Sensorimotor Control of Speech Production: Models and Data

Joseph S. Perkell

Speech Communication Group
Research Laboratory of Electronics and
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
Cambridge, Massachusetts

perkell@speech.mit.edu
http://www.rle.mit.edu/perkell
Collaborators

Margaret Denny
Frank Guenther
Harlan Lane
Melanie Matthies
Ken Stevens
Ellen Stockmann
Mark Tiede
Majid Zandipour
Jennell Vick

Research supported by the N.I.D.C.D., N.I.H.
Outline

- The control problem
- Measuring Speech
- Vowels and liquids
- Consonants (sibilants)
- Feedback and feedforward control
- Summary
Outline

• The control problem
  – What speech movements look like
  – Possible controlled variables
  – Use of models
  – Schematic view of speech movements

• Measuring Speech
• Vowels and liquids
• Consonants (sibilants)
• Feedback and feedforward control
• Summary of observations
Speech Movements

- Consider the movements of each of these structures
- Approximate number of muscle pairs that move the
  - Tongue: 9
  - Velum: 3
  - Lips: 12
  - Mandible: 7
  - Hyoid bone: 10
  - Larynx: 8
  - Pharynx: 4
- NB: The respiratory system

Observations:
- A large number of degrees of freedom
- A very complicated control problem
A Major Question: What are the controlled variables?

• Objective of Speaker:
  – To produce sounds strings with **acoustic patterns** that result in **intelligible** patterns of **auditory sensations** in the listener

• Acoustic/auditory cues depend on type of sound segment:
  – Vowels and glides:
    • Time varying patterns of formant frequencies
  – Consonants:
    • Noise bursts
    • Silent intervals
    • Aspiration and frication noises
    • Rapid formant transitions

“The yacht was a heavy one”
Possible Controlled Variables

• Auditory characteristics of speech sounds are determined by:
  1. Levels of muscle tension
  2. Changing muscle lengths and movements of structures
  3. The vocal-tract shape (area function)
  4. Aerodynamic events and aeromechanical interactions
  5. The acoustic properties of the radiated sound
• Hypothetically, motor control variables could consist of feedback about any combination of the above parameters
Modeling to make the problem approachable: DIVA

“Directions Into Velocities of Articulators” – Guenther and Colleagues

- A computational model of relations among cortical activity, motor output, sensory consequences

- **Phonemic Goals**: Projections (mappings) from premotor to sensory cortex that encode *expected sensory consequences* of produced speech sounds
  - Correspond to *regions* in multidimensional auditory-temporal and somatosensory-temporal spaces

- With acquisition, control becomes *predominantly feedforward*

- Feedback control – Uses error detection and correction - to teach, refine and update feedforward control mechanisms
Defining Phonemic Goals

- Properties of speakers’ production and perception mechanisms help to define *phonemic goal regions*
- Some of these properties are characterized by *quantal effects* (Stevens), also called *saturation effects* (Fujimura and Kakita)

![Diagram](image)

- Schematic example: A continuous change in an articulatory parameter produces two regions of acoustic stability, separated by a rapid transition

- Languages prefer such stable regions
- Individual speakers can use such regions to help produce robust acoustic cues with relatively imprecise motor commands
Speech Movements: A schematic view

• Planned and actual acoustic trajectories illustrate:
  – Auditory/acoustic goal regions
  – Economy of effort (Lindblom)
  – Coarticulation
  – Motor equivalence
  – Biomechanical saturation (quantal) effects
• When controlling an articulatory synthesizer, DIVA accounts for the first four and
  – Aspects of acquisition
  – Responses to perturbations
Outline

• The control problem
• Measuring Speech
• Vowels and liquids
• Consonants (sibilants)
• Feedback and feedforward control
• Summary of observations
Measuring Speech Production

- Acoustics – important for perception
- Spectral, temporal and amplitude measures
  - Vowels, liquids and glides:
    - Time varying patterns of formant frequencies
  - Consonants:
    - Noise bursts
    - Silent intervals
    - Aspiration and frication noises
    - Rapid formant transitions

