# Investigating lexical stress accuracy in non-native speech through real-time speech visualization: a pilot study

Kizzi Edensor Costille

# Investigating lexical stress accuracy in non-native speech through real-time speech visualization: a pilot study

*Kizzi Edensor Costille*

CRISCO, University of Caen Normandy
kizzi.edensor-costille@unicaen.fr

## Abstract

This pilot study explores the impact of visualizing real-time speech on improving lexical stress among non-native English speakers. The study introduces a real-time 3D spectrogram, where learners can see, hear and imitate a model's speech and view their own productions. Six French English learners participated in a three-phase within-subject study consisting in a pre-test, a 10-week training session using the spectrogram, and a post-test. The study questions whether visualizing speech improves lexical stress production and equips learners to handle new words post-training. An auditory analysis of pre-test and post-test results revealed a slight improvement in correct lexical stress placement, with the global mean of accurately pronounced words rising from 4 in the pre-test to 4.5 participants in the post-test. The mean for correctly pronouncing pre-test words included in the post-test, improved by 1 (from 4 to 5) but there was minimal improvement in correctly pronouncing the new words in the post test. The goal of this study is to contribute to understanding how L2 learners can improve their word stress accuracy in English and to expand our knowledge regarding multi-sensory tools' efficacy in second language learning.

**Index Terms**: prosody, word stress, L2 learners of English, visualising speech, multi-sensorial tool

## 1. Introduction

Research on second language acquisition has soared in the last decades but few studies focus solely on prosodic aspects in non-native speech. Prosody's crucial role in non-native discourse intelligibility and comprehensibility has now been established in prior research [1], and [2]. However, mastering prosodic features like lexical stress remains a challenge for language learners [3], and [4]. The perception of lexical stress by French L2 leaners has been found to be difficult because of the vast difference between English and French prosody especially in the domain of stress, to the extent that the existence of 'stress deafness' has been postulated [3]. One of the proposed explanations is that lexical stress does not exist in French, therefore making it difficult to perceive.

In French, final syllables are lengthened, whereas in English, unstressed syllables are frequently reduced to a schwa representing a real challenge for native French speakers [5]. Lexical stress misplacement alone can lead to comprehension difficulties in addition to causing segmental mispronunciations. The combination of these can create real intelligibility and comprehensibility issues. A case in point is *independent* pronounced /in'dipident/ (instead of /ˌɪndəˈpendənt/).

Since the 1970s, there has been an on-going stream of studies which have used computer-based methods to test and improve the perception and production of prosody [6], and [7]. One of the first studies carried out in the early 1980s by [7] concluded that visual feedback was more effective than auditory feedback. Major advances in speech technology have led to the increasing use of language software and technology such as computer assisted language learning (CALL) and computer assisted pronunciation training (CAPT). A few examples of more recent tools are WASP [8] and Better Accent Tutor [9]. These examples are particularly interesting regarding L2 prosody because they enable the speaker to see their productions. Pitch visualisers (such as Praat [10], and similar software) have been used in more recent research [11], and [12]. While some studies have shown the potential of visual aids in improving speech production [12], [13], [11], and [14], others found that combining sound and image led to more mixed results in learning prosody, often due to the complexity of the software used [15], but also because of the cognitive load [16]. In a previous study on intonation [17], which used the multi-sensorial tool described below, four groups had access to different types of input. For example, one group had no input and just read and recorded sentences, the second group heard the model before recording their productions, group three only saw the spectrogram and group four received multi-sensorial input (heard the model and saw the spectrogram). The results between those who only had auditory input and those who had visual and auditory input were almost identical (63,5% audio vs 64,5% visual and audio input).

Several reasons were put forward to explain this result, notably the French stress deafness hypothesis [3], and the high cognitive load of dealing with multi-sensorial input.

