Perceptual coding reliability of (L)-vocalization in casual speech data

Perceptual coding reliability of (L)-vocalization in casual speech data

Lauren Hall-Lew
University of Edinburgh

Sonya Fix
New York University

Corresponding author contact:

Linguistics and English Language
The University of Edinburgh
Dugald Stewart Building, 3 Charles St.
Edinburgh EH8 9AD
United Kingdom

Email: Lauren.Hall-Lew@ed.ac.uk

The slightly revised version, with figures, appears as:

Hall-Lew, Lauren and Sonya Fix. (2012) Perceptual coding reliability of (L)-vocalization in casual speech data. *Lingua.* 122: 794-809.

The published version is available at:

http://www.sciencedirect.com/science/article/pii/S0024384111002464

Abstract

(L)-vocalization has been receiving increasing attention in sociophonetic research but is a challenging variable to measure consistently. Acoustic measures are not typically used because velarized-(L), which is the realization most likely to vocalize, is itself extremely difficult to distinguish from a back rounded vowel based only on acoustic features. Because of this, as well as the difficulty in using articulatory measures to capture spontaneous, field-based data, sociolinguists have typically relied solely on auditory coding measures. However, the level of consistency across coders is an issue of particular methodological concern when employing auditory coding, both within and across studies. The current paper presents results from a multi-listener perception survey of (L)-vocalization coding. Phonetically and sociolinguistically trained listeners evaluated a range of productions from two ethnically diverse U.S. English communities: Columbus, Ohio, and San Francisco, California. The survey investigates inter-coder consistency with respect to both phonetic environment and speech variety, with results showing that reliability is dependent on both factors. Inter-coder disagreement is also highest for tokens rated at intermediate levels of vocalization. Given our ethnically diverse speaker sample, we further ask how the coder's perception of a speaker's ethnicity interacts with their vocalization coding decisions. Our findings bear on the methodological decisions made in research that relies on auditory coding, drawing particular attention to the challenge of designing a method sensitive to patterns of variability and social meaning that are potentially both universal and community-specific.

Keywords
variation, sociophonetics, auditory methods, (L)-vocalization, US English, ethnicity

1. Introduction

A vocalized (L) is one that phonetically resembles a back vowel, semi-vowel, or voiced glide (Wells 1982:258; Hardcastle & Barry 1989). The vocalization of (L) has been receiving increasing attention in variationist studies across the English speaking world, including the United States (Ash 1982; Bailey & Thomas 1998; McElhinny 1999; Dodsworth 2005; Durian 2008; Fix 2004, 2008; Hall-Lew 2010), Great Britain (Sivertsen 1960; Wells 1982; Trudgill 1986; Wright 1989; Hardcastle & Barry 1989; Tollfree 1999; Heselwood & McChrystal 2000, Borowsky 2001; Przedlacka 2001; Timmons et al. 2004; Stuart-Smith et al. 2006; Johnson & Britain 2007), and Australia and New Zealand (Bauer 1986, 1994; Borowsky and Horvath 1997; Borowsky 2001; Horvath & Horvath 2002). One question sociolinguistic research needs to address is if vocalization across these different varieties is phonetically analogous; are these studies even looking at 'the same' variable? This is an entirely open question, and the difficulty is that it cannot be investigated appropriately without the field agreeing on a more robust methodological approach to measuring (L)-vocalization. As vocalization continues to be of increasing sociolinguistic interest, its analysis presents a real methodological challenge that needs to be addressed.

(L) is an articulatorily complex segment, resulting in tremendous variability in production. Contemporary descriptions of (L) allophony in standard varieties of

English typically focus on the 'light'/'dark' contrast, though English varieties differ with respect to realizing this distinction (see Lawson et al. 2011 for an overview). Both variants involve a lateral lowering that defines the segment as a variant of (L). Clear or 'light' (L) is typified by articulatory contact between the tongue tip and the alveolar ridge, while velarized or 'dark' (L) may include this contact, but is typified by retraction of the tongue dorsum towards the velum. Sproat and Fujimura (1993) argue that the 'light'/'dark' distinction is best captured in terms of the timing of the coronal and dorsal gestures; in onset (L) the coronal articulation occurs before the dorsal, whereas in nucleus and coda (L) the dorsal articulation occurs before the coronal. Borowsky (2001) analyzes this observation with respect to syllable harmony, whereby an onset 'light' (L) is more consonantal, and nucleus or coda 'dark' (L)s are more vocalic; by extension, a truly vocalized (L) is one with no coronal articulation. This observation fits with the claim that vocalized (L) always develops from velarized (L) (van Reenen 1986 states this for Dutch; see also Johnson & Britain 2007).

The phonetic distinction between the most consonantal (L) and the most vocalized (L) is acoustically straightforward. More consonantal (L)s exhibit an 'antiformant' between the second and third formants, have higher second formant values, and shorter formant transitions than more vocalized (L)s (Lawson et al. 2011). The difficulty arises in distinguishing reliably between a relatively vocal (L) with velar constriction (and variable, late coronal contact) and a truly vocalized (L). The velarized (L) shares a vocalic articulatory gesture with a back rounded vowel (Sproat & Fujimura 1993; Przedlacka 2001; Gick et al. 2002), and lacks the antiformant resonances typical of 'light' (L). These facts make the variable extremely challenging to code for, especially in the environment following a back vowel (cf. Timmins et al. 2004). The precise phonetic difference between velar (L) and vocalized (L) is one of the more subtle variable distinctions in sociophonetic research and presents one of the biggest methodological challenges.

Furthermore, defining vocalization as the lack of articulatory closure is too simplistic and quite possibly incorrect (Recasens & Espinosa 2009). Since (L) is a complex segment, involving coronal, lateral, and velar contact, there is a lot of variation between tokens of velarized (L) and vocalized (L). Defining vocalization as the lack of apical contact, for example, is problematic, as many productions without apical contact still maintain velar constriction and are not vocalized (cf. Wrench & Scobbie 2003). There is also some electropalatographical evidence that lateral articulation can be maintained even with a vocalized percept (Keating, p.c.). Even labial articulation is a relevant factor, because vocalization is often, for some speakers accompanied by lip-rounding (Wrench & Scobbie 2003). In this paper we treat vocalization as the perception of the lack of an (L) where a velar (L) would be expected, paired with some kind of reduced (L) articulation.

As Hardcastle and Barry (1985) observed, "[t]he decision as to whether a given alveolar pattern is to be regarded as … a full (L) realization is to a certain degree subject to the ultimately arbitrary criteria governing the classification of borderline cases." This tremendous articulatory variability leads to a conceptualization of (L) articulations as points along a gradient continuum. In our view, vocalization represents just one range of that articulation continuum. Velarized (L) represents another range that is relatively more vocalic than the range represented by consonantal (L), and vocalized (L) is more vocalic and than velarized (L). The final

variant of the range is total deletion of (L), which can be thought of as maximally or truly vocalic (Scobbie & Pouplier 2010:241), especially if vocalization is considered to be a form of lenition. Justification for treating deletion as the most extreme end of vocalization comes from the vowel-like quality of the segment which results when syllabic (L) is vocalized, as in words like *people* and *settle*. Previous studies have also adopted the view that deletion evidences the most advanced form of vocalization (Ash 1982; Durian 2008; Scobbie & Pouplier 2010), while others have treated complete deletion as a separate 'miscellaneous' classification (Timmins et al. 2004).