"The yacht was a heavy one"

- Movements
  - Points on the tongue, lips, jaw, (velum)
  - Use an Electro-Magnetic Midsagittal Articulometer (EMMA) System
• Transducer coils are placed on subject’s articulators
• Subject reads text from an LCD screen
• Movement and audio signals are digitized and displayed in real time
• Signals are processed and data are extracted and analyzed
Analysis of EMMA data

- Algorithmic data extraction at time of minimum in absolute velocity during the vowel:
  - Vowel formants
  - Articulatory positions (x, y)
Outline

• The control problem
• Measuring Speech
• Vowels and liquids – properties of goal regions
  – Saturation effects
  – Motor equivalence
  – Changes related to hearing status
  – Relations between production and perception
• Consonants (sibilants)
• Feedback and feedforward control
• Summary of observations
An acoustic saturation effect for constriction location for /i/

- Over a range (green) of back cavity lengths, F1-F3 are relatively stable (Stevens)
- Many repetitions of /i/ in two subjects: Considerable variation of constriction location; however,
- Formants of /i/ are sensitive to variation in constriction degree

X-ray Microbeam data (Perkell & Nelson)
A biomechanical saturation effect for constriction degree for /i/

- Constriction degree and resulting formants can be *stabilized*
  - Stiffening the tongue blade
  - Pressing the stiffened tongue blade against the sides of the hard palate through contraction of the posterior genioglossus (GGp) muscles

- Constriction area (shaded) varies little, even with variation in GGp contraction (from a 3D tongue model by Fujimura & Kakita)
Stabilizing the sound output for the vowel /u/: Motor Equivalence

- Hypothesis: negative correlation between tongue-body raising and lip protrusion in multiple repetitions of the vowel
- Hypothesis is supported in a number of subjects
- The goal for the articulatory movements for /u/ is in an acoustic/auditory frame of reference, not a spatial one
- Strategy: Stay just within the acoustic goal region
Motor equivalence for /r/

- Two kinds of /r/ in American English

- Speakers use articulatory trading relations when producing the goal for /r/ (low F3) in different phonetic contexts (Guenther et al)
  - Acoustic effect of the longer front cavity (retroflexed) is compensated for by the effect of the longer and narrower constriction (bunched)
  - F3 variability is greatly decreased by these articulatory trading relations
- The goal for /r/ is auditory/acoustic
Learning and maintaining phonemic goals: Use of Auditory Feedback

- Audition is crucial for normal speech acquisition
- Postlingual deafness: Intelligible speech, but with some abnormalities
- Regain some hearing with a Cochlear Implant (CI):
  - Usually show parallel improvements in perception, production and intelligibility

Acoustic measures of contrast between /l/ and /r/
6 months after receiving a CI

- Phonemic contrast is enhanced pre- to post-implant – typical for CI users, many of whom have somewhat diminished contrasts pre-implant
Long-term stability of auditory-phonemic goals for vowels

- Typical pre- (○) and post- (□) implant formant patterns: generally congruent with normative data (□)
  - FA: some irregularity of F2 pre-implant (18 years after onset of profound hearing loss)
  - One year post-implant: F2 values are more like normative ones
- Phonemic identity doesn’t change; degree of contrast can
- Goals for vowels generally are stable
  - If they degrade from hearing loss, can be recalibrated with hearing from a CI

Data from 2 cochlear implant users
Vowel Contrasts and Hearing Status

• Compare English with Spanish CI users, CI processor OFF and ON
• Previous findings: Contrasts increase with hearing, decrease without
• Hypothesis: Because of the more crowded vowel space in English, turning the CI processor OFF and ON will produce more consistent decreases and increases in vowel contrasts in English than in Spanish

- Average Vowel Spacing (AVS) – a measure of overall vowel contrast
- Change of AVS from processor ON to processor OFF (for 24 hours)
  – AVS: decreases for the English speaker, increases for the Spanish speaker
AVS – by subject

