An MRI-based articulatory and acoustic study of American English liquid sounds /r/ and /l/
Contents

• Introduction
• Acoustic modeling of “retroflex” /r/ and “bunched” /r/
• Acoustic modeling of dark /l/ and light /l/
• Acoustic variability and discriminative power of liquids for speaker recognition
• Summary and future work
Introduction
Schematic of human speech production system

Adapted from http://www.indiana.edu/~hlw/PhonUnits/vowels.html
Properties of the liquids

- Production of /r/ usually involves two constrictions and one narrowing, and a large front cavity.

- Production of /l/ usually involves an occlusion of air flow (supralingual constriction) and air flow around one or both sides of the tongue (lateral channels).
Properties of the liquids

- Usually /r/ and /l/ have similar pattern in the first two formants (F1 and F2). However /r/ has a low F3 (below 2000 Hz), and /l/ has a high F3 (above 2500 Hz).
Why study the liquids?

- Liquids are considered to be the most difficult sounds to learn both for children and for adult English learners (Shriberg and Kent, 1982). Problems with /r/ alone can account for as much as 60% of the typical school-based clinician's caseload (Creaghead et al., 1989).

- Words containing /r/ and /l/ are frequently the source of errors in automatic speech recognition system (Espy-Wilson, 1992).

- Compared to vowels or obstruent consonants, the acoustics and vocal tract models of /r/ and /l/ are less studied.

- Compared to other sounds, /r/ and /l/ have more variable and more complex articulatory configurations across speakers.
Objectives of this study

• To better understand the acoustics and articulation of the liquid sounds in American English.

• To study the acoustic variability and the speaker discriminative power of the liquids.
Acoustic modeling of “retroflex” /r/ and “bunched” /r/
Acoustic modeling of /r/

- Delattre and Freeman (1968) divided the tongue shapes for /r/ into eight types (Types 2-7 for American English /r/).
- Two maximally distinct types: “retroflex” and “bunched”.

- Low F3 is the most salient acoustic feature for /r/. But no distinct formant pattern for F1, F2 and F3 to distinguish the "bunched" /r/ and "retroflex" /r/.
- Higher formants suggested as potential cues to tongue configuration (Espy-Wilson, 2004; Espy-Wilson and Boyce, 1999).
Two subjects in UC database  (Tiede et al. 2004)
Two Subjects

Retroflex /r/ (S1)  Bunched /r/ (S2)

Tip up  Dorsum down  Tip down  Dorsum up
Differences

• A shorter and more forward palatal constriction in the retroflex /r/.
• A larger volume of the back cavity in the retroflex /r/.
• Sharper transition between the palatal constriction and its anterior and posterior cavities for the retroflex /r/.
Retroflex /r/  
Bunched /r/

Midsagittal MR slices

Spectra of sustained sound

Spectrogram of ‘warav’
3D reconstruction of vocal tract

Sagittal view

Retroflex /r/

Bunched /r/

Axial view (view from the top)

Lips

Glottis

Lips

Glottis
Finite element analysis of 3D vocal tract

Harmonic analysis using finite element method (FEM):

- Helmholtz equation

\[
\nabla . \left( \frac{1}{\rho_0} \nabla p \right) + \frac{\omega^2 p}{\rho_0 c_s^2} = 0
\]

- Boundary conditions

  - Glottis: Normal velocity as sinusoidal signal at different frequency
  - Lips: Radiation impedance of an ideal piston or \( p = 0 \) (Pressure release)
  - Wall: Rigid
Harmonic analysis of 3D vocal tract using FEM

- Pressure isosurfaces at different frequencies

400 Hz

1500 Hz

6000 Hz
Area function extraction based on pressure isosurface

Retroflex \(/r/:\)

500 Hz

Grid line for area function extraction

Bunched \(/r/:\)

