Interspeech 2024
1-5 September 2024, Kos, Greece

Automatic Speech Recognition with parallel L1 and L2 acoustic phone models to evaluate /l/ allophony in L2 English speech production

Anisia Popescu[1,2,3], *Lori Lamel*[1,2,3], *Ioana Vasilescu*[1,2,3], *Laurence Devillers*[1,2,3]

[1]LISN, France
[2]CNRS, France
[3]Université Paris Saclay, France

anisia.popescu@universite-paris-saclay.fr

Abstract

The acoustic and articulatory characteristics of the syllable position lateral allophony in English (clear /l/ in onsets vs. dark /l/ in codas) have been well documented. The present study tests whether speech technology derived methods can be used to evaluate lateral allophony in L2 English production, by combining classic acoustic analyses and automatic speech recognition (ASR).

In this study, an ASR system is forced to choose between English and French /l/ acoustic phone models when force-aligning a corpus consisting of read English texts by 43 L2 French learners. The output is correlated with a staple measure for /l/ darkness: the difference between the second and first formants (F2-F1).

Results show that segments aligned with the French /l/ acoustic model correspond to "clearer" /l/s (i.e. higher values of F2-F1) suggesting automatic, less time consuming methods of speech processing could be used to identify L1 transfer in L2 production.

Index Terms: second language speech production, allophonic variation, acoustic models, forced alignment, pronunciation variants

1. Introduction

Acquiring native-like proficiency in a second language is no easy feat. Difficulties arise at multiple levels: lexical, semantic, syntactic, phonological. The latter is particularly difficult, especially if the second language was acquired later in life. Languages not only have different phonological inventories, but a same phoneme can have variable phonetic implementations across languages. For example, the voiceless stop /t/ is produced with different places of articulation in French (dental) vs. English (alveolar) [1]. Furthermore, within a single language phonological systems show substantial variation - the same phoneme can have different phonetic implementations depending on the context. This is also known as contextual allophony. For example, in English, /t/ appears as aspirated in word initial position (*tool* [t^{h}u:ɫ]) and as unaspirated in a word initial consonant cluster following /s/ (*stool* [stu:ɫ]). In order to acquire native-like pronunciation L2 learners must acquire not only phonemic and phonetic, but also allophonic differences.

The present paper focuses on L2 French learners' production of a well known case of allophonic variation in English: the positional allophony of the lateral consonant /l/. In English the lateral consonant allophony is conditioned by the position in the syllable: clear /l/ in onsets and dark /l/ in codas. In French there is no lateral consonant allophony, there is only one phonetic implementation of the /l/: clear /l/ in all syllable positions. To investigate if and how French learners of L2 English produce this allophony we combine classic acoustic analyses and automatic speech recognition (ASR) methods, which have increasingly been used to improve pronunciation in second language learning [2, 3, 4, 5]. ASR dictation software is used to focus on pronunciation errors at the segmental level (e.g. vowel and consonant minimal pairs within a language: for example tense/lax minimal pairs in English [3] or fricative/plosive minimal pairs in Dutch [4, 5]). In this paper we force an ASR system to choose between parallel English (L2) and French (L1) acoustic models for the same phoneme /l/ when force-aligning a corpus consisting of read English texts by 43 L2 French learners. The output of the ASR alignment is correlated to formant measures in the interest of having a better understanding of how the system chooses the different variants, and of determining whether the proposed method can be applied to identify pronunciation errors in second language learning.