- Prediction: AVS increases with the CI processor ON (hearing)
- Changes follow the predicted pattern more consistently for English than Spanish speakers
Modeling Contrast Changes: Clarity vs. Economy of Effort

- DIVA contains a parameter that changes sizes of all goal regions simultaneously – to control speaking rate and clarity (e.g., AVS)

- Shrinking goal region size – like what English speakers do with hearing
  - Produces *increased clarity* (contrast distance), *decreased dispersion*
  - Without hearing, *economy of effort* dominates

- With fewer vowels in Spanish, clarity demands aren’t as stringent
  - Acceptable contrasts may be produced regardless of hearing status, without changing goal region size
Further Insights into Goal Regions for Vowels: Relations between Production and Perception

• Close linkage between production and perception:
  – Speech acquisition, with and without hearing
  – Speech of Cochlear Implant users
  – Second-language learning (e.g., Bradlow et al.)
  – Focused studies of production & perception (e.g., Newman)
  – Mirror neurons – a more general action-perception link

• Hypothesis:
  – Speakers who discriminate well between vowel sounds with subtle acoustic differences will produce more clear-cut sound contrasts
  – Speakers who are less able to discriminate the same sound stimuli will produce less clear-cut contrasts
Production Experiment

- **Data Collection**
  - Subjects: 19 young-adult speakers of American English
  - For each subject:
    - Recorded articulatory movements and acoustic signal
    - Subject pronounced “Say ___ hid it.”;
      ___ = cod, cud, who’d or hood
    - Clear, Normal and Fast conditions

- **Analysis**
  - Calculated **contrast distance** for each vowel pair:
    - Articulatory (TB) *contrast distance*: distance in mm between the centroids of the cod and cud TB distributions
    - Acoustic *contrast distance*: distance in Hz between centroids of F1, F2 distributions for cod, cud
Perception Experiment

• Methods
  – Synthesized natural-sounding stimuli in 7-step continua – for *cod-cud, who’d-hood*
  – Each subject: Labeling and discrimination (ABX) tasks

• Results: ABX scores (2-step)
  – Ceiling effects: some 100% subjects probably had better discrimination than measured
  – For further analysis divide subjects into two groups:
    • **HI** discriminators - at 100% (above the median)
    • **LO** discriminators - (at median and below)
Results & Conclusions

• HI discrimination subjects produced greater contrast distance than LO discrimination subjects (measured in articulation or acoustics)

• The more accurately a speaker discriminates a vowel contrast, the more distinctly the speaker produces the contrast

* Difference between HI and LO groups is significant at p < .001
A Possible Explanation

- It is advantageous to be as intelligible as possible
- Children will acquire goal regions that are as distinct as possible
  - Speakers who can perceive fine acoustic details learn *auditory goal regions* that are *smaller* and *spaced further apart* than speakers with less acute perception, because
  - The speakers with more acute perception are more likely to reject poorly produced tokens when learning the goal regions
Outline

• The control problem
• Measuring Speech
• Vowels and liquids – properties of goal regions
• Consonants
  – Stability of goal regions for sibilants
  – Changes with hearing and motor loss
  – Relations between production and perception
  – Unexpected short-term drift of /ʃ/
• Feedback and feedforward control
• Summary of observations
A saturation effect for /s/ may help define the /s-ʃ/ contrast

- Production of /ʃ/ (as in “shed”)
  - Relatively long, narrow groove between tongue blade and palate
  - Sublingual space
- Production of /s/ (as in “said”)
  - Short narrow groove
  - No sublingual space
- Saturation effect for /s/
  - As tongue moves forward from /ʃ/, sublingual cavity volume decreases
  - When tongue contacts lower alveolar ridge, sublingual cavity is eliminated, resonant frequency of anterior cavity increases abruptly
  - After contact, muscle activity can increase further; output is unchanged
Long-term stability of phonemic goals for sibilants in CI users

