Investigating /l/ variation in English through forced alignment

Jiahong Yuan & Mark Liberman
University of Pennsylvania

Sept. 9, 2009
Introduction

• English /l/ is traditionally classified into at least two allophones:
  “dark /l/”, which appears in syllable rimes
  “clear /l/”, which appears in syllable onsets.

• Sproat and Fujimura (1993):
  clear and dark allophones are not categorically distinct;
  single phonological entity /l/ involves two gestures –
  a vocalic dorsal gesture and a consonantal apical gesture.

• The two gestures are inherently asynchronous:
  the vocalic gesture is attracted to the nucleus of the syllable
  the consonantal gesture is attracted to the margin (“gestural affinity”).

• In a syllable-final /l/, the tongue dorsum gesture shifts left to the syllable
  nucleus, making the vocalic gesture precede the consonantal, tongue
  apex gesture.
  In a syllable-initial /l/, the reverse situation holds.
Introduction

• Clear /l/ has a relatively high $F_2$ and a low $F_1$;
  Dark /l/ has a lower $F_2$ and a higher $F_1$;
  Intervocalic /l/s are intermediate between the clear and dark variants (Lehiste 1964).

• An important piece of evidence for the “gestural affinity” proposal: Sproat and Fujimura (1993) found that the backness of pre-boundary intervocalic /l/ (in /i - ɪ/) is correlated with the duration of the pre-boundary rime. The /l/ in longer rimes is darker.

• S&F (1993) devised a set of boundaries with a variety of strengths, to ‘elicit’ different rime durations in laboratory speech:
  Major intonation boundary: Beel, equate the actors. “|”
  VP phrase boundary: Beel equates the actors. “\W”
  Compound-internal boundary: The beel-equator’s amazing. “\C”
  ‘#’ boundary: The beel-ing men are actors. “\N”
  No boundary: Mr Beelik wants actors. “\P”
Introduction

- Figure 1 in Sproat and Fujimura (1993): Relation between $F_2 - F_1$ (in Hz) and **pre-boundary rime duration** (in s) for (a) speaker CS and (b) speaker RS.
Introduction

- Figure 4 in Sproat and Fujimura (1993): A schematic illustration of the effects of rime duration on pre-boundary post-nuclear /l/.
Introduction

- Following Sproat and Fujimura (1993), Huffman (1997) showed that onset [l]s also vary in backness. The study suggested that the dorsum gesture for the intervocalic onset [l]s (e.g., in below) may be shifted leftward in time relative to the apical gesture, which makes a dark(er) /l/.

- The data utilized in these studies contained only a few hundred tokens of /l/ in laboratory speech.

- “The relation of duration and backness can be complicated by differences in coarticulatory effects of neighboring vowels, or by speaker-specific constraints on absolute degree of backness.” (Huffman 1997).

  => Our study utilizes a very large speech corpus.

- Automatic formant tracking is error-prone, and it is time-consuming to measure formants by hand.

  => We develop a new method to quantify /l/ backness without formant tracking.
Data

- The SCOTUS corpus includes more than 50 years of oral arguments from the Supreme Court of the United States – nearly 9,000 hours in total. For this study, we used only the Justices’ speech (25.5 hours) from the 2001-term arguments, along with the orthographic transcripts.

- The phone boundaries were automatically aligned using the PPL forced aligner trained on the same data, with the HTK toolkit and the CMU pronouncing dictionary.

- The dataset contains 21,706 tokens of /l/, including

  3,410 word-initial [l]s,

  7,565 word-final [l]s, and

  10,731 word-medial [l]s.
The aligner’s acoustic models are GMM-based monophone HMMs on 39 PLP coefficients. The monophones include:

**speech segments:** /t/, /l/, /aa1/, /ih0/, … (ARPAbet)

**non-speech segments:**

{sil} silence; {LG} laugh; {NS} noise; {BR} breath;
{CG} cough; {LS} lip smack

{sp} is a “tee” model that has a direct transition from the entry to the exit node in the HMM; therefore, “sp” can have “zero” length. The tee-model is used for handling possible inter-word silence.

The mean absolute difference between manual and automatically-aligned phone boundaries in TIMIT is about 12 milliseconds.

[http://www.ling.upenn.edu/phonetics/p2fa/](http://www.ling.upenn.edu/phonetics/p2fa/)
Forced Alignment Architecture

Cepstral Feature Extraction

Gaussian Acoustic Model

Viterbi Decoder

Gaussian acoustic models of phones
Word transcripts + pronunciation dictionary

Word and phone boundaries located
Method

• To measure the “darkness” of /l/ through forced alignment, we first split /l/ into two phones, L1 for the clear /l/ and L2 for the dark /l/, and retrained the acoustic models for the new phone set.