The ways that phoneticians and sociolinguists have attempted to measure and code for the presence of (L)-vocalization are many, though a single reliable method (where 'reliability' would be similar to the level of representing a vowel by its first two formant values, for example) has not yet been found. Because of its acoustic complexity, most sociolinguistic studies use auditory, impressionistic coding, with one to (at most) three different coders. Given the challenges, how reliable is this method, and how reliable are the data that result? In order to answer these questions, and to further assist sociolinguists working on (L)-vocalization, this paper presents the results of a large-scale perceptual coding task and examines the factors that impact inter-rater reliability in impressionistic coding of (L)-vocalization. We explore the linguistic and social factors that impact a linguist's auditory coding reliability: the phonological environment of the (L) token, the social traits of the voices being coded, the social traits of the linguists doing the coding, and the interaction between all three.

2. Methods of Measurement

In theory, the phonetic production of (L) is like all phonetic variables, and can be analyzed using articulatory methods, acoustic methods, and auditory methods. Each method has inherent limitations, and this is particularly true for the study of (L)-vocalization. Whether with regard to articulatory methods' compatibility with the spontaneous, field-based data collected with sociolinguistic methodology, or the ability to distinguish vocalised (L) from velar (L) using acoustic measurements, there are several basic reasons for the preference for auditory methods in measuring (L), even though consistency between coders and across data sets is open to question.

2.1 Articulatory Methods

While articulatory measures offer a greater level of quantitative accuracy over auditory methods, at the writing of this paper they still present considerable challenges for researchers interested in casual speech obtained in field conditions. Methods such as EPG (Hardcastle & Barry 1989; Scobbie et al. 2007; Wright 1989; Wrench & Scobbie 2003; Scobbie & Pouplier 2010), EMA (Wrench & Scobbie 2003), MRI (Gick, Kang & Whalen 2002), and Microbeam (Sproat & Fujimura 1993) have all so far been limited to the measurement of articulations in the speech of subjects within the laboratory setting. Developments in Ultrasound and other technologies (Wrench & Scobbie 2003; Scobbie, Stuart-Smith & Lawson 2008) point to future ways in which articulatory measures might be taken in fieldwork settings, but their utility for truly spontaneous speech still remains to be tested, and the technologies not yet readily available to most researchers.

2.2 Acoustic Methods

In terms of acoustic methodology, there is at present no definite measure for distinguishing a vocalized token from a velar token, largely because of the challenges mentioned above. Figures 1 and 2 offer a visual example of this challenge. Both are images from PRAAT (Boersma 2003) of single-syllable utterances of an (L)-final syllable with a front vowel nucleus, which is the vocalic environment where identifying vocalization should be the easiest. Figure 1 shows a more consonantal (L) production and Figure 2 shows a more vocalized (L) production, both of the word *hill* in the fixed phrase *Say hill here*.

- - [Insert Figure 1 about here] - -

Figure 1: A spectrogram of a consonantal (L) in the word *hill.*

- - [Insert Figure 2 about here] - -

Figure 2: A spectrogram of a vocalized (L) in the word *hill*.

Figures 1 and 2 show that even when (L) follows a vowel with formant structure quite different from the formant structure of (L), it is difficult to determine which acoustic features differentiate velar (L) from vocalized (L). For example, there is no major difference in the amplitude, the height of the first and second formants, or the formant bandwidths. What differences there are even go in the opposite direction of what might be expected; for example, the more sonorant production of vocalization might be expected to have higher amplitude at the syllable coda, but in this particular example, it does not. The vocalized token in Figure 2 does have a slightly lower F2 than the consonantal token, as would be expected, but the subtlety of this difference in the environment shows just how difficult it would be to use second formant height as a reliable cue in other vowel contexts.

One method that has been proposed to measure (L)-vocalization acoustically (Dodsworth, Plichta, & Durian 2006) measures the change in amplitude between the preceding vowel and the following (L) at intervals of ten milliseconds. Greater change in amplitude between the vowel (more sonorant, high amplitude) and the (L) (less sonorant, lower amplitude) is taken to mean more (L) closure and less vocalization. Unfortunately, our own separate attempts to replicate this measure across our own two datasets have proved unsuccessful, so far. One of the main challenges is that the syllable offset will almost always have decreased global amplitude relative to the nucleus, so the comparison is only meaningful across multiple tokens of the same type (phonological environmental, speaker, sentential stress or prosody, etc.). The measure is also incompatible with measuring syllabic (L)s. This latter point was addressed by Hazen and Dodsworth (2011), who argue that vocalization for syllabic (L) may be indicated by drop in F3; again, this measure remains to be replicated across datasets.

2.3 Auditory Methods

Due to the significant challenges of articulatory and acoustic methods, most sociolinguistic variationist studies of vocalization have employed perceptual coding techniques. Previous auditory coding methods reduce the gradient range of (L) production down to a scale with at most five levels, ranging from (L) with perceived coronal contact to deleted (L). Studies that employ the popular statistical package VARBRUL or its updated version, GOLDVARB, are forced to reduce this newly created, multi-level discretization down even further to a mere binary measure. Consequently, many studies of (L) vocalization ultimately rely on results that comparing one set of generally consonantal tokens to another often wide-ranging set of generally vocalized tokens. Reducing the (L) production to a binary variable in this way is inherently problematic given what we know about its phonetic properties, and potentially interesting detail – both phonetic and social – is lost in that process. Furthermore, each study makes different choices concerning the number of individual coders used, ranging from one coder (Ash 1982; McElhinny 1999; Horvath & Horvath 2002; Fix 2004; Hall-Lew 2010), to two coders (Timmins et al., 2004; Stuart-Smith et al. 2006), to at most three coders (Dodsworth 2005; Durian 2008), although the impact of this decision on the results of a given study is not yet known.

As phoneticians continue to work toward a reliable and duplicable methodology for measuring (L) acoustically or articulatorily, sociolinguists continue to code (L)-vocalization impressionistically. Since impressionist coding remains by far the most popular method of measuring vocalization, we may ask how reliable is it within and across data sets.

3. A Cross-coder Perception Study

3.1 Motivations

Anyone who has coded (L) impressionistically knows that there is a high degree of difficulty and doubt that comes with the coding process. In addition to the phonetic subtleties, coders may feel they are coding differently for each speaker based on a given speaker's range of (L) articulation. Coders may also worry that their own linguistic knowledge and expectations about (L) variability may impact their coding, which is one reason linguists have occasionally employed external coders (Horvath & Horvath 2002, Dodsworth 2005). Furthermore, coders who come from vocalized dialects may perceive and rate tokens differently than coders who themselves do not use vocalized variants. Adding to the phonetic complexity is the observation that the vocalized variant itself may take many forms and these forms may vary from community to community: a back rounded vowel, a voiced glide, a schwa (attributed to African American communities, cf. Green 2002:120), or total deletion.

The current study addresses these concerns through a multi-coder perception study of (L) realization, based initially on Yaeger-Dror et al.'s (2009) survey of the perception of (r) realization. The goal in both their study and ours was to create a perception task that mimicked the auditory coding process sociolinguists undergo when collecting vocalization data from interview recordings, and to test for the relevant factors that impact variability within and across coders. In the present study we took recordings

from two different U.S. English speech communities and elicited responses from linguists with a range of social and linguistic backgrounds. The study asks how consistent coders are with respect to their fellow linguists and across a range of (L) token types.

3.2 Pilot Study

A two-stage pilot study was conducted to prepare the stimulus set for the main perception study. The challenge in stimuli creation was that the set had to contain a wide yet balanced range of (L) tokens, from very consonantal to very vocalized. This was a challenge because determining the level of vocalization for any given token is itself difficult (and the reason for having the study in the first place). The pilot study was thus necessary to ensure some kind of balanced representation of token types.

