Speech Production 2

Paper 9: Foundations of Speech Communication

Lent Term: Week 4

Katharine Barden
Today’s lecture

• Prosodic-segmental interdependencies

• **Models of speech production**
  – Articulatory phonology (and task dynamics)
  – DIVA model
  – Levelt’s model of lexical access in speech production
Prosodic-segmental interdependencies

• Within phonetics, ‘segmental’ and ‘suprasegmental’ aspects of speech have traditionally been analysed separately (e.g. IPA chart)

• Suprasegmental
  – (=not confined to any one segment)
  – Prosody (a less problematic term?)
  – Intonation, rhythm, stress
Prosodic-segmental interdependencies

• Prosodic prominence and prosodic grouping affect the production of segments, such as:
  
  – Vowel quality (compare English stressed vs unstressed vowels)
  
  – Magnitude of a glottal gesture e.g. more aspiration for aspirated stops in utterance-initial > phrase-initial > word-initial position

(Cooper, 1991; Pierrehumbert & Talkin, 1992)
Prosodic-segmental interdependencies

- Keating, Cho, Fougeron: ‘Domain-initial articulatory strengthening’
  - Measured by peak linguopalatal contact in electropalatography
  - E.g. More contact for tokens of French /k/ that are utterance-initial > accentual-phrase-initial > word-initial

From Fougeron (2001). Tongue-palate contact during Word-initial, Accent-Phrase-initial and Intonational-Phrase-initial tokens of French /k/
Prosodic-segmental interdependencies

• The phonetic realisation of an individual speech segment will depend in part on that segment’s position in the entire prosodic structure

• How can prosodic-segmental interdependencies be accounted for in models of speech production?
Models of Speech Production…
Task Dynamics & Articulatory Phonology

• **Task dynamics** models how to get from a goal to movements of the articulators.

• **Articulatory Phonology** makes use of the principles of task dynamics within a gestural theory of phonology
Task Dynamics

• Task dynamics describes movement in terms of the tasks to be done
• When a speaker plans the message, the goal is important, but not the details of how it is achieved (the details are automatic)
• It is these automatic details that are modelled by task dynamics
• (See Hawkins 1992 for tutorial)
Task Dynamics

Task Space (abstract)
- An abstract constriction
- Spatial coordinates and degree of an ideal unspecified constriction

Body Space (abstract)
- Vocal tract constriction
- Particular constriction degree or location

Articulator Space (equations of motion)
- Articulators potentially involved
- Movement of individual articulators to achieve target constriction
Task Dynamics

• In articulator space, a **gesture** is a **family of movement patterns** that are functionally equivalent ways of achieving the same goal
  
  e.g. family of combinations of lip and jaw activity that achieve the same degree of lip aperture

• **Coordinative structures**
  
  – functional grouping of articulators, acting in concert to achieve particular VT configurations
  
  – constrain the large number of possible combinations of articulator movements for a given speech task

• **Gestures are made using coordinative structures**
# Coordinative Structures

<table>
<thead>
<tr>
<th>Combination of articulators</th>
<th>Constrictions achieved</th>
</tr>
</thead>
<tbody>
<tr>
<td>lips, jaw</td>
<td>labial</td>
</tr>
<tr>
<td>jaw, tongue tip, tongue body</td>
<td>dental, alveolar, retroflex</td>
</tr>
<tr>
<td>jaw, tongue tip, tongue front/centre</td>
<td>palatal</td>
</tr>
<tr>
<td>jaw, tongue dorsum</td>
<td>velar, uvular</td>
</tr>
<tr>
<td>tongue, pharyngeal wall muscles</td>
<td>pharyngeal</td>
</tr>
</tbody>
</table>
Task Dynamics

• Gesture = movement towards a goal
• For speech, gesture = movement towards the target of a particular constriction, e.g. lip aperture
• A gesture is modelled as the movement of a mass attached to a spring: when the spring’s free end is moved to a target, the mass follows so that the spring returns to its equilibrium position
Task Dynamics

• Task dynamics distinguishes how we control the achievement of a goal from the details of how to achieve it
  – it automates the stage between the production goal and the movements of the articulators

• Task dynamics can account for, e.g.:
  – rapid compensation for unexpected perturbation of an articulator
  – coarticulatory blending resulting from simultaneous demands on an articulator by multiple gestures
Co-production: a sequence of three overlapping gestures