500 Hz

Grid line for area function extraction
The retroflex /r/

(a) Area function

(b) Acoustic response from 3D FEM

(c) Acoustic response based on area function

(d) Spectrum of sustained /r/ utterance
The bunched /r/

(a) Area function

(b) Acoustic response from 3D FEM

(c) Acoustic response based on area function

(d) Spectrum of sustained /r/ utterance
The retroflex /r/
The bunched /r/

(a) Area function

(b) Acoustic response from 3D FEM

(c) Acoustic response based on area function

(d) Spectrum of sustained /r/ utterance
Acoustic sensitivity function of the formants (Fant et al., 1974)

- The acoustic sensitivity $S_n$ of one specific formant $F_n$ used to analyze the formant-cavity affiliation.

- Formant change with area perturbation.

\[
\frac{\Delta F_n}{F_n} = \sum_{i=1}^{N} S_n(i) \frac{\Delta A_i}{A_i}
\]

$n$ is the formant number, and $i$ is the section number of the vocal tract area function $A$. 
The retroflex /r/

Area function

\[ \text{Area (cm}^2\text{)} \]

Distance from the glottis (cm)

F1

Sensitivity

F2

Sensitivity

F3

Sensitivity

F4

Sensitivity

F5

Sensitivity

Area function

Distance from the glottis (cm)

Helmholtz resonator

\[ F2 = \frac{c}{2\pi} \sqrt{\frac{A_i}{L_i L_f A_f}} \]
The bunched /r/

Area function

Distance from the glottis (cm)

Area (cm²)

Sensitivity

F1

F2

F3

F4

F5
Four element simple-tube model of the retroflex /r/

[Graph showing area function and acoustic response]
Four element simple-tube model of the retroflex /r/

- Four element model

- F2 comes from the front part like a Helmholtz resonator.

- F1 comes from the back cavity as a Helmholtz resonator formed by the palatal constriction and the tube behind it.
• F3, F4 and F5 are half-wavelength resonances of the cavity posterior to the palatal constriction fairly evenly spaced.

Half-wavelength tube

\[ f_n = \frac{c}{2L_b} n \]

Spacing between formants

\[ \frac{c}{2L_b} = \frac{c}{2(12)} = 1460\text{Hz} \]

Acoustic response
Seven element simple-tube model of the retroflex /r/

Area function

Distance from glottis (cm)

Acoustic response

Frequency (Hz)
Three element simple-tube model of the bunched /r/
• Three element model

- F2 comes from the front part like a Helmholtz resonator.

- When decoupled, the back cavity will act as a quarter-wavelength tube producing F1, F3, F4 and F5.

Quarter-wavelength tube

\[ f_n = \frac{c}{2L_b} (2n + 1) \]

Spacing between formants

\[ \frac{c}{2L_b} = \frac{c}{2(15.3)} = 1150Hz \]
Eight cavity simple-tube model of the bunched /r/

Area function

Distance from glottis (cm)

Acoustic response

Acoustic response (dB)

Frequency (Hz)
Four additional subjects (S3, S4, S5, S6)
Analysis of F4 and F5 for retroflex /r/
Analysis of F4 and F5 for bunched /r/

**Sustained /r/ utterances**

- **S2**: F5-F4 ≈ 700 Hz
- **S5**: F5-F4 ≈ 500 Hz
- **S6**: F5-F4 ≈ 600 Hz

**Dynamic speech (‘warav’)**

- **S2**: F5-F4 ≈ 700 Hz
- **S5**: F5-F4 ≈ 900 Hz
- **S6**: F5-F4 ≈ 600 Hz
Summary of /r/ acoustic modeling

- Both retroflex /r/ and bunched /r/ produce similar formant patterns in F1, F2, and F3.

- Both retroflex /r/ and bunched /r/ produce zeros above 5000 Hz due to cross modes at front cavity.

- Retroflex /r/ and bunched /r/ differ in F4 and F5 pattern. The spacing between F4 and F5 in the retroflex /r/ is much larger than in the bunched /r/ (around 1400 Hz vs. around 700 Hz).