All tokens come from sociolinguistic interviews from two regionally and ethnically distinct communities in the United States that variably vocalize (L). The first community is San Francisco, California, where (L)-vocalization appears to be associated with speakers of Asian heritage, particularly those of Chinese and Japanese backgrounds (Hall-Lew & Starr 2010). The second community is Columbus, Ohio, where (L)-vocalization is variably used by African Americans and white speakers. We chose productions of (L) from ten speakers, five from each region. Of the five speakers from San Francisco, four are Chinese American and one is Japanese American. Of the five speakers from Columbus, four are white and one is African American. All are female, to eliminate one social variable from the analysis.

The speakers were chosen from a much wider pool of speakers from our independent research projects (Fix 2011; Hall-Lew 2009). The five were each chosen to represent a range of (L) production that best represented the community, which was a subjective decision based only on our general impressions of each speaker's use of (L)-vocalization. The tokens were selected from the early middle of each interview, and utterances of (L) were extracted as they appeared. Tokens with background noise, or durations less than 60ms, were excluded. We also excluded all tokens containing ambisyllabic (L) (as in 'settle late'), known historically variable tokens like *palm, folk,* etc., and any obvious instances of (L)-insertion (e.g. 'wheel barrel' for 'wheel barrow'). In other words, we excluded those tokens which would also be discarded from a typical quantitative sociolinguistic analysis of (L)-vocalization. It is important to bear in mind that we were not designing a traditional Verbal Guise Test, which would have manipulated level of vocalization in a controlled manner, and excluded very ambiguous productions of (L). Rather, the goal was to mimic sociolinguistic coding in the form of a perception test, and so we tried to represent the methods that have been used to obtain vocalization data in previous and on-going variationist studies.

The quality of the preceding vowel has been previously shown to correlate with (L)-realization (Ash 1982; Recasens 1996; McElhinny 1999; Borowsky 2001; Horvath & Horvath 2002; Durian 2008; van Hofwegen 2010; but see van Reenen 1986); in most but not all data sets, non-front vowels have been shown to favor (L)-vocalization. Our final set of stimuli for the pilot study contained a representative range of preceding vowel environments. Table 1 shows this range. We also created two versions of the same stimuli set, which varied in terms of the length of speech on either side of the

(L) token. Based on personal coding experiences and feedback from other researchers who code for (L)-vocalization, we set out to test the observation that the vocalized quality of an (L) shifts according to how isolated it is in the speech stream. The two contexts are the 'sentence' context and the 'syllable' context. The former includes the syllable containing the (L) and several syllables of speech preceding and following it, approximating an intonational phrase as best as possible. The latter includes the syllable containing the (L) and only one syllable of speech preceding and following it.

- - [Insert Table 1 about here] - -

Table 1: The preceding phonological environments represented in the stimuli

The pilot study was arranged in two stages, one stage of stimuli coding between the two authors, and one stage of stimuli coding among three other sociophoneticians variably experienced with coding (L). All coding was done on a 4-point scale:

1. Definitely consonantal
2. Some vocalization, but more consonantal than vocalized
3. Stronger vocalization, more vocalized than consonantal
4. Definitely vocalized

In the first stage, the two authors cross-coded each of the 119 total tokens for the two regional token sets. Differences in coding decisions were identified and discussed, until a consensus was reached on a coding strategy. Tokens that were particularly difficult to reach consensus on were tagged and discussed at length; if consensus was not possible for linguistic reasons, tokens were included; tokens with non-linguistic problems (e.g., recording quality) were discarded. In the second stage, one of the three coders was giving the same 119 tokens and given a nominal gift for their time and effort. The other two coders were each given approximately half of the token set (62 and 57 tokens, respectively), balanced for speaker, dialect region, and phonological environment. The second-stage pilot coders provided input on time to completion, token quality, and any factors that made coding difficult.

Across all five pilot-stage listeners, there was surprisingly high consistency in coding decisions – most tokens had a standard deviation of around 0.5 on a 4-point scale. One point of difference was that the stage 1 pilot coders (the authors of the present paper) perceived a difference in (L) realization between the 'sentence' and 'syllable' token contexts. However, none of the stage 2 pilot coders found this difference to influence their coding decisions in any way, so in the main study the two contexts were always both presented to listeners back to back as a single token entry.

The pilot study results and feedback from pilot coders informed the design of the main study, which was primarily modeled on the study of (r) realization by Yaeger-Dror and colleagues (2009). One key difference between the two pilot stages and the main study was presentation of stimuli with respect to speaker. In the pilot stages, all the tokens for a single speaker were presented together, with the goal of best representing the task that a sociolinguist undertakes in coding their own interview data. However, following the design of Yaeger-Dror et al.'s (r) study, in our main

study the stimuli were pseudo-randomized such that no two adjacent tokens came from the same speaker.

3.3 The Main Study

In order to reach a wide number of professional linguists, we used an internet-based survey hosted by the website SurveyGizmo <www.surveygizmo.com>. Respondents first agreed to participate in the study and have their results used for research purposes. They then took part in a brief training task before beginning the main coding task. In the training task the participant listened to six tokens of (L) that had been consistently coded as consonantal in the pilot studies, followed by nine tokens of (L) that had been consistently coded as vocalized. The tokens represented a range of preceding vowel environments. Participants were not asked to code the training stimuli, but rather to listen and consider the way we were defining 'consonantal' and 'vocalized' in our dataset, and to become familiar with the voices of our speakers. They were also given space to comment after each token, which some used to state disagreement about the tokens being classified as straight-forwardly 'consonantal' or 'vocalized'. Following training, participants were presented with four different realizations of one word—'older'—arranged according to our 4-point vocalization coding scale. Again they listened and did not give codes to these tokens, and they could comment on the token choice if they wished.

In the main study, participants listened to each token, one per page, presented back-to-back in syllable then sentence environments. They were given the opportunity to listen to the tokens as many times as they chose.[1] They then had to assign the token a rating based on the 4-point coding scale. They were again given space to comment after each token, if they wished. At the end of the survey they were again given space to comment freely.

The main study was divided into two surveys, A and B, which were identically balanced for speaker and phonological environment. The only structural difference between A and B was that A had 60 tokens and B had 40 (different) tokens. Two short surveys were used instead of one long survey in order to avoid participant fatigue (which several participants in survey A complained of, even with only half the initial dataset) and to encourage a large number of fully completed surveys (as opposed to ones abandoned midway though). The number was reduced in survey B by eliminating the 20 tokens of (L) that followed either front tense vowels or consonants, both environments that result in a (relatively) more syllabic (L) than a syllable-coda (L) (Borowsky & Horvath 1997; Gick & Wilson 2006). The result of this decision is that all analysis of preceding vowel presented here only compares four categories of vowel:

1. front high and mid (including both tense/long FLEECE, FACE (see Wells 1982) and lax/short KIT, DRESS lexical classes in survey A and only lax/short KIT, DRESS lexical classes in survey B);
2. low back or central (including tokens of the LOT, CLOTH, and THOUGHT lexical groups, but not BATH or TRAP);
3. the (ow) or GOAT vowel;

4. the (uw) and (U) or GOOSE and FOOT vowels (combined because of variable merger in our datasets).

As the results in Figure 3 show, a survey with 40 stimuli proved to be a better length than one with 60, because while inter-rater disagreement increased for those stimuli that appeared towards the end of survey A, perhaps indicating fatigue, there was no such increase across time for the survey B.

- - [Insert Figure 3 about here] - -

Figure 3: A comparison of surveys A and B, with respect to inter-rater disagreement (standard deviation) for each token, progressing from the start of the survey to the end of the survey.