1.1. Laterals in English vs. French

Many English dialects contrast two varieties of /l/ depending on syllable position: clear /l/ in syllable onsets (*leap* [li:p]) and dark /l/ in syllable codas (*peal* [pi:ɫ]). The two varieties have been extensively described for American English [6, 7, 8, 9, 10]. The main difference is articulatory: dark /l/ is produced with a tongue dorsum retraction towards the uvular region that precedes a coronal constriction; clear /l/ is produced with a simultaneous tongue dorsum lowering (i.e. no retraction towards the uvular region) and coronal constriction. This articulatory difference translates acoustically in dark /l/ having a lower second formant (F2 $\approx$ 800-1200 Hz) and clear /l/ having higher F2 ($\approx$ 1500-2000 Hz) [6, 11]. The first formant (F1) is also different: typically higher for dark /l/ and lower for clear /l/ due to tongue height differences between the two variants [12]. A classic measure of /l/ darkness is defined as the difference between the second and first formants (F2 - F1). A smaller difference corresponds to darker /l/s and a higher difference corresponds to clearer /l/s. Figure 1 illustrates the produced formant structure of clear and dark /l/ (framed in black boxes) by a native American English speaker (F). While for clear /l/ (on the left) F2 and F1 are further apart, for dark /l/ (on the right) F2 and F1 are closer together. Languages usually have one or the other variant (clear /l/: German, Spanish; dark /l/: Catalan, Portuguese, Russian). Furthermore, dark and clear /l/s across languages do not have a binary distribution but a gradual one, with languages presenting clearer or darker lateral consonants [11].

In French, similar to most languages and contrary to English, there is no allophonic variation for lateral consonants. The /l/ is always clear, in all syllable positions. In the case of French, the /l/ is clearer (i.e., F2 - F1 values are higher) than both English

10.21437/Interspeech.2024-1014

clear and dark /l/ [11].

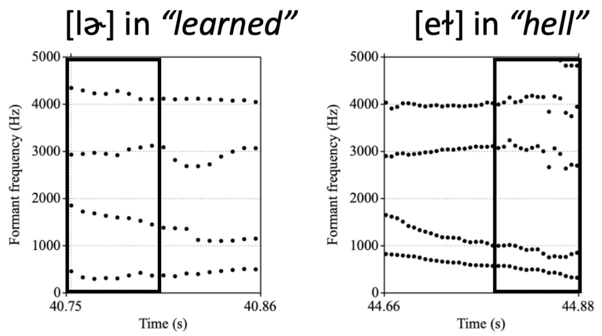

Figure 1: *Formant structure of [lɚ] vs. [eɫ] produced by a native American English speaker. Onset clear /l/ and coda dark /l/ are framed by black boxes.*

1.2. Acquisition of English lateral allophony

Contrary to the acquisition of L2 phonemic contrast, the acquisition of allophonic variation is a much less studied phenomenon. More recently several laboratory studies have focused on the non-native production of syllable position lateral allophony in English (L1 French: [13, 14]; L1 French & L1 Spanish [15]; Spanish-English bilinguals: [16]; L1 Japanese: [17]). These studies investigated acoustic and/or articulatory recordings of words containing /l/ read in isolation and in controlled carrier sentences. In the current study we look at read texts of different levels of difficulty (beginner, intermediate, advanced) to answer the following questions:

- How do French L1 speakers produce the syllable position lateral allophony in L2 English?
- Can ASR with parallel L1 and L2 acoustic models be used to identify L1 transfer in L2 production?

To our knowledge this the first study using ASR in combination with acoustic measures to evaluate the production of lateral consonant allophonic variation.

2. Methods

2.1. Corpus and acoustic analysis

To answer the questions presented above we analyzed a corpus consisting of read speech by 43 French L2 English learners. All participants (24 female and 19 male) were recorded reading the same three beginner- ("A Happy Visitor), intermediate- ("Time with Grandpa") and advanced-level ("Fried") texts available on the "English for Everyone" website (see [18]). Faithful orthographic transcriptions (including substitutions, repetitions, truncations, hesitations) were available for the entirety of the data. The acoustic signals of the productions were forced aligned using both an open source (WebMAUS) and a lab-internal aligner. Both aligners used English US as a reference language and performed similarly well. Gross misalignments of our target tokens were rare (in less that 5% of cases) and always involved /l/ in word internal position (e.g. Valerie, really). These misaligned words were not included in the analysis. The rest of the lateral consonant tokens were hand-corrected in Praat [19]. A total of 36 words containing singleton /l/ (17 in onset and 19 in coda position) were included in the analysis. Segmental duration and formant measures (F1, F2 and F3) were extracted at the midpoint of the lateral for all tokens. A measure of /l/ darkness was defined as the difference between the second and first formants (*/l/ darkness = F2-F1*).