- In both /r/’s, F2 is produced by the front cavity.
- For retroflex /r/, the palatal constriction is made with the tongue tip sharper area change around palatal constriction.
  - Palatal constriction decouples the back and front cavity.
  - F4 and F5 are mainly produced by the back cavity as half wavelength resonances.

- For bunched /r/, the palatal constriction is made with the tongue dorsum more gradual area change around palatal constriction.
  - Difficult to decouple the back and front cavity.
  - F4 and F5 are mainly produced by the back cavity as quarter wavelength resonances with area perturbation

- The acoustic data from several other speakers further validates the result on F5-F4.
Acoustic modeling of dark /l/ and light /l/
Articulation of /l/

• The number of lateral channels, linguo-alveolar contact and the tongue shape are the main concerns in articulatory configurations of /l/.

• The acoustic effects of these geometric features are not clearly understood.
Acoustics of /l/

• Usually /l/ has relatively weak energy in the F3-F5 region.

• It has been proposed that this weak energy is due to the pole-zero pair created by the lateral channels and/or the supralingual space (Fant, 1970; Prahler, 1998; Stevens, 1998; Zhang and Espy-Wilson, 2004).

• Source of zeros in /l/ production are the main focus in this study. It is based on the 3-D acoustic analysis.
The subject in the UC database
(Tiede, et al., 2004)
Midsagittal MR slice of the dark /L/ (S2)

Midsagittal MR slice of a light /L/ (S2)
Midsagittal MR slice of the dark /l/

Coronal MR slices

1
Two lateral channels without sublingual space

2
1 pathway over the tongue

3
1 pathway over the tongue

4
1 pathway over the tongue

Lateral channels
Midsagittal MR slice the light /l/

Coronal MR slices

1. Two lateral channels with sublingual space
2. Three pathways: over the tongue & 2 lateral channels
3. Two pathways
4. One pathway over the tongue

Linguopalatal contact
Supralingual space
Linguopalatal contact
Lateral channels
Sublingual space
Linguopalatal contact
Lateral channels
3D reconstruction of vocal tract

Sagittal view
- Dark /l/: Glottis, Lips, Lateral channels, Luugo alveolar contact
- Light /l/: Glottis, Linguopalatal contacts, Supralingual space, Lateral channels

Axial view
- Dark /l/: Glottis, Lips, Lateral channels, Linguo alveolar contact
- Light /l/: Glottis, Linguopalatal contacts, Supralingual space, Lateral channels
(b) Spectrum of sustained /l/ utterance

(a) Acoustic response from 3D FEM
Dark /l/

Sustained dark /l/ in "pole"

/l/ in "feel"

/l/ in "mole"
Light /l/

Sustained light /l/ in “Lee”

/l/ in “low”

/l/ in “light”
Wave propagation at different frequencies for the dark /l/

At 500 Hz

Sagittal view

Axial view (view from the top)

At 4000 Hz (cross mode posterior to the contact)
Wave propagation at different frequencies for the light /l/

Sagittal view

Axial view (view form the top)

At 500 Hz

At 2350 Hz
Wave propagation at different frequencies for the light /l/

At 2950 Hz
(the zero with supralingual cavity as a side branch)

At 4490 Hz
(the zero with cross mode posterior to the contact)
Schematics of area function vocal tract models

The dark /l/

The light /l/
Area function extraction of the dark /l/

Grid lines for area function extraction

Area function of back cavity, channel 2 and front cavity

Area function of channel 1

Acoustic response (dB)

Frequency (Hz)
Midsagittal MR slice of the dark /l/

Coronal MR slices

1. Two lateral channels without sublingual space
2. 1 pathway over the tongue
3. 1 pathway over the tongue
4. 1 pathway over the tongue