The two surveys also differed with respect to participant recruitment. For survey A, 43 professional linguists were invited by individual email to participate. For survey B, another 41 professional linguists were invited by individual email to participate, but the survey was also opened up to professional or otherwise trained linguists who were within the authors' internet-based social networks, including Facebook and Twitter accounts. As a result, survey B respondents varied more with respect to linguistic background, both personally (in terms of language/dialect background) and professionally (in terms of the kind of linguistics studied). Overall, our recruitment led to 26 linguists responding for survey A and 27 for survey B. With the different number of token stimuli for the two surveys, this yielded 2,640 judgments for 100 different tokens of variable (L) production.

We collected as much potentially relevant social information about each participant as possible. This included not only their age, sex, ethnicity, and nationality, but also their native language, native accent of English, hometown, current town, accents of English currently exposed to, experience with phonetics research, and experience coding for (L) or (r) variation. All social variables therefore constitute self-reported classifications, however specific the participant chose to be. For example, some respondents described their native accent as Standard American or Southern Standard British, whereas others were much more specific. Demographic information about all the participants is given in Table 2. Non-native English speakers' native languages included Cantonese (x2) in Survey A and Norwegian, Russian (x2), Spanish, and Swedish in Survey B. Ethnic identities of the British, Irish, and Australian respondents, as well as the Survey B non-native English speakers, were all self-described as 'white' and are listed under the ethnicity category 'European, non-/part-Jewish'. The five 'Asian' respondents include both Asian Americans and respondents born in Asia.

- - [Insert Table 2 about here] - -

4. Quantitative Results

The results address the following three sets of questions:

1. How great was the inter-rater agreement between linguists? In what ways did coders differ from one another (did they code more or less vocalization than the average)? What were the social factors about the coders that correlated with differences between them?
2. How great was the inter-rater agreement across tokens? What token-specific factors predict its likelihood of being rated more or less consistently?
3. What factors predict vocalization in this dataset? What token-specific factors predict its likelihood of being rated more or less vocalized?

4.1 Inter-rater differences between raters

As is expected for a 4-point scale, the average token rating across all coders was 2.5. This suggests that the pilot study was successful in providing a range of tokens along a continuum of vocalization. The minimum average vocalization rating for any single coder was 2.1, indicating that coder's preference to rate tokens as not vocalized, while the maximum rating average for a coder was 3.2, indicating that coder's preference towards rating tokens as more vocalized.

To quantify inter-rater difference, we calculated two dependent variables: how different a given coder was from the rest of the coders in a survey, represented by the average of the differences from the survey mean of token codes, and how variable a given coder was, across tokens, with respect to that difference, represented by the standard deviation (variability) of the differences from survey mean. In other words, we computed the average vocalization rating for each token, and then calculated the token-by-token difference between a given participant's set of codes and the respective survey's average set of codes. The average of that difference represents to what extent that coder differed from the mean, as well as the direction in which they differed: overall positive scores indicate a preference for coding tokens as more vocalized than the survey average, and overall negative scores show a preference for coding less vocalization than the average. The standard deviation of those values shows if differences on a token-by-token level were consistently in one direction (resulting in low standard deviation values) or if they went in both directions (resulting in higher standard deviation values).

- - [Insert Figure 4 about here] - -

Figure 4: Inter-rater variance for surveys A and B.

Figure 4 represents the variation between the coders with respect to the two surveys. The *x*-axis shows for each coder how different they were from the mean, and the direction in which they were different. Coders to the right of the *y*-axis were on average coding tokens as more vocalized than the mean, those on the left, less. The *y*-axis represents how variable a given coder was with respect to that difference. The coders with most variation have the highest *y*-axis values. Figure 4 shows how most of the coders cluster near each other (within the circle), indicating that most linguists were pretty consistent with one another. The individuals marked as *M, N, P,* and *Q* might be considered outliers from the overall pattern. It is also worth noting that the

overall distribution for B is slightly higher than for A, indicating slightly more variance between coders for B than for A.

Only a few coders differed from their survey mean, and as can be seen in Figure 4, they were also among the most variable across tokens. Each survey shows two individuals who seem to disagree markedly from their fellow coders. All four appear as outliers because of their high standard deviation (*y*-axis) values, values which represents the variability across tokens between their individual responses and the groups' mean responses. The two outliers in A, *M* and *N*, and one of the two outliers in B, *Q*, also show greater *x*-axis differences than the mean. This indicates that these three coders disagreed with the general coding decisions both at a token level and at the level of a more general bias towards coding for more vocalization than average (*N* and *Q*) or less (*M*).

Across the 53 respondents for both surveys, there were absolutely no social factors that we tested for that correlated with variation in individual coding practices. Note that this is contrary to what Yaeger-Dror et al. (2009) found for (r) coding—U.S. coders were found to be the most accurate compared to those from the U.K., Australia, and New Zealand. However, the comparison in that case was between the social factors of the coders, their coding decisions, and an acoustic measure of rhoticity (the height of F3). Our finding of the utter lack of difference between raters based on social or linguistic background is both encouraging and somewhat surprising, and is discussed further in section 7.

4.2 Inter-rater differences across tokens

To calculate the inter-rater reliability with respect to individual tokens, we took the standard deviation for each token across all survey respondents. Tokens with higher standard deviations had more variation in how they were rated; coders had more disagreement or confusion about how vocalized those tokens were, in comparison to tokens with a standard deviation of zero, which indicates complete agreement across all coders. Note that we did not use other typical measures of inter-rater reliability, such as Cohen's kappa correlation coefficient (Banerjee et al., 1999; Maryn 2009), because we wanted one measure of reliability per each individual token. Cohen's kappa strongly overestimates variability at the item level.

- - [Insert Figure 5 about here] - -

Figure 5: Significant correlation between token rating and its rating reliability.

The strongest predictor of inter-rater variability at the token level is the perceived amount of vocalization for that token. Figure 5 plots a token's vocalization rating, averaged across all respondents, with the token's standard deviation, for a combined dataset of both survey A and B. The tokens that were rated more similarly across all respondents are those that were coded as either definitely *not* vocalized, 1, or definitely vocalized, 4. The tokens with the most variance across coders were those with the average rating in the middle of the range, around 2 or 3. In other words, those tokens with an average of a 2 were not consistently rated as '2' to the same extent that

those tokens with an average of 1 were consistently rated as '1'. Although not unexpected, this result conclusively shows that tokens with more ambiguous (L) productions are coded less reliably than tokens with an unambiguous (L) production, across a large sample of trained linguists.

The identity of the preceding vowel was also a factor predicting inter-rater agreement for a given (L) token, across combined A and B data. The results show that inter-rater variability is higher for tokens that follow a low vowel, and lower for all other tokens (front: b = -0.144, t(89) = -2.011, $p < 0.048$, ow: b = -0.186, t(89) = -2.306, $p < 0.024$, uw/U: b = -0.192, t(89) = -2.589, $p < 0.011$). In other words, unsurprisingly, the extent of vocalization for a given (L) token is most ambiguous when following the environment it most closely resembles, acoustically: low back vowels. Note that the speakers in this study were all members of communities in which the low back vowel merger (THOUGHT lowering to LOT) is in progress, with individually variable production of that merger. So, although the roundedness that accompanies vocalization makes (L) more acoustically similar to (back, rounded) /ɔ/ than to (central, unrounded) /a/, the acoustic similarity between (L) and /a/ was sufficient to interfere with the perception of vocalization.

The specific survey, A versus B, also correlated with inter-rater variability at the level of the token. The results show that the overall variance of Survey B was higher than Survey A, in the sense that there was more variance between raters per token in B than per token in A (SurveyB: b = 0.138, t(89) = 2.646, $p < 0.010$). This result could be due to our less controlled recruitment methods, and the greater variability in demographic profiles among the Survey B linguists. On the other hand, the greater variability might be an artifact of there being 2/3 of the number of tokens in B as are in A.