A real-time 3D spectrogram tool, Englishville [18], was used in this study. It uses a real-time 3D spectrogram as a base and allows the capture of the audio stream so that it can be recorded on a server. This tool is then integrated in a website that enables corpus recording, setting up and participation in experiments. L2 learners can see (and hear) the spectrogram of the recorded utterances, repeat them and compare their own productions with the model. This can be considered as visual feedback [16], as learners are able to use the visual model to improve their own productions.

This article examines the preliminary results from a pilot study where non-native learners of English used Englishville to practice lexical stress pronunciation. The results presented here are part of a larger project researching the impact of this software on word-stress and intonation pronunciation. The research was carried out using a within-subject design [19], (i.e., the participants are compared to each other, and data comparison takes place within the group of participants, with each participant serving as their own baseline) to examine the learning effect [19]. During the training period (in the form of drills), learners saw and heard a model spectrogram and audio, recorded their own productions whilst visualising their own spectrogram. They could then listen back to both and record their production again if necessary. Previous research has

shown that using visual aids often yields better results in training sessions when learning prosody. This can even have lasting effects on speech production in general [20]. Therefore, it was hypothesized that the training done during the drills would have a lasting impact on the post-test experiment.

The research comprised three phases. Initially, a pre-test required participants to record a list of 30 words without visual or auditory aid. The words were taken from a list of 76 words which had been identified as problematic for French L2 learners of English. This list was made up from 56 words (collected from personal teaching experience), and 20 words from a published list [21, pp.111-119]. The words were then sorted out per number of syllables (2, 3 and 4 syllables).10 words of 2, 3 and 4-syllables were selected for the pre-test (30 words in total). Following the same criteria, 30 words were selected for the post-test to which the 30 words from the pre-test were added (60 words in total). Some examples are *village, separately, Japan, Britain, effort, independent, harmonious*. Certain stress-neutral suffixes were added to some of the words to assess how they would be produced. Subsequently, the participants were asked to complete a 10-week training session which consisted in weekly drills on words and sentences with varied lexical stress patterns. 120 words and 40 sentences were included in the drills. There were 40 2, 3 and 4-syllable words. The words from the pre-test (henceforth considered as 'known' words) were included in the drills but the 30 'unknown' words that the participants recorded later in the post-test were not included in the drills. The model speaker was a British female for all the recordings. The study concluded with a post-test of 60 words, including 30 from the pre-test. While learners had visual and auditory support during training, both pre-test and post-test were conducted without aids.

This study focuses on the following research questions: Does visualising speech enhance learners' word stress production? Through their training, are the learners better equipped to deal with the new words included in the post test?

## 2. Experiment

### 2.1. Corpus

Each participant recorded a total of 90 words (30 in the pre-test and 60 in the post-test). They were recorded in repetitive mode by six French learners of English.

The explanation the participants received before starting the drills was the following: "The text to be read appears on the screen and you will also hear it. You will also see the corresponding 3D spectrogram that appears in real-time. In the spectrogram, the colours red/orange correspond to high intensity indicating the stressed syllables of the word (and sentence). The green/yellow colours indicate low intensity corresponding to unstressed syllables. You can also see the movement of the tone of voice in the spectrogram, in other words, this means that you can see if your voice is moving in the same direction as the model, if the tone/pitch of your voice is going down or up. Try and imitate the model as much as possible, i.e., you should try to have red parts and green parts in the same places as the model. This corresponds to full/strong sounds (red) and weak/reduced sounds (green)".

### 2.2. Speakers

The pilot study involved six French English learners at a B2 proficiency level, corresponding to the French Baccalaureate level. They were enrolled in their first year of a BA in English at university and were all volunteers.

