INTERSPEECH 2009 BRIGHTON

Investigating /l/ Variation in English through Forced Alignment

Jiahong Yuan, Mark Liberman

University of Pennsylvania, USA

jiahong@ling.upenn.edu, myl@cis.upenn.edu

Abstract

We present a new method for measuring the "darkness" of /l/, and use it to investigate the variation of English /l/ in a large speech corpus that is automatically aligned with phones predicted from an orthographic transcript. We found a correlation between the rime duration and /l/-darkness for syllable-final /l/, but no correlation between /l/ duration and darkness for syllable-initial /l/. The data showed a clear difference between clear and dark /l/ in English, and also showed that syllable-final /l/ was less dark preceding an unstressed vowel than preceding a consonant or a word boundary.

Index Terms: gestural phonology, forced alignment, variation

1. Introduction

The distinction between dark and clear /l/ in English has long been observed [1-2], and the two variants have traditionally been classified as allophones of the same phoneme [3-4]. Generally speaking, the dark /l/ appears in syllable rimes and the clear /l/ in syllable onsets. In a classic study of English /l/, Sproat and Fujimura (1993) argued that the clear and dark allophones are not categorically distinct [5]. They proposed that the 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 whereas the consonantal gesture is attracted to the margin ("gestural affinity"). When producing a syllable-final /l/, the tongue dorsum gesture shifts left to the syllable nucleus, making the vocalic gesture precedes the consonantal, tongue apex gesture. When producing a syllable-initial /l/, the reverse situation holds. As an important piece of evidence for their proposal, Sproat and Fujimura (1993) found that the backness of pre-boundary intervocalic /l/ is correlated with the duration of the pre-boundary rime. The /l/ in longer rimes is darker. Their explanation was that when the rime is short, the tongue dorsum gesture may not have enough time to reach its full target, and therefore the /l/ is lighter. They also found that the lightest pre-boundary /l/ can be as light as the prevocalic and clear /l/, and they explained this result with "gestural separation", i.e., conflicting tongue dorsum gestures will avoid temporal overlap. In very short rimes, the tongue dorsum gesture of the syllable-final /l/ may follow the apex gesture, as in clear /l/, avoiding a clash with the preceding (tautosyllabic) vowel's dorsum gesture. Sproat and Fujimura's study focused on the intervocalic syllable-final /l/. Huffman (1997) extended the investigation to intervocalic onset /l/ following schwa, e.g. in *below* [6]. Her study showed that intervocalic onset /l/ also varies in backness, suggesting that the dorsum gesture for the onset /l/ may also be shifted leftward, which is contradictory to the "gestural affinity" principle in Sproat and Fujimura's model. To provide a phonetic account of onset /l/ backness, Huffman (1997) proposed that "gestural separation" is stress sensitive and tends to keep the dorsum gesture of /l/ at a distance from the dorsum gesture of a neighboring strongly stressed vowel, no matter the vowel is in the same syllable with /l/ or not. Her explanation of onset /l/ backness was that if gestural affinity and gestural separation make conflicting predictions for an onset /l/, gestural separation may override gestural affinity, so that the tongue dorsum gesture of /l/ in words such as *below* can move leftward, making a dark /l/. The main difference between Sproat and Fujimura (1993) and Huffman (1997) is that gestural separation operates only within syllables in Sproat and Fujimura (1993) whereas it can operate across syllables in Huffman (1997).

The relation between timing and quality for intervocalic /l/ revealed in these studies is very compelling. The data utilized in the studies contained, however, only a few hundred tokens of /l/ in laboratory speech. In this study, we revisit the relation between timing and quality for /l/ using a large speech corpus, and compare the results with previous studies. In previous studies, the difference between F_2 and F_1 has been used as an acoustic measure of the darkness of /l/. The light /l/ has a relatively high F2 and a low F1, whereas the dark /l/ has a lower F2 and a higher F1 [7-8]. But because automatic formant tracking is error-prone (especially LPC analysis for sounds such as /l/ in which antiformants exist), and because it is time-consuming to measure formants by hand, we propose a new method for measuring the darkness of /l/.