4.3 Vocalization differences across tokens

While we were testing the factors that correlate with inter-rater reliability across raters and across tokens, we decided to also conduct a typical sociolinguistic analysis, to the extent that was possible. In other words, what are the factors that correlate with the extent of (L)-vocalization for the tokens in this (albeit limited) dataset? Although we are looking at two distinct speech communities, there may be interesting commonalities across them, as well as interesting differences between them.

The strongest effect on vocalization rating was the identity of the preceding vowel (Figure 6). Specifically, those (L)s following front vowels were rated as less vocalized than those following all other vowels (front: b = -0.6747, t(89) = -3.107, $p <$ 0.003). This result obtained for both Columbus and San Francisco datasets, and supports previous findings in other studies of vocalization in English (Durian 2008; Borowsky 2001) and Romance languages (Recasens 1996). Note that this finding singles out front vowels from all other vowels. In contrast, when preceding vowel was considered with regard to inter-rater variance, low vowels were the most variable across raters.

- - [Insert Figure 6 about here] - -

Figure 6: Correlation between preceding vowel and vocalization rating.

As discussed in the previous section, a token's vocalization was a significant predictor of its inter-rater variance; the reverse situation also applies. Those tokens that were heard as being least vocalized or most vocalized were rated most consistently, and those tokens that were more ambiguous were, not surprisingly, those with the greatest standard deviation (stdev: $b = 0.8538$, $t(89) = -2.564$, $p < 0.012$). This was again true for both the Columbus and San Francisco datasets.

- - [Insert Figure 7 about here] - -

Figure 7: Correlation between region and vocalization rating.

Vocalization was differentially rated between the Columbus and San Francisco speakers, however; a speaker's dialect region was a significant predictor for a token's vocalization rating. As seen in Figure 7, San Franciscan speakers were overall rated with lower vocalization scores than the Ohio speakers (sanfran: $b = -0.4645$, $t(89) = -2.988$, $p < 0.004$). This is an interesting difference, the analysis of which is beyond the scope of the present paper. However, we would hypothesize that this difference is not reflecting a difference in the rate of a change-in-progress; in other words, we would not argue for what might be the expected interpretation of Columbus speakers being further along in a change towards vocalization than San Franciscans. Rather, despite vocalization being widespread across English varieties, we posit very different linguistic and social motivations for vocalization in these two communities, both interacting with change-in-progress in different ways. In Columbus, the vocalized variant is linked to race (Durian 2008; Fix 2008, 2011), class (Fix 2011), and urbanity (Dodsworth 2005), while in San Francisco, the vocalized variant correlates with Asian American ethnicity and, among Asian Americans, with a speaker's age of acquisition of English; it correlates with race but in an entirely different way as in Columbus (Hall-Lew 2010).

The factors that did not correlate with vocalization are the survey itself (A or B) and whether the (L) was syllable-final (64 tokens) or in a consonant cluster (36 tokens). The former result is encouraging, because despite the finding that the responses to B were more variable than for A, this result suggests that we did design parallel surveys that represented a similar range of (L) productions. The latter result may be due to skew, since almost twice as many tokens were coda (L) than were coda-cluster (L); on the other hand, it may suggest that vocalization in these communities is perceptually unrelated to whether or not the (L) is syllable final, contrary to Borowsky's (2001) predictions that coda-clusters should show more vocalization than codas due to syllable structure constraints.

5. Qualitative Results

Written feedback from our participants further informs the interpretation of our quantitative results, suggesting potential sources of inter-coder variability and

potential reasons for the difficulty of the task. Not every coder provided comment, and some only commented on individual tokens (if the sound quality was particularly bad, for example). Most of those who did comment did so at the very end of the survey, and their comments are thematically summarized here. If we were successful in our goal of representing the real experiences of sociolinguists coding (L) data, then the feedback received is not (entirely) the fault of the survey design, but can shed light on current techniques of (L)-vocalization coding in sociolinguistics.

The most frequent comments were about the general difficulty of coding for (L)-vocalization, a well-known complaint that was, of course, the inspiration for the present study. Some participants complained of the specific difficulty in matching (L) productions to a single dimension, a clear-to-vocal continuum, arguing that (L) was varying along more than one dimension: coronal, lateral, and velar contact were all independently at issue. In some ways, these comments meshed with our own observations about the dialect-specific patterns of vocalization in our stimuli, which are discussed further below. Other coders questioned the validity of our discretization having four levels: some wished for fewer levels (because they felt unable to distinguish between four levels), while others wished for more than four (because they felt they could distinguish).

Only four of the coders specifically commented on our providing two versions of the token: the sentence level and the syllable level. However, while two of coders preferred the longer version, feeling that more context helped them better perceive the (L) quality, the other two coders in fact preferred the short, syllable-length version, feeling it allowed them to focus more specifically on the (L) and thus better perceive its quality. Several other coders commented on the duration of sample, both in terms of the stimuli length and the length of the syllable nucleus in which the (L) appeared. Coders noted that shorter syllables and faster rates of speech made coding more challenging, again highlighting the fact that vocalization coding is a particularly acute challenge for the kinds of data sociolinguists analyze.

The training stage of the survey highlighted the fact that many linguists do not agree on a single perceptual definition of vocalization, and that many of those who participated in our study did not agree with the scale established in the pilot study stage. Some of the non-American coders remarked on the very velarized production of the American coda (L), such that the envelope of (L) variability in U.S. English is already shifted more towards vocalization than in other varieties of English. One coder commented on the training page, in reference to the 'strong' (L)s we presented, "All of the tokens on this page sound pretty vocalized to me." Two other coders wrote:

> *One thing I've been noticing about Canadian English is that Ls here are less velarized than in US English, which may have made me judge these American talkers as being more vocalized than raters located in the US.*
>
> *I don't think I mean the same thing by l-vocalization as you do. I would consider the way a lot of Standard Amer. Eng. pronunciations in coda position come out, with no alveolar closure, to be vocalized, but they fall at 1-2 on your scale. So I started listening for the vowel differences that go with the dialects, as well as for the l itself.*

Only two participants explicitly commented on the role that their own dialect background and exposure might have played in their coding; only one said that they relied on their own production of vocalization in making coding decisions.

However, the use of vowel quality as a cue to vocalization was a common strategy – one that has important implications for how we interpret the effects of vowel environment reported previously. One coder noted a difference in perception depending on whether the vowel was tense or lax. Another coder specifically noted the difficulty with low vowels: "I'm not sure I can tell the difference between /aw/ and /awl/." A third coder commented on their variable perception of formant trajectories as being related to the vowel "(so I coded them as 4)" versus related to "the articulation of the liquid (so I coded them as 2 or 3)."

Perhaps the most interesting issue raised in the qualitative results was the effect of perceived speaker identity on perceived token vocalization. This was a factor tested for in Yaeger-Dror, et al.'s (2009) study on (r), namely, would other cues to dialect in a stimulus, such as 'British' vowel patterns heard by an American linguist, influence coding decisions, such as the American hearing more (r)-vocalization than is actually produced? The issue is a rather serious one, given that sociolinguists who do their own coding not only know the ethnic or regional identity of their speakers, but sometimes also have a very intimate, ethnographic awareness of these speakers. To what extent do our expectations influence our judgments?

The present study cannot answer this question directly, but qualitative feedback from participants suggests a clear further direction of study. While we did not specifically choose stimuli based on the other dialect features in the speech stream (nor were any 'famous' voices used; cf. Yaeger-Dror et al. 2009), many coders commented that the presence of any salient features possibly influenced their coding decisions. Participants attributed these effects to both the ethnicity and regional dialect of the speakers, which were conflated in this study: Ohio voices were, more or less, correlated with features associated with African American varieties, and San Francisco voices were correlated with Asian American speakers. One participant said, "I feel like perceived ethnolect/ethnicity influenced my judgments," and another, "definitely an effect of overall dialect impression when I was coding, which I tried to correct for." One coder said, even more explicitly:

> *I was probably more likely to code AAVE tokens as vocalized, because I know that it is a very common feature of AAVE. Tried not to do that, but it's unconscious, you know?*

Based on receiving feedback like this in survey A, we piloted a small ethnic identification task for survey B.