2.2. Lateral acoustic models

Both French and English acoustic models were trained on similar amounts and types of data (French: 7 million word tokens, 102k word types; English: 7 million word tokens, 86k word types). Each model is a 3-state left-to-right continuous density HMM with Gaussian mixtures with up to 32 Gaussians per state. Silences are modeled by a single state with 256 Gaussians. The same cepstral (PLP) [20] and pitch (F0) features were used for the acoustic parameterization, similar to [21]. The acoustic models are all word-, context- and speaker-independent monophone models. This implies that the English lateral /l/ model contains acoustic features that correspond to the phoneme /l/ (i.e., both clear /l/ and dark /l/ allophones).
By running both French and English lateral consonant acoustic models in parallel we force the ASR system to choose the best fitting phone model (either the English /l/ or the French /l/) for each individual /l/ in the corpus. Figure 2 shows the spectrogram and alignment of the word *cell* for two different participants. The ASR system chooses different acoustic models for the two productions (French /l/ (**L**) on the left and English /l/ (**l**) on the right).

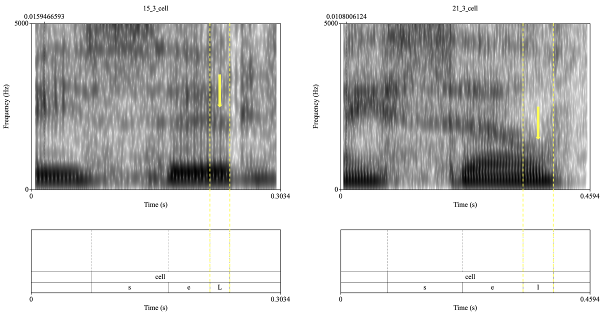

Figure 2: *Spectrograms of the word cell for participant 15_3 (left) and 21_3 (right). Yellow lines delimit the /l/ in each signal. Yellow arrows point towards F2. On the left the system chooses a French /l/ (no downward trend of F2). On the right the system chooses a English /l/ model (the F2 trajectory has a downward trend).*

2.3. Statistical analysis

To test whether */l/ darkness* (F2-F1 measures) and *acoustic model* (English vs. French) are correlated we ran a linear mixed model (*lme4* [22]) with */l/ darkness* as a response variable. Along with *acoustic model* we included other predictors, known to influence lateral formant structure: *syllable position* (onset vs. coda), *biological gender* (male vs. female) and *vocalic context* (front vs. mid vs. back vowels). A measure of *overall accent* was also added as a predictor. This variable was defined by the word-error-rate (WER) calculated using the WER() function in Matlab from the output of the unbiased ASR alignment (i.e. ASR without reference text transcription). The WER scores ranged from 29% to 59%. The same ASR system obtains WER scores of 5%-10% for noisy non-laboratory (e.g., broadcast news and telephone conversation speech data) and < 1% for laboratory *native* speech. Finally, interactions between *acoustic model* and (i) *syllable position* as well as (ii) *overall accent* were also included as fixed factors. The random factor structure included Participant with random intercept and slope for *acoustic model*.

2.4. Predictions

We make the following predictions based on previous literature (acoustic studies [11, 6]) for each of our model variables and interaction terms:

Individual variables:

- acoustic model: We expect lower F2-F1 in the case of English /l/ models
- syllable position: We expect lower F2-F1 in the case of coda positions
- biological gender: We expect overall higher values in female participants
- vocalic context: We expect the higher F2-F1 values in front vowel and the lower in back vowel context
- global accent: Lower scores (i.e., more native like pronunciation) are expected to correlate with lower F2-F1 measures in all syllable positions (French /l/ is expected to be clearer than both English lateral allophones)

Interaction terms:

- acoustic model * syllable position: We expect higher differences between the English (l) and French (L) lateral models in coda position indicating the production of clear /l/ in coda position.
- acoustic model * global accent: We expect higher differences between the English (l) and the French (L) lateral models for lower scores of WER (i.e., more native like accent).

3. Results

Results will be presented in two stages. First we describe the output of the forced alignment with parallel French and English /l/ acoustic models and then we discuss the linear model results.