Forced alignment has been widely used for automatic phonetic segmentation in speech recognition and corpus-based concatenative speech synthesis [9-10]. This task requires two inputs: recorded audio and (usually) word transcriptions. The transcribed words are mapped into a phone sequence in advance by using a pronouncing dictionary, or grapheme to phoneme rules. If a word has multiple pronunciations in the pronouncing dictionary, forced alignment will choose the most probable one for the acoustic observation. Furthermore, likelihood scores are associated with the forced alignment procedure, i.e., the log likelihood of the aligned segment to be a particular phone whose acoustic model is normally trained from thousands of tokens. These two characteristics of forced alignment provide a new perspective for investigating phonetic and phonological variation in speech.

In the following sections we first introduce the data set, then we describe and validate our method for measuring /l/-darkness through forced alignment. The results are presented in Section 4, followed by conclusions and discussions in Section 5. Finally, Section 6 summarizes the study.

2. Data

We utilized the 2001 term of the SCOUS corpus, which in full includes more than 50 years of oral arguments from the Supreme Court of the United States. Only the Justices' "clean" turns (i.e., the turns that have no noise, laughter, etc., based on the transcripts) were used for this study. The data set contained a total of 21,706 tokens of /l/.

The phone boundaries were automatically determined using the Penn Phonetics Lab Forced Aligner [11], whose acoustic models were trained on the same data set using the HTK

toolkit [12] and the CMU American English Pronouncing Dictionary [13]. The acoustic models are GMM-based, monophone HMMs. Each HMM state has 32 Gaussian Mixture components on 39 PLP coefficients (12 cepstral coefficients plus energy, and Delta and Acceleration).

3. A new method for measuring /l/-darkness

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. In training, the word-initial [l]'s (e.g., *like*) and the [l]'s in the word-initial consonant clusters (e.g., *please*) were categorized as L1 (clear); the word-final [l]s (e.g., *full*) and the [l]'s in the word-final consonant clusters (e.g., *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 a data set that had been set aside for the testing purpose. In test, all [l]'s were treated as ambiguous, the forced aligner determined whether a [l] was L1 or L2. The results for the word-initial and word-final [l]'s are shown in Table 1.

Table 1. *Classification of /l/ through forced alignment*.

	L1	L2	
L1	2987	235	(training data)
L2	414	6757	
			⇐ classified by word position
L1	169	19	
L2	23	371	(test data)
	⇑		
	classified by the aligner		

Table 1 shows that, for example, 2987 of the 3222 (2987+235) word-initial [l]'s in the training data were classified as L1, the clear /l/, by the aligner. 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. These numbers suggest that forced alignment can be used to determine the darkness of /l/.

To compute a score that can measure the degree of /l/-darkness, we ran forced alignment twice. First, all [l]'s were aligned using the L1 model, and then, using the L2 model. The difference between the likelihood scores resulted from L2 alignment and L1 alignment - the *D* score - measures the darkness of /l/ (Eq. 1). The larger the *D* score is, the darker the /l/.

$$D(l) = \log p(l \mid L_2) - \log p(l \mid L_1) \quad \text{(Eq. 1)}$$

Figure 1 draws the histograms of the *D* scores of all /l/ tokens in the data set. L1 and L2 were classified by forced alignment as above. We can see that, as expected, most L1's have negative *D* scores whereas most L2's have positive *D* scores. In the following we use the *D* score to investigate the variation of /l/ in the data set.

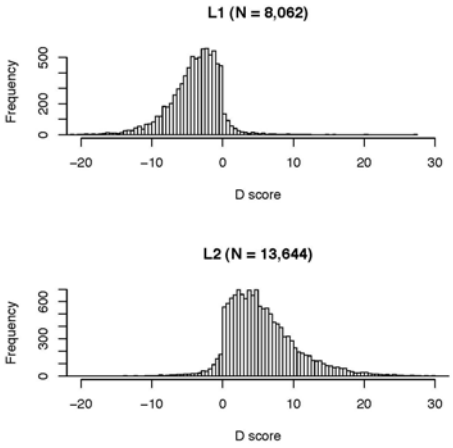

Figure 1. Histograms of D-scores for L1 and L2.