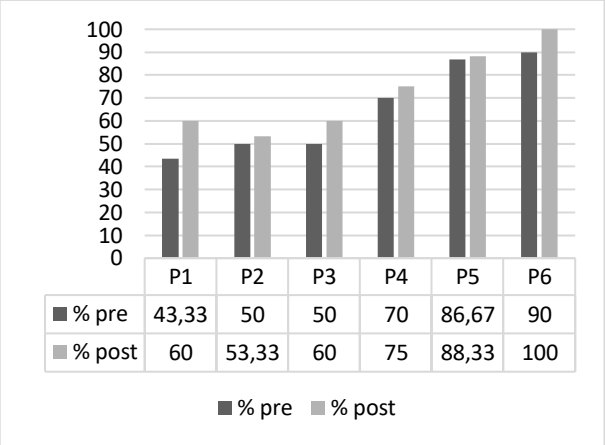
### 2.3. Auditory Analysis of the pre-test and post-test

An auditory analysis of the pre-test and post-test was carried out. For each word produced by the six participants the productions were compared to the model following certain criteria: the number of syllables realised, if the lexical stress was heard as being on the correct syllable and if it wasn't on which syllable it had been placed.

## 3. Results

### 3.1. Global results pre-test and post-test

As can be seen in figure 1, the results per participant are heterogeneous.




| | P1 | P2 | P3 | P4 | P5 | P6 |
|---|---|---|---|---|---|---|
| % pre | 43,33 | 50 | 50 | 70 | 86,67 | 90 |
| % post | 60 | 53,33 | 60 | 75 | 88,33 | 100 |

Figure 1: Correct stress (%) per participant for the pre-test and post-test (all words).

However, all participants showed improvement in the number of correctly stressed words in the post-test. This increase was also observed in the global average of correctly stressed words which was 65% in the pre-test and rose to 72.78% in the post-test. On average the participants progressed by 7.78%. The global mean of accurately pronounced words regarding lexical stress rose from 4 participants (out of 6) in the pre-test to 4.5 participants in the post-test.

As mentioned previously, the participants' results were heterogeneous. For example, in the pre-test, Participant 1 (P1) correctly pronounced less than 50% of the words, whereas P6 realised 90% correctly. The participants all improved from one test to the next, but their progression was also heterogeneous. P1, who got the lowest score in the pre-test improved the most (16.67%). P6 – the participant whose lexical word stress accuracy was the highest – also improved by 10%.

### 3.2. Comparing results of words from the pre-test and post-test separately

The inclusion of the 30 (known) words from the pre-test in the post-test served multiple purposes: to assess potential improvement in stress placement from the pre-test, and to compare the production of these known words with the unknown words in the post-test.

The mean for correctly pronouncing the known words in the post-test improved by 1 (from 4 participants in the pre-test to 5 in the post-test). The mean for the unknown words in the post-test was, however, identical to the global mean in the pre-test (4).

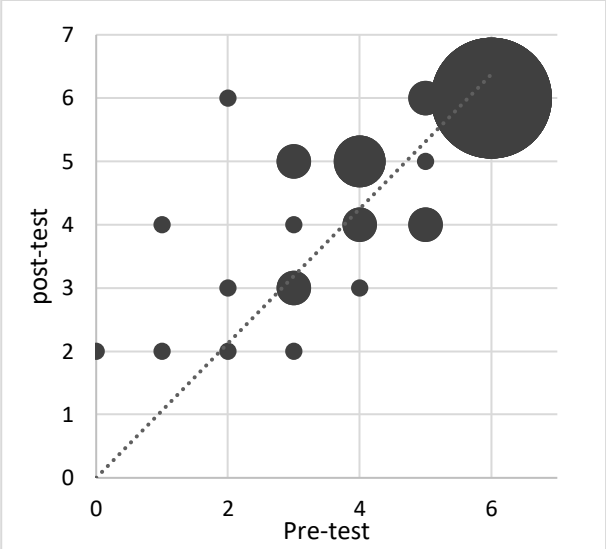



Figure 2: the number of participants with correct stress production for the same pre-test words in the pre-test and post-test.

