007, Nat Rev Neurosci
Le Médecin guérit Phantastie
Purgeant aussi Par drogues la folie.
The goal of the present study:

- Linguistic analysis cannot proceed directly from the sound waveform itself, but is preceded by a stage of acoustic analysis that maps patterns of sound energy onto intermediate, perhaps invariant, representations of features, phonemes, or syllables.

- (i) isolate the auditory analysis of speech, in particular its temporal attributes. (ii) identify a locus of speech-specific acoustic analysis that is a candidate precursor to linguistic processing.

- by manipulating foreign speech that has no lexical-semantic or syntactic content for our listeners.

- Hypothesis: neurons underpinning speech analysis are tuned to the particular structures that occur in natural speech, such that they respond more to sound signals that have naturalistic speech structure than to those in which such structure is disrupted.

- Notably, speech is richly structured in time, with systematic organization at multiple timescales.
A Sequence of Object-Processing Stages Revealed by fMRI in the Human Occipital Lobe
Grill-Spector et al. (1998), *Human Brain Mapping*
Regions subserving speech-specific analysis should exhibit an increasing response to speech quilts as the segment length increases, because this manipulation increases the temporal extent over which the signal contains naturalistic speech structure.

The structure of a quilted signal is similar to that of the source signal within a segment and across a segment’s border, but differs from the source at larger scales for source signals that contain large-scale dependencies.
Long term power spectra are very similar between source and quilt signal ...
... but the modulation spectra are a bit different, especially in the syllable range.

Simoncelli & McDermott texture synthesis
Regions subserving speech-specific analysis should exhibit an increasing response to speech quilts as the segment length increases, because this manipulation increases the temporal extent over which the signal contains naturalistic speech structure.

The structure of a quilted signal is similar to that of the source signal within a segment and across a segment’s border, but differs from the source at larger scales for source signals that contain large-scale dependencies.
Synthetic signals (20s in length) were generated that matched either 1) the envelope marginal statistics and modulation power or 2) the envelope marginal statistics, modulation power, and cochlear correlations of each of the 20s speech source signals used for speech quilts.
Synthetic signals (20s in length) were generated that matched either 1) the envelope marginal statistics and modulation power or 2) the envelope marginal statistics, modulation power, and cochlear correlations of each of the 20s speech source signals used for speech quilts.
Regions subserving speech-specific analysis should exhibit an increasing response to speech quilts as the segment length increases, because this manipulation increases the temporal extent over which the signal contains naturalistic speech structure.

The structure of a quilted signal is similar to that of the source signal within a segment and across a segment’s border, but differs from the source at larger scales for source signals that contain large-scale dependencies.
Synthetic signals (20s in length) were generated that matched either 1) the envelope marginal statistics and modulation power or 2) the envelope marginal statistics, modulation power, and cochlear correlations of each of the 20s speech source signals used for speech quilts.
Regions subserving speech-specific analysis should exhibit an increasing response to speech quilts as the segment length increases, because this manipulation increases the temporal extent over which the signal contains naturalistic speech structure.

The structure of a quilted signal is similar to that of the source signal within a segment and across a segment’s border, but differs from the source at larger scales for source signals that contain large-scale dependencies.
One possible account: responses are driven by prosody, and in particular the **prosodic pitch variation** that occurs in natural speech. Such patterns are largely preserved in quilts made from long segments of speech, but are severely disrupted in quilts made from short segments.

To test the importance of pitch variation, we presented quilts made from noise-vocoded speech. We generated noise-vocoded versions of each of our source speech recordings by imposing the envelopes of 10 frequency bands (ERB-spaced, covering the entire spectrum) on noise.
Potential explanation for this response plateau is that the regions we have identified integrate sound information at time scales up to about half a second.

An alternative explanation is that the response asymptote reflects the statistics of speech. Specifically, it is conceivable that foreign speech does not contain substantial acoustic dependencies past half a second, such that 480 ms quilts and 960 ms quilts are largely acoustically equivalent.
Functional ROIs revealed by a parcellation algorithm. The five color-coded fROIs are rendered onto the smoothed surface (top) and on individual axial slices (bottom) of a template brain (avg152Ts.nii in SPM). The response to the six different segment length conditions (normalized with respect to the L960 localizer condition) is plotted for each of the five fROIs. The response pattern is similar across ROIs.
**Quilts**

- Preserve statistical structure of speech (without linguistic structure)

**Bilateral STS**

- Sensitivity to speech
- Specificity to speech
- Intrinsic temporal window of <500 ms

**An intermediate level between acoustics and linguistic analysis**

- posit a stage/level of acoustic-phonetic analysis that generates intermediate representations prior to linguistic computation
Summary

There is reason to posit a stage/level of acoustic-phonetic analysis that generates intermediate representations prior to linguistic computation.