4. Results

4.1. Rime duration and /l/-darkness

We first study the /l/ in syllable rimes. Although such /l/ typically follows a primary-stress vowel (denoted as '1'), it can precede a word boundary (denoted as '#'), a consonant within the word (denoted as 'C'), or a non-stress vowel within the word (denoted as '0'). Figure 2 plots the average *D* scores of the /l/ for different rime durations, grouped by the type of the segment that /l/ precedes. The duration of C in 1_L_C is irrelevant in Sproat and Fujimura's model (whether it is part of the rime or not), therefore, it was excluded when measuring the rime duration for 1_L_C.

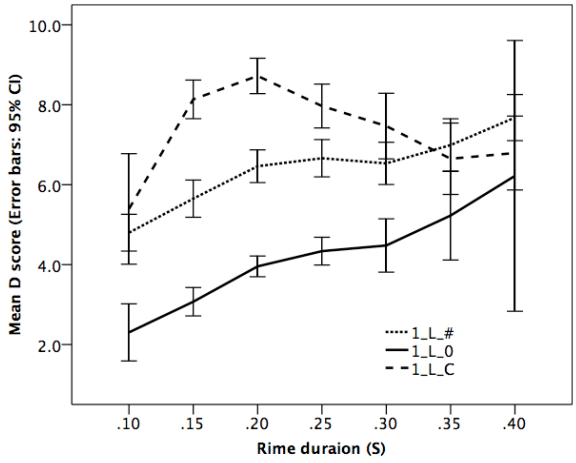

Figure 2. *Relation between rime duration and darkness for syllable-final /l/. The x-axis represents duration, ".10" means below .10s, ".15" means between .10 and .15 seconds, etc.*

We can see from Figure 2 that the /l/ in longer rimes has larger *D* scores, and hence is darker. This result is consistent with Sproat and Fujimura (1993). Figure 2 also shows that the rime duration being equal, the /l/ preceding a non-stress vowel (1_L_0) is less dark than the /l/ preceding a word boundary (1_L_#) or a consonant (1_L_C). This result presents a problem to both 'gestural affinity' (Sproat and Fujimura 1993) and 'gestural separation' (Huffman 1997), which we will discuss in Section 5.

From Figure 2 we can also see that relationship between the rime duration and darkness for the /l/ in 1_L_C is non-linear. For shorter rimes the correlation is positive whereas for longer ones it is negative; the /l/ reaches its peak of darkness when the rime (more precisely, the stressed vowel and /l/) is about 150-200 ms.

Finally, Figure 2 shows that the syllable final /l/ was always dark ($D > 0$), even in the rimes that were very short, i.e., less than 100 ms. This result is contradictory to Sproat and Fujimura's finding that the syllable-final /l/ in very short rimes can be as clear as the canonical clear /l/.

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). In Figure 3, 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/. We can see from Figure 3 that there is a clear distinction between the intervocalic syllable-final /l/ and syllable-initial /l/: The former has positive *D* scores and shows a correlation between darkness and rime duration (i.e., the duration of /l/ and its preceding vowel) whereas the latter has negative *D* scores and shows no correlation between darkness and the duration of /l/ and its preceding vowel. Figure 3 further demonstrates that in our data although the syllable final /l/ was less dark in shorter rimes, it was distinct from the clear /l/.

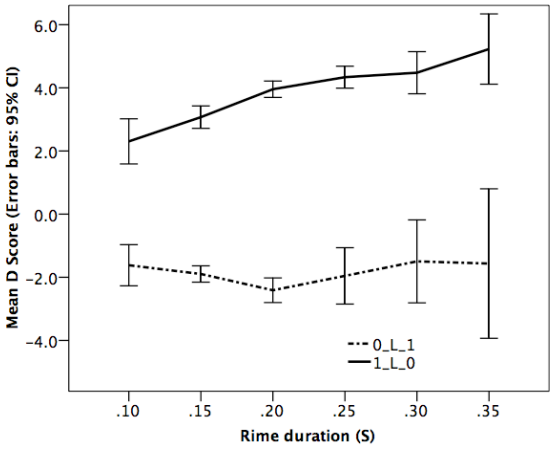

Figure 3. *Relation between rime duration and darkness for the intervocalic /l/. The x-axis represents duration, ".10" means below .10s, ".15" means between .10 and .15 seconds, etc. The "rime" duration for 0_L_1 is the duration of 0 and /l/.*

4.2. /l/ duration and /l/-darkness

This section investigates the correlation between the duration of /l/ and its darkness. Figure 4 plots the correlation between /l/ duration and darkness for the intervocalic syllable-final /l/ and syllable-initial /l/ respectively.