Table 1: *Percentages of /l/ occurrences aligned with either the English (l) or the French (L) acoustic model.*

Acoustic Model	**Onset**	**Coda**
l: English /l/	44.3%	65.8%
L: French /l/	55.7%	34.2%

Table 1 shows the percentages of English vs. French /l/ acoustic models chosen in onset and coda position by the ASR system. For onset /l/ the French model is chosen in 55.7% of cases. For coda /l/ the English acoustic model is preferred in 65.8% of cases. Table 2 shows the distribution of /l/ darkness values (F2-F1 measures) for the English vs. the French /l/ acoustic model per syllable position.

For both English and French /l/ models onset values are higher than coda values. F2-F1 measures are also lower for /l/s identified by the system as being more English like (the ASR system chose the English /l/ model as a better fit to the acoustic output). These results indicate that to a certain degree L2 English French learners produce darker /l/s in coda position. However, some speakers retain more French-like production of the lateral (46%) independent of syllable position.

Table 2: *Distribution of F2-F1 (Hz) measures per acoustic model and syllable position.*

Acoustic model	**Syllable position**	Min.	Mean	Max.
l: English /l/	onset	277	1068	2563
	coda	117	689	1968
L: French /l/	onset	494	1373	2677
	coda	134	1114	2564

Figure 3 shows the degree of /l/ darkness (F2-F1) as a function of the French (L) vs. English (l) acoustic model and syllable position (onset vs. coda) for female and male speakers. Three patterns can be observed:

- lateral segments detected by the system as more French-like have higher F2-F1 values (/l/ is clearer)
- lateral segments detected by the system as more English-like exhibit a more pronounced difference between onset and coda /l/s
- female speakers have overall higher formant values than male speakers

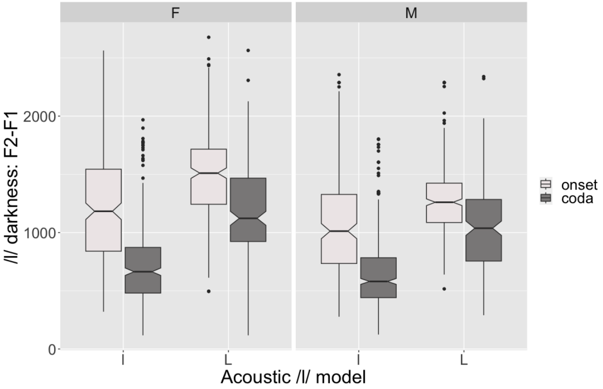

Figure 3: */l/ darkness (F2-F1) as a function of /l/ acoustic model (English (l) vs. French (L) and syllable position (onset vs. coda)*

These patterns are confirmed by the results of the linear mixed model. The presentation of the results follows the same structure as the one in the predictions sections.

Individual variables

Acoustic model: Laterals identified by the ASR system as being more French-like (**L**) have significantly higher F2-F1 values (i.e clearer /l/s) that those detected as more English-like (**l**) (Est. 253.36, t-value $\sim$ 8.591, p-value < 0.001).

Syllable position: Coda /l/ is overall darker that onset /l/ (Est. -272.31, t-value $\sim$ -2.700 , p-value < 0.01).

Biological gender: As expected male speakers exhibit overall lower formant values (Est. -167.81, t.value $\sim$ -3.822 p-value < 0.001).

Vowel context: Vowel position has a significant effect on /l/ darkness with laterals in a back vowel context being darker (lower F2-F1) than in front vowel contexts (Est.156.2295, t-value $\sim$ 6.183, p-value < 0.001). (Figure 4)

Global accent: WER score, the only continuous variable in our model, does not have a significant effect on /l/ darkness.

Interaction terms:

- **Acoustic model Syllable position**: The interaction between the chosen acoustic model and the syllable position is significant: lateral segments detected as more English-like exhibit a higher difference in F2-F1 values in coda than in onset position (Est. 153.09, t-value $\sim$ 4.109, p-value $<$0.001).
- **Acoustic model Global accent**: No significant interaction was found (p-value $\sim$ 0.346).