The graph shows the relationship between the number of participants who correctly pronounced the lexical stress of the words that appeared in both pre-test and post-test. Each circle represents one or more words. The larger the circle, the more words it represents. The circles on the straight line indicate the words where there was no improvement in the pronunciation of lexical stress between the pre-test and the post-test. The further each circle is above this line, the greater the progression, as can be observed with the circle on the top left side of figure 2. This circle corresponds to the word *character* which was correctly pronounced by 2 participants in the pre-test but by all 6 participants in the post-test. The words which were the most mispronounced in the pre-test were generally pronounced with greater accuracy in the post-test. Underneath the line, we can observe the words for which the participants did not (or barely) improved from one test to the next. Overall, correct pronunciation of stress increased from an average of 65% in the pre-test to 75.89% in the post-test (known words only). The global average of correctly pronounced words was 72.78% in the post-test (all words) which decreased to 70.56% for the unknown words alone.

### 3.3. Stress-neutral suffixes production

It has been noted that in general, the lexical stress of the known words was more accurately pronounced than the unknown words in the post-test.

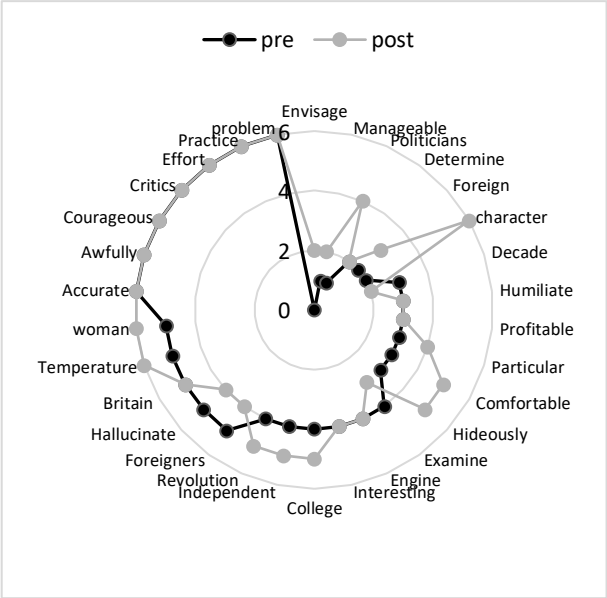



Figure 3: the number of correct lexical stress productions per word for the same pre-test words in the pre-test and post-test.

Figure 3 represents the number of accurate lexical stress production per word in both tests. On the left side of figure 3, it can be observed that words such as *problem*, *critics* were correctly produced by all participants in both tests. In the post-test (light grey line) the improvement on certain words can be seen clearly, e.g., *character*. In the pre-test, the word *foreign* was correctly stressed by 2 participants and *foreigners* was correctly stressed by 5 participants. However, despite being in the drills, the correct pronunciation of *foreign* (which had now become a known word) only improved by 1 participant in the post-test, and *foreigners* was accurately produced by only 4 participants. Both words had been included in the drills and therefore correct pronunciation was expected to increase. In the post-test, the new (but derived) word *foreignness* was less accurately stressed. Another stress neutral suffix was added to the word *effort* in the post-test (*effortlessness*). All participants accurately pronounced the stress on *effort* both in the pre-test and post-test but only 2 participants correctly pronounced *effortlessness* in the post-test. This shows that despite training, French learners of English were unable to correctly stress stress-neutral suffixes. It is possible that they did not know that these suffixes had no effect on word stress and did not notice or manage to learn this during the drills.

## 4. Discussion

The main objective of this pilot study was to investigate the effect of real-time speech visualization on lexical stress production in non-native speech. The global results show that there was a small improvement in the post-test results in producing correct word stress.