- Subjects 1 and 3: good distinctions between /s/ and /ʃ/ pre-implant –
  - Typical, decades following onset of hearing loss
- Subject 2: reversed values and distorted productions pre-implant
  - After about 6 months of implant use, sibilant productions improved
- These precisely differentiated articulations are usually maintained for years without hearing
  - Possibly because of the use of somatosensory goals – e.g. pattern of contact between tongue, teeth and palate
Responses to abrupt changes in hearing and motor innervation

An NF2 patient with sudden hearing loss, followed by some motor loss

- Two surgical interventions
  - OHL: Onset of a significant hearing loss (especially spectral) from removal of an acoustic neuroma
  - Hypoglossal nerve transposition surgery → Some tongue weakness

- /s-/j/ contrast: Good until second surgery, when contrast collapsed

- Hypothesis: Feedforward mappings invalidated by transposition surgery
  - Without spectral auditory feedback, compensatory adaptation (relearning) was impossible – as might be possible with hearing
  - Somatosensory goal deteriorated without auditory reinforcement

Spectral median for /s/ and /j/ vs. weeks in an NF2 patient
Relations Between Production and Perception

• Hypothesis: The sibilants, /s/ and /ʃ/, have two kinds of sensory goals:
  – Auditory: particular distribution of energy in the noise spectrum
  – Somatosensory: e.g., patterns of contact of the tongue blade with the palate and teeth

• Speakers will vary in their ability to discriminate /s/ from /ʃ/
• Speakers use contact of the tongue tip with the lower alveolar ridge for /s/ to help differentiate /s/ from /ʃ/
  – This will also vary across speakers
• Across speakers, both factors, ability to discriminate auditorily between the two sounds and use of contact (a possible somatosensory goal), will predict the strength of the produced contrast
Methods (with the same 19 subjects as the vowel study)

• Production experiment – each subject:
  – Recorded: acoustic signal, and contact of the under side of the tongue tip with the lower alveolar ridge - with a custom-made sensor
  – Subject pronounced, “Say___ hid it.”; ___ = sod, shod, said or shed”
  – Clear, Normal and Fast conditions

• Analysis – calculated:
  – Proportion of time contact was made during the sibilant interval
  – Spectral median for /s/ and /ʃ/
  – Acoustic contrast distance:
    • Difference in spectral median between /s/ and /ʃ/

• Perception experiment - each subject:
  – Labeled and discriminated (ABX) between synthesized stimuli from a seven-step said to shed continuum
Production results: Use of tongue-to-lower-ridge contact

− 12 subjects (left of vertical line) are classified as Strong (S) for use of contact difference (c) between /s/ and /ʃ/
− The remaining subjects are classified Weak (W) for use of contact difference
Discrimination results

- 2-step ABX Scores
  - Nine subjects had percent correct = 100; categorized as HI discriminators (right of line)
  - 10 subjects had percent correct < 100; categorized as LO discriminators
Produced contrast distance is related to

- Ability to discriminate the contrast
  * difference is significant, $p < .01$
- Use of contact difference

• Interactions
  - Speakers with good discrimination and use of contact difference: best contrasts
  - Speakers with one or the other factor: intermediate contrasts
  - Speakers with neither factor: poorest contrasts
Rapid drift of spectral median for /ʃ/

- A CI “on-off” experiment

- Observations
  - Vowel SPL increased rapidly with CI processor off, decreased with processor on
  - Spectral median drifted upward toward /s/ during the 1000 seconds with processor off – Surprising, since the goals are usually stable
  - Hearing one aberrant utterance when the processor was turned on, speaker overcompensated to restore an appropriate /ʃ/
  - Extremely narrow dental arches (and movement transducer coil on tongue) may have made it difficult for speaker to rely on somatosensory goal
  - He may have had to rely predominantly on auditory feedback to maintain feedforward control on an utterance-to-utterance basis
Outline

• The control problem
• Measuring Speech
• Vowels and liquids – properties of goal regions
• Consonants – sibilants
• Feedback and feedforward control
  – Feedback and jaw movement perturbations
  – Inaudible gestures and feedforward control
  – Sensorimotor adaptation
• Summary of observations
Responses to unanticipated jaw perturbations