6. Vocalization and Ethnic Identity

In this task, participants heard ten additional tokens at the end of the survey, one from each of the ten speakers represented in the survey. Rather than rate the vocalization of (L) for those ten tokens, survey B participants were asked to code for the ethnic

identity of the speaker. This was a force-choice task with only three options: African American, Asian American, or European American. We did not want to add the ethnic identification task to each token of the survey, which would have made the time to completion twice as long. Instead, we put the task at the end of the survey, thinking that coders would have built up increasingly stronger perceptions about the speakers as they progressed through the survey (one participant from survey A had, in fact, commented: "I think I started to form expectations about your individual speakers by about halfway through"). We then compared the overall ethnic identification ratings, for each speaker, with the average vocalization rating for each speaker (on a scale of 1 = never heard as vocalizing, to 4 = always heard as vocalizing).

- - [Insert Figure 8 about here] - -

Figure 8: Perception of speaker ethnicity correlates with perception of speaker vocalization.

The results show a clear correlation between the perceived ethnicity of the speaker and the average vocalization rating for a speaker, at least the racial level of perceived white versus perceived non-white. The speakers who were most often heard as either African American or Asian American were also heard as vocalizing more often than the speakers heard as European American. This result is shown in Figure 8, which presents the results for the Columbus speakers on the left and the San Franciscans on the right. The number next to the speaker name indicates that speaker's average vocalization rating. It is important to remember that it is the perception of ethnicity, and not the speaker's 'actual' ethnicity, that forms this correlation. 'Lori' is African American; 'Tiffany', 'Paula', 'Monica', and 'Pam' are European American (but have had sustained close social contacts with African Americans—in the case of Tiffany and Paula, since early childhood); the other five speakers (the San Franciscans) are all Asian American. As can be seen in Figure 8, perceptions of speakers ethnicities varied quite a bit with respect to the speakers' own ethnic identity. In short, these results support the arguments made elsewhere (Fix 2008; Hall-Lew 2010; Hall-Lew & Starr 2010) that (L)-vocalization in these communities indexes ethnic meanings. However, the results from this task are quite limited, and the implications for perceived ethnicity on the general coding decisions still remain to be seen.

7. Discussion

The most encouraging result from our analysis is that most of the linguists surveyed were in relatively close agreement with one another, particularly with respect to the (L)s heard as the most consonantal and those heard as the most vocalized. This suggests that perceptual coding is a valid measure for (L)-vocalization studies, to a certain extent. Unfortunately, those tokens that are probably the most sociolinguistically interesting – exhibiting intermediate stages of realization – are those that the coders agreed on the least. This is not necessarily a surprising finding, but it does present a formidable challenge to vocalization research.

The results also showed that linguists from both vocalizing and non-vocalizing dialect regions appear to be able to equally reliably code for (L)-vocalization. We found no consistent difference in coding strategies between American and British coders, or between native and non-native English speakers. Coders with more years of sociophonetic training were no more likely to conform to the decisions of the sample mean than coders with minimal training. These results contrast, in part, with Yaeger-Dror et al.'s (2009) findings for (r) coding, which suggest that a linguist's exposure to (r) variability correlate with their rating of (r) vocalization. While the results from the present study are in some ways very encouraging, it is worth noting that several coders did diverge substantially from the mean. However, they did so for no discernable social reasons. Until we are able to identify any factors that predict an individual's ability to code vocalization in agreement with the mean, this finding must serve as a cautionary note for all researchers who use auditory methods for (L) coding. Our results suggest that employing multiple coders over the same subsample of data, and spending time coming to a group consensus with respect to coding decisions, is certainly warranted.

8. Future Directions

Based on feedback from our participants in the coding survey, we have charted next steps to work towards a better understanding of how (L) realizations are perceived and how this impacts the reliability of auditory coding processes. We would like to examine the effect of other phonetic factors on (L) realization perception, such as syllable duration or the position of the syllable in the intonational phrase. Longer (L) syllables may be easier to rate, and phrase-medial tokens may be more difficult to rate, and more likely to vocalize, than phrase-final tokens.[2] The frequency of the lexical item is another potential correlate of vocalization (see van Reenen 1986:192; Scobbie & Pouplier 2010). One of the most obvious phonetic factors that remains to be tested is the influence of the following segment, which has been found in previous studies to correlate with extent of vocalization (Borowsky 2001; Stuart-Smith et al. 2006).

Another factor that seemed to impact coder perception of (L)-vocalization was their perception of the speakers' ethnic and regional identities. This effect can be straightforwardly investigated in the future with the addition of a between-subjects test using the same stimuli but asking participants to identify the ethnicity of the speaker for each token. Analyzing dialect-specific differences in *how* (L) is vocalized—for example, whether (L) becomes vocalized as a back rounded vowel versus a schwa, as it is sometimes realized in African American English—will also help to inform how dialect-specific differences in the production of vocalized (L) impact how the variable is coded. Ultimately, we hope that our combination of two sets of (L)-vocalization examples from two distinct English dialect areas will lead to a consideration of dialect-specific differences in how (L) is vocalized in global varieties of English.

9. Implications for the Development of Acoustic Measurements

A triangulation of methodological approaches – auditory, acoustic, and articulatory – to the quantification of (L)-vocalization is the ideal outcome of this research. The

results from the present study of auditory coding suggest a few important factors for the development of an acoustic method of analysis, many of them related to the issues presented for further consideration in the previous section.

One of our clearest results was that the identity of the preceding vowel influenced both the perception of a token's vocalization and the consistency of a token's rating across multiple coders. Development of an acoustic measure of (L)-vocalization may find that this points to an acoustic difference between realizations of (L) across vowel environments. Syllable-final consonantal (L) has a strong co-articulatory effect on the preceding vowel (West 1999), and vocalization would result in the loss of the coronal and lateral aspects of that co-articulation, which would have different implications for different vowels. Since velar (L) has the formant structure of a back rounded vowel, at least in some dialects (Gick, Kang, & Whalen 2002), co-articulatory changes to a preceding back, rounded vowel would be less than changes to a preceding front, unrounded vowel. Any acoustic (or articulatory) method that is developed for the coding of (L)-vocalization may need to depend directly on the identity of the preceding vowel. Just as sociophonetic studies of vowels must always take into account the features of the following consonant, so too must studies of (L) take into account the features of the preceding vowel.

Similarly, we argue here that any method of measurement must take into account dialect differences in the realization of vocalized (L). For example, in our datasets, San Franciscans never vocalize to schwa, while Columbus speakers variably do; San Franciscans also rarely vocalize after front vowels, while Columbus speakers variably do. In relation to preceding vowel effects, regional differences in vowel production also complicate (L)'s conditioning of vowel production. For example, many mergers are conditioned by a following (L) (Thomas 2001), and (L) has been found to block back vowel fronting (Luthin 1987), although there may be an interactive relationship between increased rates of (L)-vocalization and back vowel fronting (Dodsworth 2005). We should also be careful to not assume that (L)-vocalization constitutes the same sociolinguistic variable in every variety of English. Not only is it questionable to treat vocalization in Australian, British, and American contexts as being comparable evidence of a 'global' sound change, but the results from previous work (Fix 2008; Hall-Lew 2010), suggest that vocalization in two different areas of the United States results from two entirely different sociolinguistic situations. It may be that future acoustic or articulatory methods are more appropriate for some varieties than others; in any case, we are only at the beginning of a discussion about what exactly constitutes the envelope of variation that we call '(L)-vocalization'.

