Parallèles 36(1) Avril 2024

Converting semiotic signs into a linguistic code: Implications for language learners' oral skills

Marga Navarrete

University College London

Abstract

Research on the use of audio description (AD) in foreign language education has only been developed over the last decade. This paper introduces an experimental study on the potential of active audio description in learners' oral skills, focusing on the quantitative data which evidenced that oral productive skills are enhanced by this AVT mode. The study involved 81 undergraduate students of Spanish enrolled in a British University. During the ten-week intervention, participants were required to complete collaborative AD tasks. A range of instruments were used for data collection, which include pre- and post-questionnaires, rubrics, pre- and post-tests based on recordings of spontaneous conversation and observation notes by the teacher-researcher. There were two observers who revised all data collection tools, and three external evaluators assessed the potential enhancement of oral productive skills in learners' pre-tests and post-tests. Such a wide choice of tools was essential to allow the triangulation of data, and thus, to guarantee a greater reliability and consistency of the results obtained. Intonation, speed, and stress demonstrated the most significant improvement following the intervention, while the reduction of prolonged pauses was minimal, with the lowest rating among all examined features.

Keywords

Didactic audio description (DAD), foreign language teaching, oral production skills, experiment, quantitative data

 DOI:10.17462/para.2024.01.06

1. Introduction

This article discusses an experiment that aimed at confirming the following hypothesis: students enjoy active audio description (AD) tasks, they find them beneficial not only for their language learning in general, but also for productive oral skills (fluency, pronunciation and intonation), which is enhanced in spontaneous speech thanks to the use of didactic AD practice (DAD). The study took place in the third cycle of a larger study from which relevant qualitative and quantitative data was gathered and analysed. As such, it followed a cyclical structure in accord with action research principles, a methodology based on evaluation and reflection to implement required changes. Thus, lessons learnt from each cycle were applied to improve the quality of the data collected in each subsequent stage. Although its focus evolved with objectives and research questions, it culminated with this main experiment, which successfully responded to both questions formulated during the course of the study. The methodology used for the first and second cycles mainly provided information about learners' perceptions on AD practice. However, it was not until the final cycle (cycle 3) when there was enough relevant and consistent data to respond the second question which queried about the potential enhancement on oral productive skills in spontaneous speech.

1. What are the learners' perceptions when completing DAD tasks?
2. How does DAD practice impact oral production skills (fluency, pronunciation and intonation) in spontaneous speech?

The main objectives of this study were to analyse the impact of DAD practice on oral skills in terms of fluency, pronunciation and intonation, and also, the learner perceptions after completing their tasks. A secondary objective was to outline a series of guidelines for DAD which were extracted once data was analysed. The goal was to contribute to the community of language practitioners interested in an effective use of this audiovisual translation (AVT) mode. Due to space restrictions, this article focuses on the second research question. This is because the qualitative data obtained with regards to learners' perceptions clearly confirmed the conclusions reached in previous experiments (Navarrete, 2020). Participants' reflections towards AD practice were positive and encouraging. They valued how oral productive skills, as well as grammar and vocabulary, were enhanced. They also found their course tasks enjoyable and fun, and they appreciated the way the course was taught. However, on the negative side, some learners were not able to see the link between the course and certain areas of assessment.

Firstly, this article introduces the theoretical background of DAD, followed by the methodological rationale and the design of this study. This includes the mixed-methods strategies employed for data collection and analysis of results obtained. Secondly, it discusses the context and participants of the experiment as well as the resources and procedures used for structuring the lessons in a coherent way. Finally, it analyses the quantitative results obtained which evidenced the enhancement of oral production skills in language learning.

2. Audio description as a young sibling of didactic audiovisual translation

AD is a mode of AVT generally used for making video content accessible to blind and visually-impaired audiences. It aims at facilitating access to visual content by translating it into a verbal narration (Walczak & Fryer, 2017). AD is included in the AVT category of revoicing which comprises all modes that involve oral narration, dubbing, voice-over and AD (Lertola, 2019). In foreign language education (FLE), this exercise is often referred to didactic AD (DAD) as it is an active practice where the language learner inserts a narration into the original soundtrack of a clip to describe information transmitted visually, thus converting images into words. As

noted by Salway (2007), the scholarly examination of the connections between images and words has been ongoing for centuries, and the practice of AD combines them by incorporating spoken language discourse. Fryer (2016, p. 3) highlighted that "[u]nlike subtitling, dubbing, or voice-over, AD does not involve an existing text that requires translation from one language to another". She cites Braun's (2008, p. 2) characterisation of this process as an "intersemiotic, intermodal, or cross-modal translation or mediation" alluding to Jakobson (1959), who originally coined the term "intersemiotic" to describe forms of translation where the information does not originate solely from the translated channel but draws from other sources. Fryer also clarifies that in this context, "modal" suggests additional layers of meaning, encompassing spoken expressions, written communications, music, or sound effects. This is perhaps one more reason why this form of AVT is particularly beneficial for language learners. As students transform a visual concept into verbal expression without the need for intervention through their native language, they must construct their own scripts from scratch and record them using their own voices. This mode is often portrayed as an artistic practice due to the creative nature of translating visual elements into a linguistic code whilst the learner becomes a social agent that mediates between the clip and others, using aural discourse to interpret what can be seen (Navarrete, 2020, 2022; Navarrete & Bolaños García-Escribano, 2022).