We can see from Figure 4 that, again, clear and dark /l/ show different patterns. 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 onset /l/. This result is different from Huffman (1997). In Huffman (1997), there was a correlation between /l/ duration and darkness for the onset /l/, and to provide an account of the correlation she refined 'gestural separation' to a stronger form.

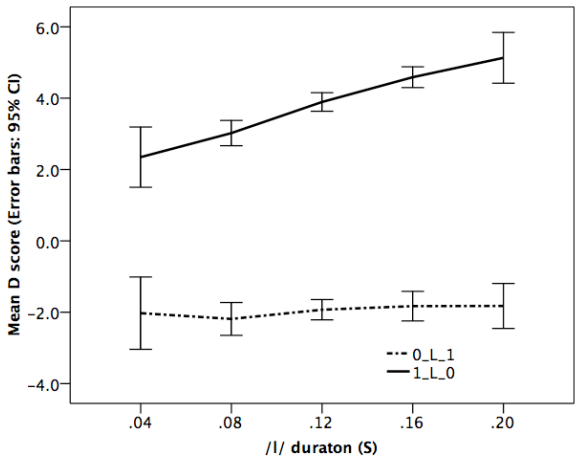

Figure 4. *Relation between /l/ duration and darkness for the intervocalic /l/. The x-axis represents duration, ".04" means below .04s, ".08" means between .04 and .08 seconds, etc.*

5. Conclusions and Discussions

The acoustic evidence for the 'gestural affinity' principle (i.e., the vocalic dorsal gesture is attracted to the nucleus of the syllable whereas the consonantal apical gesture is attracted to the margin) in Sproat and Fujimura (1993)'s model is that they found a correlation between the rime duration and /l/-darkness for the syllable-final /l/, using laboratory speech. In this study we found the same correlation from analyzing many more /l/ tokens from natural speech (about 50-100 times the number of tokens analyzed in Sproat and Fujimura 1993), especially for the /l/ preceding a word boundary or a non-stress vowel, as shown in Figure 2.

Figure 2 also shows that the syllable-final /l/ preceding a non-stress vowel was less dark than preceding a word boundary or a consonant. This result cannot be explained by 'gestural affinity'. According to 'gestural affinity', the leftward shift of the tongue dorsum gesture of /l/ (making a dark /l/) only depends on the duration of the rime. Our data suggested, however, that the rime duration being equal, the /l/ preceding a non-stress vowel was always less dark than preceding a word boundary or a consonant. Can 'gestural separation" explain the result? The 'gestural separation' principle has two forms: to avoid a clash with the gesture of the vowel in the same syllable, as proposed in Sproat and Fujimura (1993); or to avoid the neighboring strongly stressed vowel (either within or across syllables), as proposed in Huffman (1997). Neither of them can explain why the syllable-final /l/ preceding a non-stress vowel is less dark than preceding a word boundary or a consonant, because both of them predict that it is the vowel preceding the /l/ (the vowel is in the same syllable with /l/ and also strongly stressed) not the one following /l/ that is responsible for 'gestural separation'.

At least two possibilities remain to account for the finding that the syllable-final /l/ preceding a non-stress vowel was less dark than preceding a word boundary or a consonant. First, the /l/ preceding a non-stress vowel becomes less dark due to the coarticulatory effect of the vowel. Huffman (1997) showed, for example, that the relation of duration and backness for the English onset /l/ could be complicated by differences in coarticulatory effects of neighboring vowels [6]. The second hypothesis is that the syllable-final /l/ preceding a non-stress vowel is ambisyllabic [14]; it belongs to both the rime of the preceding syllable and the onset of the following syllable. Further studies are needed to test these hypotheses. In

addition, speaker and dialect variation might also play a role [6, 15].