The degree of the participants' heterogeneity was unexpected. They were all first-year students at university; therefore, it was presumed they would all have a similar level. Despite this, the global average of correctly stressed words increased from the pre-test (65%) to the post-test (72.78%) and all the participants progressed albeit to different degrees (7.78% on average). Three levels of learners can be observed in the pre-test: poor,

average and excellent. Those with an accuracy score of 50% or less (P1, P2 and P3) can be classed as having a poor level, next, P4 whose accuracy score was 70%, can be classed as average, followed by two excellent learners (P5 and P6) who produced the most accurate lexical stress (86.67 and 90% respectively). However, if we consider the fact that the words in the pre and post-test were some of the most difficult for French learners of English in terms of lexical stress and were also cognates, it is possible to consider that those who scored 50% are better than just 'poor' speakers. The amount of time each participant spent practising the drills was not recorded therefore the amount of time spent on the drills cannot be correlated with each participants' progression. It is conceivable that those who spent more time practising simply progressed more than those who spent less time. The participant who got the lowest score in the pre-test (P1) improved the most (16.67%) and P6, who could already be considered as excellent in the pre-test (90% of word stress accuracy), did even better in the post-test (a 10% increase). All participants were able to improve their lexical stress productions demonstrating that real-time speech visualization can enhance both low level learners (poor) and very proficient speakers' productions.

As can be seen in Figure 2, the words which were the most mispronounced in the pre-test were often pronounced with greater accuracy in the post-test. It is unclear why the pronunciation of certain words, despite being in the drills, were not pronounced accurately in the post-test. The words chosen for this test were ones that are frequently mispronounced by French learners of English, even at an advanced level, maybe some of their pronunciations are simply fossilized and they were unable to correct them, even with training. Certain words were cognates, which could cause the interference from French to be even stronger than for words that were not. Another possible explanation lies in the hypothesis that French speakers are in fact stress deaf [3]. However, the participants in this study did progress to a degree, which could either signify that they were capable of perceiving and producing lexical stress (and therefore are not stress deaf), or that they were able to use the spectrogram to determine where the word was stressed. Overall, known words in comparison with unknown words were pronounced better - known words increased from an average of 65% of correctly stressed words in the pre-test to 75.89% in the post-test. The global average of correctly pronounced words was 72.78% in the post-test (for all words) but this decreased to 70.56% when the unknown words alone were analysed. Therefore, we can say that for unknown words there was a 5-percentage point improvement from the pre-test to the post-test.

Regarding stress neutral suffixes, it was expected that practising and producing these words in the pre-test and in the drills would help learners to correctly produce lexical stress when a stress neutral suffix was added in the post-test. This was not the case for *foreignness* or for *effortlessness* even though the words *foreign*, *foreigners* and *effort* were in the pre-test. This implies that the addition of a stress-neutral suffix to a word results in its treatment as an unknown word.

## 5. Conclusions

In this study, we sought to determine if learners can improve their lexical stress accuracy through real-time speech visualization. Consistent with previous research investigating the effectiveness of visual aids in the acquisition of prosody [12], [13], [11], and [14], a small progression in the accuracy of lexical stress productions was observed.

Regarding the second research question which was to investigate whether the drills could have a positive effect on learners, enabling them to accurately produce the new words included in the post test, the results were mixed. Unknown words were pronounced less accurately (70.56%) than the known words (75.89%), but even the unknown words were slightly better produced than the words in the initial pre-test (65%) (which were also unknown before the participants took the pre-test). Therefore, contrary to previous findings [19], it is unclear if the training had a lasting effect in the post-test.

Despite the limited number of participants, the results remain encouraging. The results of this study tend to confirm other researcher's claims [22], that computer technology can help second language learners more accurately perceive and produce prosodic features. Building upon the results of this pilot study, a larger-scale experiment is underway to test these findings on which it will be possible to run a paired samples t-test. With a greater sample size, the outcomes will be less susceptible to the influence of diverse participant characteristics. If the results are confirmed, it would challenge the assumption that French natives are deaf when perceiving and producing lexical word stress, while confirming the positive effect of visualizing speech.

## 6. References


[1] Munro, M. J., & Derwing, T. M. 1995. Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. Language Learning, 45(2), 73–97.