To investigate the neural basis of this analysis, we used sound quilts – stimuli constructed by shuffling segments of foreign speech, approximately preserving speech properties at short timescales while disrupting them at longer scales. We manipulate the extent of natural speech structure by varying the quilt segment length.

Using fMRI, we identified bilateral regions of the superior temporal sulcus whose responses to speech quilts increased with segment length. This effect did not occur for non-speech quilts, suggesting tuning to speech-specific temporal structure.

When examined parametrically, the response to speech quilts plateaued at segment lengths of ~500 ms. Quilts made from time-compressed speech yielded a similar plateau despite the increase in stimulus structure per unit time.

Our results identify a locus of speech analysis in human auditory cortex – STS, bilaterally - with an intrinsic temporal limit between syllables and words.
The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts

Tobias Overath, Josh H. McDermott, Jean Mary Zarate, David Poeppel

Nature Neuroscience, 2015

Duke  MIT  NPG

Department of Psychology, Center for Neural Science, NYU
Figure 2

A. Tonotopy Measured with Pure Tones

B. Component Voxel Weights Plotted in Anatomical Coordinates

C. Summary Map: Outlines of Regions with High Weight

D. Component Response Profiles to All 165 Sounds Colored by Category

E. Average Component Response to Different Categories
Figure 7. Summary of spatial distribution of voxels' tuning properties.

http://journals.plos.org/ploscompbiol/article?id=info:doi/10.1371/journal.pcbi.1003412
Outline

• Warm-up for specialization in the auditory domain

• A tantalizing hint from neuropsychology

• An fMRI experiment on specialization for speech, in detail

• The modulation of speech
A ~ 200 ms window analyzes the input signal -- The syllable as primitive?

Ohala 1972

Greenberg & Arai 2004

http://www.phonetik.uni-muenchen.de/Bas/BasPHONSTATeng.html

Pellegrino et al. 2011
Speech Modulation Spectrum

A

- **Normalized amplitude**
- Frequency (Hz)

- **Sentences (TIMIT)**
- **Conversations (Switchboard)**
- **Interview (Buckeye)**
- **Audiobooks**

B

- **Normalized amplitude**
- Frequency (Hz)

- **American English**
- **Mandarin Chinese**
- **French**

Single-Instrument Music Modulation Spectrum

A

- **Normalized amplitude**
- Frequency (Hz)

- **Violin**
- **Viola**
- **Cello**
- **Bass**
- **Plano**
- **Guita**

B

- **Normalized amplitude**
- Frequency (Hz)

- **Symphony**
- **Rock**
- **Jazz**
- **Single-Instrument Music**
- **Speech**

*Ding, Patel & Poeppel, rejected 2015 a-c*
A Marr’s eye view: Levels of analysis

| implementational | Hypothesized implementational (neurobiological) infrastructure |
## A Marr’s eye view: Levels of analysis

<table>
<thead>
<tr>
<th>Implementational</th>
<th>Hypothesized <strong>implementational (neurobiological) infrastructure</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithmic</td>
<td>Hypothesized <strong>computational primitives</strong> [domain general]</td>
</tr>
<tr>
<td>Representational</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• constructing spatiotemporal objects (streams, gestures)</td>
</tr>
<tr>
<td></td>
<td>• extracting relative pitch</td>
</tr>
<tr>
<td></td>
<td>• extracting relative time</td>
</tr>
<tr>
<td></td>
<td>• discretization</td>
</tr>
<tr>
<td></td>
<td>• sequencing - concatenation - ordering</td>
</tr>
<tr>
<td></td>
<td>• grouping - constituency - hierarchy</td>
</tr>
<tr>
<td></td>
<td>• establishing relationships - local/long-distance</td>
</tr>
<tr>
<td></td>
<td>• coordinate transformations</td>
</tr>
<tr>
<td></td>
<td>• prediction</td>
</tr>
<tr>
<td></td>
<td>• synchronization - entrainment - turn-taking</td>
</tr>
<tr>
<td></td>
<td>• concurrent processing over different levels</td>
</tr>
</tbody>
</table>
**A Marr’s eye view: Levels of analysis**