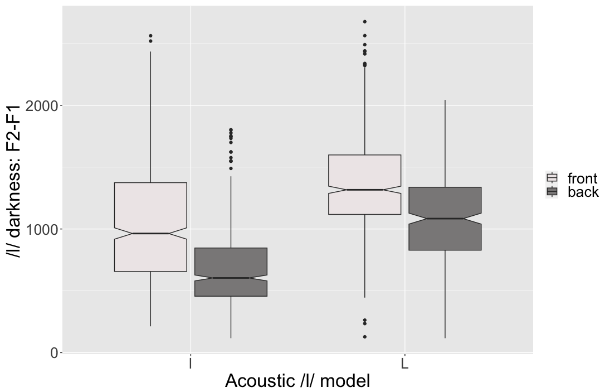

Figure 4: */l/ darkness (F2-F1) as a function of /l/ acoustic model (English (l) vs. French (L) and vowel context (front vs. mid vs. coda)*

In summary, all of our predictions, except for the one related to *global accent* were confirmed. Results show that the ASR's choice of the /l/ acoustic model correlates with the acoustic measures of /l/ darkness. Overall the French /l/ acoustic model is chosen in both onset and coda syllable position for /l/ varieties exhibiting a formant structure corresponding to clearer /l/s (i.e. higher F2-F1 measures).

4. Discussion

The present paper investigates L2 English French learners' production of onset vs. coda lateral allophonic variants combining acoustic measures and ASR with pronunciation variants (French vs. English /l/ acoustic models). Results show that French learners distinguish onset vs. coda /l/s by producing darker laterals in coda position. However not all learners produce English native-like laterals: 46% of lateral productions are detected as being closer to the participants native language (i.e. more French-like /l/). This is in line with previous studies that show that while learners move away from their L1 production, they differ from native speakers [15, 17]. Results also show an effect of vowel context on the formant structure of the lateral. This suggest there is a significant degree of coarticulation. Effects of coarticulation on lateral consonants are also found for native speakers: [23] coin the term "coarticulatory resistance" stating that dark /l/ is more resistant to coarticulation than clear /l/. No effect of global accent, defined based on WER scores, was found. This is not surprising since global accent does not necessarily represent the allophonic variation of interest here. A WER measure targeting /l/ tokens specifically could yield different results.

4.1. Limitations

The present paper only looks at static acoustic measures (formant structure at the midpoint of the lateral). For a better interpretation of how the ASR system chooses between English and French /l/ acoustic models more fine-grained acoustic measures are needed, such as dynamic measures (formant trajectories) and MFC coefficients. A dynamic analysis of the formant analysis would also shed more light on coarticulatory patterns. The current acoustic lateral model for English does not differentiate between clear and dark /l/ combining acoustic features from both positional variants. To specifically evaluate the production of dark /l/ one could add a third acoustic model corresponding to a language that only has dark /l/, and exhibits similar formant structure as the English dark /l/. Possible candidates would be Dutch or Portuguese, both having similarly dark /l/s to English [11]. Finally, the study would benefit from expanding the current analysis to include data from English native speakers, thus comparing native and non-native productions. All presented limitations are currently being addressed.

5. Conclusion

The current study sought to apply a, to our knowledge, novel approach, derived from ASR technology, to evaluate the pronunciation of syllable position allophonic variation of English L2 learners. Using both L1 and target L2 lateral acoustic models in parallel allowed the ASR system to detect pronunciations that diverge from native-like productions. Correlations of the ASR system output with traditional formant measures confirmed that elements detected as less native-like correspond to more L1-like formant structures (i.e. higher measures of F2-F1, a staple indicator of /l/ darkness). These results suggest that ASR with parallel L1 and L2 acoustic models , a less time consuming approach than classic acoustic measurements or native raters of accents in L2, can be used to detect L1 transfer in L2 production. Correctly identifying the mispronounced allophones, by providing immediate and automatic feedback, is a first step in bettering the pronunciation of L2 learners, who can focus on targeted pronunciations.

6. References

[1] S. N. Dart, "Articulatory and acoustic properties of apical and laminal articulations," *UCLA Working Papers in Phonetics*, vol. 79, 1991.

[2] A. Guskaroska, "Asr-dictation on smartphones for vowel pronunciation practice," *Journal of Contemporary Philology*, vol. 2, no. 2, pp. 45–61, 2020.