Based on Farnetani (1997: Fig. 12.9)
Articulatory Phonology

- Articulatory Phonology is a phonological theory that has gestures as primitive phonological units

- It uses task dynamic principles to model speech production
Articulatory Phonology

Based on Browman and Goldstein (1990: Fig. 19.1)
Gestural Score

• represents the entire control plan to produce the particular utterance

• specifies for each gesture its:
  – duration
  – location in the sequence of gestures
  – relative timing in the overall plan
Schematic gestural score for ‘pad’

Browman and Goldstein (1992), from Fig. 2
Articulatory Phonology

Based on Browman and Goldstein (1990: Fig. 19.1)
Articulatory Phonology

• Claim: Coarticulation and ‘phonological’ processes can be explained by the relative timing and overlapping of gestures
AP: What is a gesture?

“While gestures are primitive phonological units, they do not correspond to either features or segments. Rather, they sometimes give the appearance of corresponding to features, and sometimes to segments.”

(Browman and Goldstein, 1992)
AP: What is a gesture?

- **bad** would typically be considered to differ from **add** by the presence of a segment.
- **bad** and **pad** would be considered to contrast only in a single feature (voicing).
- In AP, both of these contrasts are conveyed by the presence or absence of a single gesture.
AP: What is a gesture?

Aspiration
- \([p^h]\) feature [+spread glottis] vs. /h/ (segment)
- In AP, both modelled as a single glottal spreading gesture

\[ \text{pad} \]

\[ \text{had} \]
Articulatory Phonology

• ‘Insertion’ may be a change in the timing of gestures
  – e.g. epenthetic stop in *something*
  
  /sʌmθɪŋ/ → [sʌmpθɪŋ]

  If denasalisation precedes the release of the closure, a short period of oral closure will result

• ‘Deletion’ may just be complete overlap
‘Deletion’ as overlapping gestures

List Production
“perfect, memory”

Phrasal Production
“perfec(t) memory”

- (TT: tongue tip; TR: tongue rear; LL: lower lip)
- In the phrasal production: /m/ closure overlaps /t/ release, making it inaudible; /t/ gesture is present nevertheless

Slide courtesy of J. Perkell

U. Tokyo X-ray μ-beam
(Fujimura et al. 1973)
Evaluation of task dynamics and articulatory phonology

• Task dynamics appealing because separates *plan from skill* in the control of movement

• Articulatory phonology appealing because *phonetic units pushed into phonology*
  – no distinction between phonetic and phonological planning
  – gestural score = phonology of the language
    no phonemes, only gestures

• Coarticulation modelled as *coproduction of gestures*
Evaluation of task dynamics and articulatory phonology

- Rate and style changes also accounted for by overlapping of gestures
- More recent work includes ‘prosodic’ gestures that shrink or stretch other gestures at prosodic boundaries
- Systematic and explicit, testable
- However...
Evaluation of task dynamics and articulatory phonology

• Many simplifying assumptions
e.g. masses of different articulators and aerodynamic properties not considered

• Is articulatory phonology really phonology?
  – some phonological processes can’t be described
e.g. vowel alternations: sincere-sincerity
  – can all phonological processes be described in terms of the reorganisation of gestures?
e.g. /r/ insertion in non-rhotic accents:
  law [r] and order, the idea [r] of it
Evaluation of task dynamics and articulatory phonology

• Does articulatory phonology solve the ‘translation problem’?
  – solves, if assume motoric perception (i.e. we can perceive gestural units)
  – If assume auditory representations, need to explain translation of auditory/perceptual objects into gestural ones
Speech Perception: **Direct Realism**

- The listener perceives the speaker’s motoric gestures
- The gestures of direct realism are like those of articulatory phonology

- See Pickett (Chpt 15) for more info, and how Direct Realism compares to Motor Theory
DIVA (Directions Into Velocities of Articulators)

- (Reading: Guenther and Perkell 2004; Guenther 1995)
- A model that uses what is heard and felt to drive what is produced, based on
  - connectionist (‘neural network’) modelling (Guenther)
  - detailed articulatory research (Perkell)
  - aspects of Keating’s window model
DIVA

• Production of speech sounds is governed by targets for what each sound should sound and feel like

• Targets are regions (not points) in a planning space (cf. Keating’s window model)

• The planning space has auditory and orosensory dimensions

• Target regions are learned by matching sounds produced during babbling to auditory experience
DIVA: learning of target regions

a) The first time /i/ is produced, the learned target is simply the configuration of the vocal tract when the sound was produced.

b) The second time /i/ is babbled, the target region is expanded to encompass both vocal tract configurations used to produce the sound.

c) Schematised versions of /i/ and /p/ after many productions of each sound during babbling.