Lateral channels
Area function extraction of the dark /l/

Grid lines for area function extraction
Area function extraction of the light /l/: Method 1
Area function extraction of the light /l/: Method 2
Simple 3D vocal tract model I

H: 1.4 cm, W: 2.8 cm, L: 18 cm, T: 1 cm, block width: 1.4 cm, block starting location: 4.8 cm from the outlet.
Simple 3D vocal tract model II

- Sagittal view
  - H: 1.4 cm, W: 2.8 cm, L: 18 cm
  - Block width: 1.4 cm
  - Block starting location: 5 cm from the outlet

- View from the top
  - Symmetrical lateral channels

Amplitude response (dB) vs. Frequency (Hz)
- H/5
- 2H/5
- 3H/5
- 4H/5
- H
Simple 3D vocal tract model II

- **Sagittal view**
  - 4 cm
  - Outlet
  - Inlet
  - Variable height $h$

- **View from the top**
  - Asymmetrical Lateral channels

- Dimensions:
  - $H$: 1.4 cm, $W$: 2.8 cm, $L$: 18 cm, block width: 1.4 cm, block starting location: 5 cm from the outlet

- Frequencies:
  - 3340 Hz
  - 3480 Hz
  - 4630 Hz
  - 48630 Hz
Pressure isosurfaces for asymmetrical channels (h=H)

(a) At 3340 Hz (zero)

(b) At 3480 Hz (pole)
Acoustic responses for asymmetrical channels with different lengths ($h=H$)
Summary of /l/ acoustic modeling

• The dark /l/ and the light /l/ have similar patterns in F1-F3, but differ in the number and locations of the zeros in spectrum below 6000 Hz.

• For the dark /l/, the zero below 6000 Hz is produced by the cross mode posterior to the linguo-alveolar contact.

• For the light /l/, the three zeros below 6000 Hz are produced by the asymmetrical channels, the supralingual cavity as a side branch and the cross mode posterior to the linguo-alveolar contact.
Summary of /l/ acoustic modeling

- The simple vocal tract models show that, in order to get a zero in the region of F3-F5, the lateral channels have to be asymmetrical and about 3-6 cm long. In addition, a narrow constriction or a complete closure is also required.
Acoustic variability and discriminative power of liquids for speaker recognition
Approach

• Perform ANOVA (Analysis of variance) analysis on the Buckeye database (Pitt et al. 2005).

• Perform speaker identification experiments based on statistical Gaussian mixture models (GMMs) for the discriminative power of the phonemes.
Buckeye Database (Pitt et al. 2005)

• A corpus of spontaneous American English speech with 307,000-word
• 40 speakers from central Ohio, USA (half male and half female)
• About 30-60 minutes of conversation speech for each speaker
• Sampling rate >16 kHz
• Phonemically labeled
Analysis of variance (ANOVA)

- Acoustic parameters: mel-frequency filter bank (MFB) energy (31 coefficients)

- The inter/intra-speaker variance ratio (F-ratio)

  F-ratio definition 1
  \[ F = \text{trace} \left\{ S_w^{-1} S_b \right\} \]

  F-ratio definition 2
  \[ F = \frac{\text{trace} \left\{ S_b \right\}}{\text{trace} \left\{ S_w \right\}} \]

  Inter-speaker variability
  \[ \text{trace} \left\{ S_b \right\} \]

  Intra-speaker variability
  \[ \text{trace} \left\{ S_w \right\} \]

  Inter-speaker scatter matrix
  \[ S_w = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \left( X_{ij} - \overline{X}_i \right) \left( X_{ij} - \overline{X}_i \right)^T \]

  Intra-speaker scatter matrix
  \[ S_b = \sum_{i=1}^{N} n_i \left( \overline{X}_i - \overline{X} \right) \left( \overline{X}_i - \overline{X} \right)^T \]

  \[ \overline{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij} \quad \overline{X} = \frac{1}{N} \sum_{i=1}^{N} \overline{X}_i \]