In summary, we argue that any acoustic approaches to measuring (L)-vocalization must be adaptable to varying *internal linguistic environments* as well as varying *speech communities.* Further, it is not obvious that any acoustic method developed in the future will be any more fine-grained a measure than impressionistic coding already is. In part this is because, as sociolinguists, we want to measure differences in (L) realization that index social differences, so acoustic measures may need to vary from study to study, depending on both the range of production of (L) in the community in question, as well as the phonetic and social salience of the different variants. Acoustic measures would best be designed around the question of how fine-grained differences in (L) realization must be before those differences become unable to carry social meaning.

10. Conclusions

Increased interest in (L)-vocalization as a sociolinguistic variable encourages us to pursue more reliable methods of measurement and coding. For numerous reasons, sociolinguists interested in vocalization have preferred auditory coding rather than articulatory or acoustic measures. However, consistency across coders both within and between studies is of particular methodological concern with auditory coding because decisions are more impressionistic than with articulatory or acoustic measures. In order ensure consistency across multiple coders, researchers must operationalize the range of (L) variation in their data and set unambiguous criteria for doing so. Overall, the results of our survey task show an encouragingly high degree of reliability in impressionistic coding of (L)-vocalization across coders and data sets. In our sample, most of the participants were in relatively close agreement with one another, particularly with respect to the most consonantal and the most vocalized (L)s. This suggests that perceptual coding is indeed a valid measure for (L)-vocalization studies, except for those productions phonetically in-between consonantal (L) and vocalized (L). Unfortunately, those 'in-between' productions are by far the most frequent, as well as being in many ways the most sociolinguistically interesting. While it is not surprising that these productions were rated with the least amount of inter-coder consistency, it does present a formidable challenge to vocalization research.

Encouragingly, linguists from both vocalizing and non-vocalizing dialect regions, as well as non-native English speakers, appear able to reliably code for (L)-vocalization. The coders who diverged substantially from the mean did so for no discernable social reasons. These results suggest that a researcher can code for (L)-vocalization whether or not they are an outsider to the community under analysis. Again, this is likely to be most true at a coarse level of distinction: consonantal (L) versus vocalized (L). It remains an open question if more subtle or ambiguous realizations can be coded equally reliably by vocalizers and non-vocalizers, native and non-native English speakers, or community insiders and community outsiders. Any such differences between coders may be mitigated within a single study by including an initial stage of cross-coding a subset of the data until consensus is reached; the utility of such training remains to be thoroughly tested. There also remains a crucial question of comparability between different (L)-vocalization studies. The multi-stage piloting and multi-coder tasks exemplified in this paper may be useful as a reference point by others coding (L) to help to clarify and quantify the range and direction of differences between individual coders. Furthermore, the actual stimuli used in the perception task presented here may be used by other research teams for calibrating their own coding decisions.

The quantitative evidence in the present paper shows that, on average, linguists from a wide range of backgrounds can impressionistically code (L)-vocalization data from two different speech communities with an encouraging level of reliability. At the same time, the results also show individual variability in coding strategy: some linguists consistently judge all tokens as more, or less, vocalized than other linguists judge those tokens to be. In other words, even highly trained linguists differ in their perception of variants of (L). Although no social factors about the coders correlated

significantly with these directional preferences, the qualitative feedback from the coders' comments are highly suggestive, namely indicating that a linguist's knowledge about a speaker's regional and ethnic background may be likely to influence their coding decisions. Not only is such information always known, a priori, by every coder of every (L)-vocalization study so far, but salient sociolinguistic features which co-occur in the speech stream provide further, on-line cues to speaker identity. Our quantitative evidence show a correlation between the perceived ethnicity of a speaker and the average extent of vocalization for that speaker, but it is difficult to disentangle two interpretations of this correlation: that (L)-vocalization is a robust indexical cue to non-white ethnic meanings in U.S. English, or that linguists expect speakers with ethnically-marked speech styles to vocalize more often than speakers whose speech is less marked. Our research suggests that both are likely to be true, but the ability to argue for the former depends crucially on the ability to control for the effects of the latter. The implication here is that sociolinguists must critically examine our coding methodologies, regardless of the variable under study or the method of analysis used. Every linguist brings both professional knowledge and community-specific sociolinguistic knowledge to bear on their coding decisions. The success of our analyses depends on developing an awareness of how that knowledge is potentially affecting those decisions. Studying particularly challenging variables, such as (L)-vocalization, encourages us to develop more rigorous strategies to do so.

As we continue to refine methods for coding the production of (L)-vocalization in spontaneous speech, we come closer to being able to ask: What is (L)-vocalization, sociolinguistically? To what extent is it phonetically the same across regional dialects? How do we best describe those differences in phonetic realization that we perceive? Do all communities with vocalization exhibit the same range of phonetic variability? In what ways does vocalization index similar social meanings, across communities? In what ways do phonetic differences across communities index social differences? How subtle can phonetic variation be for it to carry socioindexical meaning? We hope the present paper is one step forward in developing tools that allow us to address these questions, encouraging the continued development of sociophonetic methodologies for identifying, measuring, interpreting, and representing instances of (L)-vocalization across the English-speaking world.

Footnotes

1. We unfortunately did not collect information regarding how many times a listener played a particular token, and some of the variance in our data may be due to some stimuli being listened to more often than other stimuli, or some linguists listening to all stimuli more often than other linguists. Thanks to Jane Stuart-Smith for making this point, which we hope to address in future work.
2. Thanks to Tamara Rathcke, as well as two of our anonymous survey coders, for making this point, which we hope to address in future work.

Acknowledgments

Different versions and stages of this work were presented at the following conferences: *New Ways of Analyzing Variation 39* (NWAV39), *Experimental*

Approaches to Perception and Production of Language Variation (ExAPP), and *The 85th Annual meeting of the Linguistics Society of America* (LSA). Subsets of the work were also presented to the Glasgow University Laboratory of Phonetics and Newcastle University's School of English Literature, Language & Linguistics. We would like to thank all the audiences of these venues for their questions and comments. Most importantly, we would like to thank the linguists who participated in this research.

References

Ash, S. 1982. The vocalization of /l/ in Philadelphia. Doctoral dissertation, University of Philadelphia, Pennsylvania, PA.

Bailey, G., and E. R. Thomas. 1998. Some aspects of African-American Vernacular English phonology. In Salikoko S. Mufwene, John Rickford, John Baugh, and Guy Bailey, (eds.), African American English. London: Routledge. 85-109.

Banerjee, M., M. Capozzoli, L. McSweeney, & D. Sinha. 1999. Beyond Kappa: A Review of Interrater Agreement Measures. The Canadian Journal of Statistics / La Revue Canadienne de Statistique, 27(1): 3-23.

Bauer, L. 1986. Notes on New Zealand English phonetics and phonology. English World Wide, 7, 225-258.

Bauer, L. 1994. English in New Zealand. In R. Burchfield (ed), The Cambridge History of the English Language. English in Britain and Overseas: Origins and Development, 5. Cambridge: Cambridge University Press. 82-429.

Boersma, P. 2003. Praat, 4.0. http://praat.org.

Borowsky, T. 2001. The vocalisation of dark-l I Australian English. In David Blair and Peter Collins (eds), English in Australia. Amsterdam/Philadelphia: John Benjamins. 69-87.

Borowsky, T., and B. Horvath. 1997. L-vocalisation in Australian English. In Frans Hinskens, Roeland van Hout, and W. Leo Wetzels (eds.), Variation, change and phonological theory. Amsterdam: John Benjamins. 101–123.