[3] S. Inceoglu, H. Lim, and W. H. Chen, "Asr for efl pronunciation practice: segmental development and learners' beliefs," *J. Asia TEFL*, vol. 17, no. 3, p. 824–840, 2020.

[4] H. Strik, K. Truong, F. Wet, and C. Cucchiarini, "Comparing different approaches for automatic pronunciationerror detection," *Speech Communication*, vol. 51, p. 845–852, 2009.

[5] C. Cucchiarini, A. Neri, and H. Strik, "Oral proficiency training in dutch l2: the contribution of asr-based corrective feedback," *Speech Communication*, vol. 51, p. 853–863, 2009.

[6] R. Sproat and O. Fujumura, "Allophonic variation in english /l/ and its implications for phonetic implementation," *Journal of Phonetics*, vol. 21, pp. 291–311, 1993.

[7] W. J. Hardcastle and W. Barry, "Towar articulatory-acoustic models for liquid approximants based on mri and epg data. part i. the laterals," *Journal of the International Phonetics Association*, vol. 15, pp. 3–17, 1989.

[8] C. Browman and L. Goldstein, “Gestural syllable position effects in american english,” *Producing Speech: Contemporary Issues*, vol. 15, pp. 3–17, 1989.

[9] S. Narayanan, A. Alwan, and K. Haker, “Towards articulatory-acoustic models for liquid approximants based on mri and epg data. part i. the laterals,” *Journal of the Acoustical Society of America*, vol. 101, pp. 1064–1077, 1997.

[10] M. Proctor, R. Walker, C. Smith, T. Szalay, L. Goldstein, and S. Narayanan, “Articulatory characterization of english liquid-final rimes,” *Journal of Phonetics*, vol. 77, 2019.

[11] D. Recasens, “Articulatory and acoustic properties of apical and laminal articulations,” *Speech Communication*, vol. 54, no. 3, pp. 368–383, 2012.

[12] G. Fant, *Acoustic Theory of Speech Production*. The Hague: Mouton, 1960.

[13] H. King and E. Ferragne, “The effect of ultrasound and video feedback on the production and perception of english liquids by french learners,” in *Phonetics and Phonology in Europe PaPE2017*, Köln, Germany, 2017.

[14] ——, “The dark side of the tongue: The feasibility of ultrasound imaging in the acquisition of english dark /l/ in french learners,” in *Ultrafest VII*, Hong Kong, 2015.

[15] L. Colantoni, A. Kochetov, and J. Steele, “Articulatory insights into the l2 acquisition of english /l/ allophony,” *Language and Speech*, pp. 1 – 33, 2023.

[16] J. Barlow, “Age of acquisition and allophony in spanish-english bilinguals,” *Frontiers in Psychology*, vol. 5, p. Article 288, 2014.

[17] T. Nagamine, “Acquisition of allophonic variation in second language speech: An acoustic and articulatory study of english laterals by japanese speakers,” in *Proceedings of Interspeech 2022*, 2022, pp. 644–648.

[18] S. Kobylyanskaya, “Speech and eye tracking features for l2 acquisition: A multimodal experiment,” in *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners’ and Doctoral Consortium*, M. M. Rodrigo, N. Matsuda, A. I. Cristea, and V. Dimitrova, Eds. Cham: Springer International Publishing, 2022, pp. 47–52.

[19] P. Boersma and D. Weenink, “Praat: doing phonetic by computer,” *From http://www.praat.org/*, 2022.

[20] H. Hermansky, “Perceptual linear prediction (plp) analysis for speech,” *J. Acoust. Soc. Amer.*, vol. 87, 1990.

[21] T. Fraga-Silva, J.-L. Gauvain, and L. Lamel, “Lattice-based unsupervised acoustic model training,” in *In ICASSP’11, 36th International Conference on Acoustics, Speech and Signal Processing*, Prague, Czech Republic, 2011.

[22] D. Bates, M. Mächler, B. Bolker, and S. Walker, “Fitting linear mixed-effects models using lme4,” *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[23] R. Bladon and A. Al-Bamerni, “Coarticulation resistance in english /l/,” *Journal of Phonetics*, vol. 4, no. 2, pp. 137–150, 1976.