[2] Munro, M. J., & Derwing, T. M. 1998. Evidence in favor of a broad framework for pronunciation instruction. Language Learning, 48(2), 393–410.

[3] Dupoux, E., Sebastián-Gallés, N, Navarrete, E., Peperkamp, S. 2008. Persistent stress 'deafness': the case of French learners of Spanish. Cognition 106, 682-706.

[4] Horgues, C. 2008. French Learners of L2 English: Intonation Boundaries and the Marking of Lexical Stress. Research in Language 11 (1).

[5] Dan Frost. The Perception of Word Stress in English and French: Which cues for native English and French speakers? 2009, EPIP1 (English Pronunciation: Issues and Practices, Université de Savoie, Chambéry, France. pp.57-73.

[6] James, E. (1976). The acquisition of prosodic features of speech using a speech visualizer. International Review of Applied Linguistics in Language Teaching (IRAL), 14(3), 227–243.

[7] de Bot, K. (1983). Visual feedback of intonation: Effectiveness and induced practice behavior. Language and Speech, 26, 331–350.

[8] WASP https://www.speechandhearing.net/laboratory/wasp/

[9] Komissarchik, E., Komissarchik J. 2000. Better Accent Tutor – Analysis and Visualization of Speech Prosody. Proceedings of InSTILL, 86–89.

[10] Boersma, P. (2001). PRAAT, a system for doing phonetics by computer. Glot International, 5(9/10), 341–345.

[11] Olson, D. J. 2014. Phonetics and technology in the classroom: A practical approach to using speech analysis software in second language pronunciation instruction. Hispania, 97(1), 47–68.

[12] Imber, B., Maynard, C., & Parker, M. 2017. Using Praat to increase intelligibility through visual feedback. In M. O'Brien & J. Levis (Eds.), Proceedings of the 8th Pronunciation in Second Language Learning and Teaching Conference (pp. 195–213). Iowa State University.

[13] Kartushina, N., Hervais-Adelman, A., Frauenfelder, U. H., & Golestani, N. 2015. The effect of phonetic production training with visual feedback on the perception and production of foreign

speech sounds. The Journal of the Acoustical Society of America, 138(2), 817–32.
[14] Gorjian, B., Hayati, A., & Pourkhoni, P. 2013. Using Praat software in teaching prosodic features to EFL learners. Procedia - Social and Behavioral Sciences, 84, 34–40.
[15] Setter, J., Stojanovik, V., & Martínez-Castilla, P. 2010. Evaluating the intonation of non-native speakers of English using a computerized test battery. International Journal of Applied Linguistics, 20(3): 368–385.
[16] Olson, D. J. 2022. Visual feedback and relative vowel duration in L2 pronunciation: the curious case of stressed and unstressed vowels. In J. Levis & A. Guskaroska (eds.), Proceedings of the 12th Pronunciation in Second Language Learning and Teaching Conference, held June 2021 virtually at Brock University, St. Catharines, ON.
[17] Edensor Costille, K. 2023. Englishville: A new way of practising prosody. In A. Henderson & A. Kirkova-Naskova (Eds.), Proceedings of the 7th International Conference on English Pronunciation: Issues and Practices (pp. 62-69). Université Grenoble-Alpes.
[18] Edensor-Costille, K. 2020, May 6 Englishville. https://demo.englishville.ovh/
[19] Seltman, H. J. (2012). Experimental Design and Analysis. Pittsburgh: Carnegie Mellon University.
[20] Derwing, T. M., & Rossiter, M. J. 2003. The effects of pronunciation instruction on the accuracy, fluency, and complexity of L2 accented speech. Applied Language Learning, 13, 1–17.
[21] Chabert, E. 2018. Bien prononcer l'anglais – Manuel d'anglais oral pour les francophones. Génération5.
[22] Hardison, D. M. 2004. Generalization of computer-assisted prosody training: Quantitative and qualitative findings. Language Learning and Technology, 8, 34–52.