<table>
<thead>
<tr>
<th>Implementational</th>
<th>Hypothesized implementational (neurobiological) infrastructure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithmic</td>
<td>Hypothesized computational primitives [domain general]</td>
</tr>
<tr>
<td>Representational</td>
<td>• constructing spatiotemporal objects (streams, gestures)</td>
</tr>
<tr>
<td></td>
<td>• extracting relative pitch</td>
</tr>
<tr>
<td></td>
<td>• extracting relative time</td>
</tr>
<tr>
<td></td>
<td>• discretization</td>
</tr>
<tr>
<td></td>
<td>• sequencing - concatenation - ordering</td>
</tr>
<tr>
<td></td>
<td>• grouping - constituency - hierarchy</td>
</tr>
<tr>
<td></td>
<td>• establishing relationships - local/long-distance</td>
</tr>
<tr>
<td></td>
<td>• coordinate transformations</td>
</tr>
<tr>
<td></td>
<td>• prediction</td>
</tr>
<tr>
<td></td>
<td>• synchronization - entrainment - turn-taking</td>
</tr>
<tr>
<td></td>
<td>• concurrent processing over different levels</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Representational</th>
<th>Hypothesized representational primitives: language [domain specific]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computational</td>
<td>• feature (articulatory)</td>
</tr>
<tr>
<td></td>
<td>• phoneme</td>
</tr>
<tr>
<td></td>
<td>• syllable</td>
</tr>
<tr>
<td></td>
<td>• morpheme</td>
</tr>
<tr>
<td></td>
<td>• noun-phrase, verb-phrase, etc...</td>
</tr>
<tr>
<td></td>
<td>• clause</td>
</tr>
<tr>
<td></td>
<td>• sentence</td>
</tr>
<tr>
<td></td>
<td>• discourse/narrative</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Representational</th>
<th>Hypothesized representational primitives: music [domain specific]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computational</td>
<td>• note (pitch and timbre)</td>
</tr>
<tr>
<td></td>
<td>• pitch interval (consonance/dissonance)</td>
</tr>
<tr>
<td></td>
<td>• octave-based pitch scale</td>
</tr>
<tr>
<td></td>
<td>• pitch hierarchy (tonality)</td>
</tr>
<tr>
<td></td>
<td>• discrete time interval</td>
</tr>
<tr>
<td></td>
<td>• beat</td>
</tr>
<tr>
<td></td>
<td>• meter</td>
</tr>
<tr>
<td></td>
<td>• motif/theme</td>
</tr>
<tr>
<td></td>
<td>• melody/satz</td>
</tr>
<tr>
<td></td>
<td>• piece</td>
</tr>
</tbody>
</table>
Outline  Conclusions

• Warm-up for specialization in the auditory domain
  Screams occupy a special acoustic regime

• A tantalizing hint from neuropsychology
  Pure word deafness suggests double dissociation

• An fMRI experiment on specialization for speech, in detail
  Quilts show sensitivity and specificity for speech in STS

• The modulation of speech
  Speech and music have overlapping but distinct modulations
Thanks to support from NIH, NSF, ARO, AFOSR, Max-Planck Society
Test case: pure word deafness (Poeppel 2001)

Table 2  *PWD patients’ performance on phonetic tasks*

<table>
<thead>
<tr>
<th>Consonant ID</th>
<th>Vowel ID</th>
<th>VOT boundaries</th>
</tr>
</thead>
<tbody>
<tr>
<td>voicing</td>
<td>place</td>
<td></td>
</tr>
<tr>
<td>Saffran et al.</td>
<td>&lt;&lt;&lt;</td>
<td>&lt;</td>
</tr>
<tr>
<td>Auerbach et al.</td>
<td>&lt;</td>
<td>&lt;&lt;&lt;</td>
</tr>
<tr>
<td>Miceli</td>
<td>&lt;</td>
<td>&lt;&lt;&lt;&lt;&lt;</td>
</tr>
<tr>
<td>Tanaka et al.</td>
<td>&lt;&lt;&lt;</td>
<td>&lt;&lt;&lt;</td>
</tr>
<tr>
<td>Yaqub et al.</td>
<td>&lt;</td>
<td>&lt;&lt;&lt;&lt;&lt;</td>
</tr>
<tr>
<td>Praamstra et al.</td>
<td>&lt;&lt;&lt;&lt;&lt;</td>
<td>&lt;&lt;&lt;</td>
</tr>
</tbody>
</table>

Key: < minor impairment; <<< medium impairment; <<<<< severe impairment. ID: identification. VOT: voice onset time judgment.
Individual subjects

Legend:

<table>
<thead>
<tr>
<th>Legend</th>
<th>Left Hemisphere</th>
<th>Right Hemisphere</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>Mean</td>
<td>Mean</td>
</tr>
<tr>
<td>Session 1</td>
<td>Session 1</td>
<td>Session 1</td>
</tr>
<tr>
<td>Session 2</td>
<td>Session 2</td>
<td>Session 2</td>
</tr>
<tr>
<td>Session 3</td>
<td>Session 3</td>
<td>Session 3</td>
</tr>
</tbody>
</table>

Graphs show BOLD response relative to 960 ms localizer across different segment durations for individual subjects.