  \( \overline{X} \): the vector of acoustic parameter
  \( n_i \): the number of tokens of speaker i
  \( N \): the number of speaker
F-ratios

• The rank of average F-ratio 1 in descending order is:
  Nasals > vowels > /r/ > /l/ > glides > fricatives > affricates > stops

• The rank of average F-ratio 2 in descending order is:
  Nasals > /r/ > /l/ > glides > vowels > fricatives > affricates > stops
Inter- and intra-speaker variability

• The rank of average inter-speaker variability in descending order is:
  /r/ > /l/ > vowels > glides > nasals > fricatives > affricates > stops

• The rank of average intra-speaker variability in descending order is:
  affricates > stops > fricatives > /r/ > vowels > /l/ > glides > nasals
F-ratio of each MFB coefficients

• /r/ and /l/ have the maximum F-ratios in the range of coefficients 17 to 25, which have the center frequencies between 2 and 4.2 kHz. This can also be observed in many other sounds such as vowels and nasals.
Speaker identification experiment

• Close set speaker identification task (40 speakers) for each phoneme

• 75% of the tokens in the training set, and 25% of the tokens in the test set

• Speaker model: 512 Gaussian Mixture Model (GMM)

• Universal Background Model (UBM) trained for speaker model adaption
Speaker identification experiment result

- The rank of average identification accuracy in descending order is:
  
  /r/ > glides > /l/ > affricates > fricatives > stops > nasals > vowels
Comparison of speaker recognition accuracy between two different sets of MFB coefficients

- The average identification accuracy is 37% for the first 11 MFB coefficients, and 41% for the last 11 MFB coefficients.
Comparison of speaker recognition accuracy between two different sets of MFB coefficients

- The average identification accuracy is 42% for the first 16 MFB coefficients, and 47% for the last 16 MFB coefficients.
Summary of acoustic variability and discriminative power

• The inter-speaker variability of /r/ is larger than any other phonemes except /sh/, /s/ and /zh/. On average, /r/ and /l/ have larger inter-speaker variability than any other broad phonetic class.

• The F-ratio value of liquids depends on how it is defined. On average, liquids have larger F-ratios than glides, stops, affricates, and fricatives, but smaller than nasals.
Summary of acoustic variability and discriminative power

• The ranking of the average discriminative power for liquids and other broad phonetic classes is:

/r/ > glides > /l/ > affricates > fricatives > stops > nasals > vowels

• The F-ratio measure is not a consistent indicator of speaker recognition performance. This is probably due to the difference between the one Gaussian mixture assumption in the F-ratio calculation and the multiple Gaussian mixture model assumed in the speaker identification task.
Summary and future work
Summary

- Two articulatory configurations for each liquid sound were studied (a "retroflex" /r/ vs. a "bunched" /r/, and a light /l/ vs. a dark /l/).
- Based on MR images, FEM analysis has been performed for acoustic analysis of the 3-D vocal tract models.
- The retroflex /r/ and the bunched /r/ show similar patterns of F1-F3 but very different spacing between F4 and F5. This can be explained by the difference in the palatal constriction between them.
- The dark /l/ and the light /l/ have similar patterns in F1-F3, but differ in the number and locations of zeros in the spectrum. This can be explained by the difference in the linguo-alveolar and linguopalatal contacts between them.
- On average, /r/ and /l/ have larger inter-speaker variability than any other broad phonetic class, and have larger discriminative power than any other broad phonetic class except glides.
Future work

• Vocal tract modeling of intermediate tongue shapes for /r/ and /l/ in the UC database

• 3-D tongue model

• Vocal tract dynamics based on dynamic MR imaging

• Automatic segmentation of the 3-D vocal tract

• Superresolution image processing

• 3-D speech synthesis

• Phoneme discriminative power analysis using other acoustic parameters
Thanks!
References

References

References