Dodsworth, R. 2005. Linguistic Variation and the Sociological Imagination. PhD Thesis, The Ohio State University, Columbus, OH. Ann Arbor: ProQuest/UMI. (Publication No. 3192844)

Dodsworth, R., B. Plichta, & D. Durian. 2006. An Acoustic Study of Columbus /l/ Vocalization. Paper presented at NWAV 35, The Ohio State University.

Durian, D. 2008. The Vocalization of /l/ in Urban Blue Collar Columbus, OH African American Vernacular English: A Quantitative Sociophonetic Analysis. OSUWPL Volume 58: 30-51.

Fix, S. 2004. /l/ vocalization and racial integration of social networks: Sociolinguistic variation among whites in a Columbus Ohio community. Poster presented at NWAV 33, Ann Arbor, MI.

Fix, S. 2008. Beyond Stereotypes: White women, black worlds, linguistic variation and style. Paper presented at NWAV 37, Houston, TX.

Fix, S. 2011. "Dark-Skinned White Girls": Linguistic and Ideological Variation Among White Women with African American Ties in the Urban Midwest. PhD thesis, New York University, New York, NY. Ann Arbor: ProQuest/UMI.

Gick, B., A. Min Kang, and D. H. Whalen. 2002. MRI evidence for commonality in the post-oral articulations of English vowels and liquids. Journal of Phonetics, 30, 357-371.

Gick, B., and I. Wilson. 2006. Excrescent Schwa and Vowel Laxing: cross-linguistic responses to conflicting articulatory targets. In L. Goldstein, D. H. Whalen and C. T. Best (eds). Laboratory Phonology, 8: 635-59.

Green, L. 2002. African American English: A Linguistic Introduction. Cambridge: Cambridge University Press.

Hall-Lew, L. 2009. Ethnicity and Phonetic Variation in a San Francisco Neighborhood. PhD thesis, Stanford University, Stanford, CA. Ann Arbor: ProQuest/UMI. (Publication No. 3382940)

Hall-Lew, L. 2010. L-vocalisation in Chinese American English. Paper presented at Sociolinguistics Symposium 18, 1-4 September, Southampton, UK.

Hall-Lew, L. and R. L. Starr. 2010. Beyond the 2nd Generation: English use among Chinese Americans in the San Francisco Bay Area. In a special issue on "Social and Linguistic State of 2nd Generation Americans," English Today, 26(3):12-19

Hardcastle, W., and W. Barry. 1985. Articulatory and perceptual factors in /l/ vocalizations in English. Journal of the International Phonetic Association. 15: 3-17

Hazen, K. and R. Dodsworth. 2011. Following L over hill and dale: Changes in L-vocalization through, space, time, and methods. Paper presented at Methods in Dialectology 14, University of Western Ontario, London, Ontario, Canada.

Heselwood, B., and L. McChrystal. 2000. Gender, accent features and voicing in Panjabi-English bilingual children. Leeds Working Papers in Linguistics and Phonetics 8: 45-70.

Horvath, B. M., and R. J. Horvath. 2002. The geolinguistics of /l/ vocalization in Australia and New Zealand. Journal of Sociolinguistics 6: 319-346.

Johnson, W., and David Britain. 2007. L-vocalisation a natural phenomenon: explorations in sociophonology. Language Sciences, 29, 294-315.

Lawson, E., J. Stuart-Smith, J. M. Scobbie, M. Yaeger-Dror, and M. Maclagan. 2011. Liquids. in M. Di Paolo and M. Yaeger-Dror, eds., Sociophonetics: A Student's Guide. New York: Routledge. 72-86.

Maryn, Y., N. Roy, M. De Bodt, P. Van Cauwenberge, and P. Corthals. 2009. Acoustic measurement of overall voice quality: A meta-analysis. Journal of the Acoustical Society of America, 126(5): 2619–2634.

McElhinny, B. 1999. More on the Third Dialect of English: Linguistic Constraints on the use of three phonological variables in Pittsburgh. Language Variation and Change, 11, 171-195

Przedlacka, J. 2001. Estuary English and RP: Some recent findings. Studia Anglica Posnaniensia, 36, 35-50.

Recasens, D. 1996. An articulatory-perceptual account of vocalization and elision of dark /l/ in the Romance Languages. Language and Speech, 39(1), 63–89.

Recasens, D., and A. Espinosa. 2009. The Role of the Spectral and Temporal Cues in Consonantal Vocalization and Glide Insertion. Phonetica, 67, 1-24.

Scobbie, J. M. and M. Pouplier. 2010. The role of syllable structure in external sandhi: An EPG study of vocalization and retraction in word-final English /l/. Journal of Phonetics, 38, 240-259.

Scobbie, J. M., J. Stuart-Smith, & Eleanor Lawson. 2008. Looking variation and change in the mouth: developing the sociolinguistic potential of Ultrasound Tongue Imaging Research Report for ESRC Project RES-000-22-2032.

Scobbie, J. M., M. Pouplier, and A. A. Wrench. 2007. Conditioning factors in external sandhi: An EPG study of English /l/ vocalisation. In Proceedings of the XVIth ICPhS, Saarbrucken, 441–444.

Scobbie, J. M., and A. A. Wrench. 2003. An articulatory investigation of word-final /l/ and /l/-sandhi in three dialects of English. In Proceedings of the XVth ICPhS, Barcelona, 1871–1874.

Sivertsen, E. 1960. Cockney Phonology. Oslo: Oslo University Press.

Sproat, R. and O. Fujimura. 1993. Allophonic variation in English /l/ and its implications for phonetic implementation. Journal of Phonetics, 21, 291-311.

Stuart-Smith, J., C. Timmins and F. Tweedie. 2006. Conservation and innovation in a traditional dialect: L-vocalization in Glaswegian, English World Wide, 27:1, 71-87.

Thomas, E. 2001. An Acoustic Analysis of Vowel Variation in New World English. A Publication by The American Dialect Society (85). Duke: Duke University Press.

Timmins, C., F. Tweedie and J. Stuart-Smith. 2004. Accent change in Glaswegian (1997 corpus): Results for consonant variables. Department of English Language, University of Glasgow.

Tollfree, L. 1999. South East London English: Discrete versus continuous modelling of consonantal reduction. In P. Foulkes, & G. J. Docherty (Eds.), Urban voices (pp. 163–184). London: Arnold.

Trudgill, P., 1986. Dialects in Contact. Blackwell, Oxford.

van Hofwegen, J. 2010. Apparent-time evolution of /l/ in one African American community. Language Variation and Change.

van Reenen, P. 1986. The vocalization of /l/ in Standard Dutch, A Pilot study of an ongoing change, in Frits Beukema and Aafke Hulk, Linguistics in the Netherlands. Foris: Dordrecht, 189-98.

Wells, J. C. 1982. Accents of English. Cambridge: Cambridge University Press.

West, P. 1999. Perception of distributed coarticulatory properties of English /l/ and /r/. Journal of Phonetics, 27:405-426

Wrench, A. A., and J. M. Scobbie. 2003. Categorising vocalisation of English /l/ using EPG, EMA and ultrasound. In Proceedings of the sixth international seminar on speech production, Sydney.

Wright, S. 1989. The effects of style and speaking rate on /l/ vocalisation in local Cambridge English. York Papers in Linguistics 13: 355-365.

Yaeger-Dror, M., T. Kendall, P. Foulkes, D. Watt, Jillian Oddie, P. Harrison, and C. Kavanagh. 2009. Perception of ‘r’ by trained listeners. Paper presented at the 83rd meeting of the Linguistics Society of America, January, San Francisco, CA.