• In training, word-initial [l]’s (e.g., like, please) were categorized as L1 (clear); the word-final [l]s (e.g., full, felt) were L2 (dark). All other [l]’s were ambiguous, which could be either L1 or L2.

• During each iteration of training, the ‘real’ pronunciations of the ambiguous [l]’s were automatically determined, and then the acoustic models of L1 and L2 were updated.

• The new acoustic models were tested on both the training data and on a data subset that had been set aside for testing. During the tests, all [l]’s were treated as ambiguous – the aligner determined whether a given [l] was L1 or L2.
Method

- An example of L1/L2 classification through forced alignment:
Method

• If we use word-initial vs. word-final as the gold standard, the accuracy of /l/ classification by forced alignment is 93.8% on the training data and 92.8% on the test data.

```
L1   L2
L1  2987  235  (training data)
L2  414   6757

⇐ gold-standard by word position
```

```
L1   L2
L1  169   19  (test data)
L2  23    371

↑
classified by the aligner
```

• These results suggest that acoustic fit to clear/dark allophones in forced alignment is an accurate way to estimate the darkness of /l/.
Method

• To compute a metric to measure the degree of /l/-darkness, we therefore ran forced alignment twice. All [l]’s were first aligned with L1 model, and then with the L2 model.

• The difference in log likelihood scores between L2 and L1 alignments – the D score – measures the darkness of [l]. The larger the D score, the darker the [l].

The histograms of the D scores –

(Word-medial /l/s were classified as L1 or L2 by forced alignment.)
Results

- To study the relation between rime duration and /l/-darkness, we use the [l]s that follow a primary-stress vowel (denoted as ‘1’).

- Such [l]s can precede a word boundary (‘#’), or a consonant (‘C’) or a non-stress vowel (‘0’) within the word.
Results

From the figure we can see that:

- The /l/ sounds in longer rimes have larger $D$ scores, and hence are darker. This result is consistent with Sproat and Fujimura (1993).

- The rime duration being equal, the /l/ sounds preceding a non-stress vowel (1_L_0) are less dark than the /l/ sounds preceding a word boundary (1_L_#) or a consonant (1_L_C).

- The relation between the rime duration and darkness for the /l/ in 1_L_C is non-linear. The /l/ reaches its peak of darkness when the rime (more precisely, the stressed vowel and /l/) is about 150-200 ms.

- The syllable final /l/ sounds are always dark ($D > 0$), even in very short rimes, i.e., less than 100 ms. This result is contradictory to Sproat and Fujimura (1993)’s finding that the syllable-final /l/ in very short rimes can be as clear as the canonical clear /l/. 
Results

• To further examine the difference between clear and dark /l/, we compare the intervocalic syllable-final /l/ (1_L_0) with the intervocalic syllable-initial /l/ (0_L_1).

• The “rime” duration of the intervocalic syllable-initial /l/ (0_L_1) was measured as the duration of the non-stress vowel plus /l/, to be comparable to the intervocalic syllable-final /l/.

Yuan & Liberman: Interspeech 2009
Results

From the figures we can see that:

- The intervocalic syllable-final [l]s have positive $D$ scores whereas the intervocalic syllable-initial [l]s have negative $D$ scores.

- There is a positive correlation between darkness and rime duration (i.e., the duration of /l/ and its preceding vowel) for the intervocalic syllable-final [l]s, but no correlation for the intervocalic syllable-initial [l]s.

- For the intervocalic syllable-final /l/, there is a positive correlation between /l/ duration and darkness. No correlation between /l/ duration and darkness was found, however, for the intervocalic syllable-initial /l/.

- These results suggest that there is a clear difference between the intervocalic syllable-final and syllable-initial /l/ s.
Conclusions

- We found a strong correlation between the rime duration and /l/-darkness for syllable-final /l/. This result is consistent with Sproat and Fujimura (1993).
  We found no correlation between /l/ duration and darkness for syllable-initial /l/. This result is different from Huffman (1997).

- We found a clear difference in /l/ darkness between the 0_1 and 1_0 stress contexts, across all values of V+/l/ duration and of /l/ duration.

- We found that the syllable-final /l/ preceding a non-stress vowel was less dark than preceding a consonant or a word boundary. Also, there was a non-linear relationship between timing and quality for the /l/ preceding a consonant and following a primary-stress vowel. Such an /l/ reaches its peak of darkness when the duration of the stressed vowel plus /l/ is about 150-200 ms.
  Further research is needed to confirm and explain these results.