(In collaboration with David Ostry)

• Robot used to perturb jaw movements
  – Triggered by downward movement
• 50 repetitions/utterance e.g., “see red”
  – 5 perturbed upward (resistive)
  – 5 downward (assistive)

• Formants begin to recover 60-90 ms after perturbation; jaw does not

• Two other subjects were similar
• Evidence of within-movement, closed-loop error correction
Persistence of inaudible gestures

U. Tokyo X-ray µ-beam (Fujimura et al. 1973)

List Production
“perfect, memory”

Phrasal Production
“perfec(t) memory”

- Phrasal: /m/ closure overlaps /t/ release, making it inaudible; /t/ gesture is present nevertheless (c.f. Browman & Goldstein; Saltzman & Munhall)
- Findings replicated and expanded with 21 speakers
- Explanation (DIVA): Frequently used phonemes, syllables, words become encoded as feedforward command sequences
Sensorimotor adaptation

- Thesis project of Virgilio Villacorta
  - Based on work of Houde & Jordan, but with voiced vowels

- Subjects partially compensate by shifting F1 in opposite direction; Shift is formant-specific
- Mismatch between expected and produced auditory sensations → Error correction
- 20 subjects: Varied in amount of compensation
- Next step: A relation between perceptual acuity and amount of compensation?
Feedforward vs. Feedback Control and Error Correction

• In DIVA, feedback and feedforward control operate simultaneously; feedforward usually predominates
• Feedback control intervenes when there is a large enough mismatch between expected and produced sensory consequences (sensorimotor adaptation results)
• Timing of correction:
  – With a long enough movement, correction is expressed (closed loop) during the movement (e.g., “see red”)
  – Otherwise, correction is expressed in the feedforward control of following movements (e.g., /ʃ/ when CI turned on)
• Additional Examples of error correction
  – Closed-loop responses to perturbations (see Abbs, others)
  – Feedforward error correction with, e.g., dental appliances
  – Responses to combined perturbations (cf. Honda & Murano)
  – All are compatible with DIVA’s use of feedback
Outline

• The control problem
• Measuring Speech
• Vowels and liquids – properties of goal regions
• Consonants – sibilants
• Feedback and feedforward control
• Summary of observations
Summary of Observations from a Modeling Perspective

• Highest level control variables for phonemic movements
  – Auditory-temporal and somatosensory-temporal goal regions
• Goal regions encoded in CNS
  – Projections (mappings) from premotor to sensory cortex:
    Expected sensory consequences of producing speech sounds
• Goal regions defined partly by articulatory and acoustic saturation effects
  – Most vowels: goals primarily auditory; saturation effects, acoustic
  – Consonants: both auditory and somatosensory goals; saturation effects, primarily articulatory (e.g., any consonant closure)
• Articulatory-to-acoustic motor equivalence (/u/, /r/)
  – Help stabilize output of certain acoustic cues
  – Evidence that goals are auditory
Summary (continued)

• Auditory feedback (CI users)
  – Used to acquire goals
  – Needed to maintain appropriate motor commands with vocal-tract growth, perturbations

• Goals are usually stable, even with hearing loss

• Clarity vs. economy of effort
  – Tradeoff evident when hearing (CI) is turned on, off

• Relations between production and perception
  – Better discriminators produce more distinct sound contrasts
  – Better discriminators may learn smaller, more distinct goal regions

• Feedback and feedforward control
  – Frequently used sounds (syllables, words) are encoded as feedforward commands
  – Responses to perturbations: intra-gesture are closed loop; inter-gesture are via adjustments to subsequent feedforward commands
• Components have hypothesized correlates in cortical activation
• Hypotheses can be tested with brain imaging
• Can quantify relations among phonemic specifications, cortical activity, movement and the speech sound output
THANK YOU
REFERENCES


Not all speakers exhibit motor equivalence for /u/.

- Individual anatomy may influence the implementation of phonemic goals and movement planning.