Through matching babbling with auditory experience, the model learns that the target for /i/ allows large variation in lip aperture, where as /p/ requires strict control of this dimension.

Guenther and Perkell (2004), Fig. 2
DIVA

Guenther and Perkell (2004: Fig. 1)
DIVA

Learned auditory and orosensory mappings for each phoneme

Guenther and Perkell (2004: Fig. 1)
Producing a sound involves computing directions from the current position in planning space (where you are now) to the desired position (where you want to be)

Guenther and Perkell (2004: Fig. 1)
The planning directions are converted into movement directions in articulator space

- Because of the degrees of freedom, there are many ways to do this mapping

- DIVA chooses the mapping that requires the least articulatory movement

- Coordinative structures used

Guenther and Perkell (2004: Fig. 1)
Finally, a ‘go’ signal sends the commands for movement directions to the motor neurons controlling muscles.

- the strength of this signal affects when movements begin, and how fast they are

Guenther and Perkell (2004: Fig. 1)
Ongoing articulation is controlled by two kinds of feedback:

- direct auditory and tactile/proprrioceptive feedback (actual sound and feel)
- an ‘efference copy’ of the articulatory commands (expected sound and feel)

Guenther and Perkell (2004: Fig. 1)
Anticipatory: occurs when the target region for the current sound overlaps with the target for the next sound – the region of overlap is used as the target.

The word *coo*. The target regions for /k/ and /u/ overlap on the dimension of lip protrusion. When pronouncing /k/, the target is shrunk to include only the overlapping portion.

Guenther (1995), Fig. 15
DIVA – accounting for coarticulation

- **Carryover**: occurs because the model always calculates the trajectory that requires least movement - it chooses a vocal tract configuration which is as close as possible to the previous configuration.

/k/ target in *Luke vs leak*.

There is a fairly wide range of acceptable tongue body positions for /k/ on the front-back dimension.

Approaching the target for /k/ from /u/ in *Luke* leads to a final tongue position further back than when approaching from /i/ in *leak*.

Guenther and Perkell (2004), Fig. 3
DIVA – accounting for changes in rate and style

- During slow speech, target regions are shrunk to achieve better accuracy, and/or movement velocities are adjusted by changing the strength of the ‘go’ signal.
DIVA – perturbation effects

• Direction-to-velocity nature of the orosensory-to-articulatory mapping means that compensation for perturbation occurs automatically to achieve auditory and orosensory goals
DIVA – articulatory synthesis demos

1. DIVA producing the vowels in "bet", "beet", "bat", "but", and "boot".
2. DIVA producing the same vowels while the jaw is held fixed.
3. DIVA producing the same vowels without using its jaw, lips, or larynx.

Source: http://speechlab.bu.edu/demos.php
DIVA – demos

1. DIVA producing the vowels in "bet", "beet", "bat", "but", and "boot"

Source: http://speechlab.bu.edu/demos.php
DIVA – demos

2. DIVA producing the same vowels while the jaw is held fixed

Source: http://speechlab.bu.edu/demos.php
DIVA – demos

3. DIVA producing the same vowels without using its jaw, lips, or larynx

Source: http://speechlab.bu.edu/demos.php
Evaluation of DIVA

• Auditory feedback = a production-perception link
• More plausible from a learnability perspective
• Use of on-line auditory feedback supported by evidence that auditory feedback affects production:
  – changes in speech targets in response to a change in hearing status (see Guenther and Perkell, 2004)
Evaluation of DIVA

• computationally implemented, therefore explicit and somewhat testable

• input to model normally phonemes
  – influences of metrical and prosodic structure, morphology, word boundaries, etc. will also need to be accounted for
  – in principle, ‘target regions’ could be stretched or shrunk in accordance with prosodic structure, but no details have been modelled
Evaluation of DIVA

• Account of coarticulation is similar to Keating’s window model – similar strengths, and comes closer to explanation because accounts for how targets are learned.