Our data showed that the syllable final /l/ was always dark, even in very short rimes. This result is contradictory to Sproat and Fujimura (1993)'s finding, and it presents a challenge to their claim that clear and dark /l/ in English are a single phonological entity. Furthermore, as shown in Figure 3 and 4, our data showed a clear difference between clear and dark /l/ in English with respect to the correlation between timing and quality: the quality of the dark /l/ was correlated with both the rime duration and the duration of /l/ whereas the quality of the clear /l/ showed no correlation with timing.

Finally, we found that when following a primary-stress vowel and preceding a consonant, the /l/ is most dark when the duration of the vowel plus /l/ is about 150-200 ms, the darkness decreases gradually when the duration becomes either shorter or longer. Future research is needed to determine whether this non-linear relationship is an artifact of the current study or a reality of English speech production.

6. Summary

We present a new method for measuring /l/-darkness in natural speech through forced alignment. We use it to investigate the variation of English /l/ in a large speech corpus, and compare the results with Sproat and Fujimura (1993) and Huffman (1997).

We found that in our data there was a 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).

The data showed a clear difference between clear and dark /l/ in English, which presents a challenge to the claim that the two variants of /l/ are a single phonological entity. The data also showed that the syllable-final /l/ preceding a non-stress vowel was less dark than preceding a consonant or a word boundary. This result cannot be explained by "gestural affinity" or "gestural separation".

Finally, we found a non-linear relationship between timing and quality for the /l/ preceding a consonant and following a primary-stress vowel. Such /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 examine this result.

7. Acknowledgements

We would like to thank Jerry Goldman and the team of the OYEZ project for providing the data. This work was partially supported by NSF award 0325739.

8. References

[1] Sweet, H., *The Sounds of English*. Oxford: Clarendon Press, 1908.

[2] Jones, D., *An Outline of English Phonetics*. Cambridge: W. Heffer and Sons, 1947.

[3] Halle, M. and Mohanan, K. P., "Segmental phonology of model English*", Linguistic Inquiry*, 16, 57-116, 1985.

[4] Giegerich, H., *English Phonology: An Introduction*. Cambridge: Cambridge University Press, 1992.

[5] Sproat, R. and Fujimura, O., "Allophonic variation in English /1/ and its implications for phonetic implementation", *Journal of Phonetics*, 21, 291–311, 1993.

[6] Huffman, M. K., "Phonetic variation in intervocalic onset /l/'s in English", *Journal of Phonetics*, 25, 115–141, 1997.

[7] Lehiste, I., *Acoustical Characteristics of Selected English Consonants*. The Hague: Mouton, 1964.

[8] Olive, J., Greenwood, A. and Coleman, J., *Acoustics of American English: A Dynamic Approach*. New York: Springer, 1993.

[9] Wightman, C. and Talkin, D., "The Aligner: Text to speech alignment using Markov Models," in J. van Santen, R. Sproat, J. Olive, and J. Hirschberg (ed.), *Progress in Speech Synthesis*, Springer Verlag, New York, pp. 313-323, 1997.

[10] Hosom, J. P., *Automatic Time Alignment of Phonemes Using Acoustic-Phonetic Information*. PhD thesis, Oregon Graduate Institute of Science and Technology, 2000.

[11] Yuan, J. and Liberman, M., "Speaker Identification on the SCOTUS corpus", *Proceedings of Acoustics 08*, 5687-5690, 2008.

[12] The HTK toolkit: http://htk.eng.cam.ac.uk/

[13] The CMU American English Pronouncing Dictionary: http://www.speech.cs.cmu.edu/cgi-bin/cmudict/

[14] Kahn, D., *Syllable-based generalizations in English phonology*, PhD thesis, MIT, 1976.

[15] Carter, P. and Local, J., "F2 variation in Newcastle and Leeds English liquid systems", *Journal of the International Phonetic Association*, 37, 183-199, 2007.