• Arguably achieves its success by relying too much on perceptual system:
  – auditory categories are essentially pre-existing in the model, but how are these learnt?
Levelt’s Model

- Model with origins in psycholinguistics
- Evidence mainly from reaction time experiments (especially ‘naming’ paradigm – time taken to say a word after presentation of a word or picture stimulus). Also from speech errors and patients with brain damage).
- Mainly a theory of lexicalisation (encoding of concepts into word forms)
- Has been implemented as a connectionist model WEAYER++
conceptual preparation in terms of lexical concepts

lexical concept

lexical selection

lemma

morphological encoding

morpheme

phonological encoding

syllabification

phonological word

phonetic encoding

phonetic gestural score

articulation

lemmas
MENTAL LEXICON
word forms

SYLLABARY

Self-monitoring

Based on Harley (2001: Fig. 12.8)
Levelt’s model

Processing is divided into several stages and is serial (one stage must be complete before the next can start)

Based on Harley (2001: Fig. 12.8)
Levelt’s model

First stage: activation of lexical concept, e.g. ESCORT

Based on Harley (2001: Fig. 12.8)
Levelt’s model

Lexical concepts activate word forms indirectly, via an intermediate lemma level which contains syntactic information.

Based on Harley (2001: Fig. 12.8)
Levelt’s model

Once a lemma has been activated and selected, the word form (phonological shape of the word) can be retrieved

Based on Harley (2001: Fig. 12.8)
Levelt’s model

Phonological encoding involves activating:

- the word’s morphological make-up (e.g. <escort> <ing>)
- its metrical structure (no. of syllables, and location of main stress, e.g. $\sigma\sigma$ for <escort>)
- a list of its segments

Based on Harley (2001: Fig. 12.8)
conceptual preparation in terms of lexical concepts

- lexical concept
  - lexical selection
    - lemma
      - morphological encoding
        - morpheme
          - phonological encoding
            - syllabification
              - phonological word
                - phonetic encoding
                  - phonetic gestural score
                    - articulation
                      - sound wave

Based on Harley (2001: Fig. 12.8)

**Levelt’s model**

If the word has more than one morpheme, its morphemes may be combined into a new metrical shape.

Segments are inserted into the metrical shape, one by one, to build up phonological syllables ‘on-the-fly’.

e.g. the segments of the morpheme <escort> form the syllables [ə][skɔt] for escort, but [ə][skɔ][tɪŋ] for escorting.
Levelt’s model

Preparing to articulate a word can only begin when the speaker knows the word’s metrical pattern, its number of syllables, and its segments.
Levelt’s model

The model has little to say about how words are actually articulated, except to propose that frequent syllables in a language are stored in a syllabary as “overlearned” (highly practised) patterns of gestures.

Based on Harley (2001: Fig. 12.8)
Evaluation of Levelt’s model

• Appealing for its broad scope
  – extending from conceptual and syntactic properties of words to some aspects of articulation
• But does not model actual speech movements
• Treatment of concepts not detailed from semantic perspective
• Controversies:
  – Is processing serial?
  – Is the lemma level necessary, or does processing proceed directly from concepts to phonological word forms?
Evaluation of Levelt’s model

• Relatively sophisticated in its treatment of metrical and morphological structure

• The syllables and morphemes that a phoneme participates in affect the time course of planning an utterance, which could affect the articulatory details

⇒ it could help describe how patterns of systematic phonetic variation arise in production, e.g. fine phonetic differences in ‘mis’ of mistakes and mistimes dependent on whether <mis> functions as an affix
Evaluation of Levelt’s model

- However, some vital aspects of phonology and phonetics are left out, such as intonation and its integration with segmental structure.
- The *syllabary* idea is probably too simple to explain actual articulatory data
  - Why should gestural patterns be stored only for syllables?
  - In practice, syllable plans would need contextual modification: coarticulation can spread between syllables and no *one* linguistic unit has been found that constrains its spread.
Summary

The different models have **complementary strengths**: 

- task dynamics’ sophisticated approach to modelling movement  
- DIVA’s use of auditory feedback and control  
- attention to higher-level linguistic structure in Levelt’s model
Next lecture…

• Speech and the brain: introduction