The first attempt for a pedagogical practice of AVT in the language setting using the subtitling mode was by Díaz-Cintas (1995, 1997) which was followed by Talaván's works (2006a, 2006b). That was the beginning of a long list of experimental studies that not only used subtitling, especially at the early stages, but also, revoicing in foreign language education (FLE). For further information, it is advisable to check both compendiums by Lertola (2019) and Talaván (2020) who examine in detail relevant studies carried out from then.

Most experimental studies on DAD have been released just over the last decade focusing on the impact of this mode of AVT on different language skills and competences, such as lexical competence (Ibáñez Moreno & Vermeulen, 2013); integrated skills (Ibáñez Moreno & Vermeulen, 2014; Ibáñez Moreno & Vermeulen, 2017); oral production skills (Ibáñez Moreno & Vermeulen, 2015, 2016, 2021; Navarrete, 2018; Talaván & Lertola, 2016); writing skills (Calduch & Talaván, 2018; Talaván *et al.,* 2022); morphology (Schaeffer-Lacroix, 2020); media literacy (Herrero & Escobar, 2018); and learners' perceptions (Bausells-Espín, 2022). In addition, some didactic proposals (Cenni & Izzo, 2016; Navarrete, 2018, 2020, 2023) and work on intercultural competence (Ibáñez Moreno & Vermeulen, 2017) have been published. Finally, the aforementioned works (Navarrete, 2020; Navarrete & Bolaños García-Escribano, 2022) have focused on providing a methodological framework for DAD that is aligned to the CEFR Companion Volume (Council of Europe, 2020).

Authors	Date	Features examined	Participants	Language	Level
Ibáñez Moreno and Vermeulen	2015	Speaking skills	16	English	B1
Ibáñez Moreno and Vermeulen	2016	Speaking skills	12	English	B1
Ibáñez Moreno and Vermeulen	2021	Speaking skills	28	English	B2
Talaván and Lertola	2016	Speaking skills	30	English	B1
Navarrete	2018	Speaking skills	6	Spanish	B1-B2

Table 1. Main studies on oral production skills (source: author)

As seen in Table 1, Ibáñez Moreno & Vermeulen's studies (2015, 2016) involved a restricted number of participants. They developed an application called "Videos for Speaking" (VISP) and deliberated on activities within it that employed DAD techniques to improve oral accuracy and fluency. While these studies did not primarily aim to measure quantitative improvement, their qualitative analysis of the outcomes provided insights into students' perceptions regarding the potential of DAD practice in enhancing oral skills. The same is true when they compared (2021) two different usages of the above application. The findings indicate that concerning language practice, VISP demonstrates equal effectiveness whether used as a classroom support tool or as a standalone application for independent use outside the classroom. However, in terms of attitudinal aspects, the students who integrated the app into their classroom activities displayed a more favourable attitude toward the app compared to those who employed it independently, with the latter group showing less enthusiasm and motivation regarding the app's utility and advantages.

Talaván & Lertola's quasi-experimental study (2016) was undertaken with 30 participants, divided into two groups, the experimental and the control group. The data analysed in the oral production test demonstrated the enhancement of this skill. A further data source comprising the feedback provided by the students confirmed these findings as the results were triangulated comparing both sets of data. However, the authors claimed that these findings are context-specific and should not be broadly applied. Despite a small participant pool, the combination of data sources and various data collection methods yielded a fairly robust design, increasing the likelihood of successful replication and broader applicability.

In a small-scale experimental research effort, Navarrete's study (2018) explored how DAD enhanced the oral skills of six undergraduate students who were studying Spanish as a foreign language at the B1 level. The investigation comprised an initial oral assessment, an evaluation of AD tasks, and the distribution of two questionnaires. Students recorded podcasts and completed AD assignments, followed by peer discussions in the classroom. Both tasks were assessed using the same criteria, which encompassed pronunciation, fluency, vocabulary, and grammar. The research closely examined improvements in fluency, pronunciation, and intonation. While the study recognised its limitations, the author found the positive student reactions to DAD practice promising and indicative of potential for further investigation.

3. Sample and context

In the upcoming sections, we will examine the study that is the focus of this article. Its specifications will be drawn before delving into the discussion of the results. The experiment took place from January to April 2017. Initially, 81 students participated in the experiment, 64 completed the pre-questionnaire, 37 the post-questionnaire and 46 students were recorded for both the pre-task and post-task. All students were studying their first year at University College London (UCL), and they were 18–19 years old. They would have done an A-level in Spanish in the previous year or an equivalent course if they had studied secondary education in a foreign country. They all joined the course with ranging levels of proficiency around B1 (CEFR, 2020). The participants were students from different departments (School of European Studies Languages and Culture, Economics and Business, BA in Modern Languages, etc.). The Spanish language module as a whole is divided into four components (grammar and communicative skills, oral skills, translation from and into English-Spanish), and it runs for two terms (10 weeks each). The experiment was carried out within the Translation from English into Spanish component. There were nine groups with an average of 9 students per group taught by the teacher-researcher. This took place over a 10-week period, consisting of a session per week, where students had to complete eight AVT collaborative tasks spread out throughout the course.

4. Data gathering tools

A number of instruments were used to collect relevant data. The methodology and the data collection instruments used in the previous cycles were essential for the design of the main experiment. These include pre- and post-questionnaires, rubrics, pre- and post-tests and observation notes by the teacher-researcher. There were additional participants in this cycle: two observers revised all data collection tools, and three external evaluators assessed the potential enhancement of oral productive skills in learners' pre- tests and post-tests. Such a wide choice of tools was essential to allow the triangulation of data, and thus, to guarantee a greater reliability and consistency of the results obtained. With 46 participants having completed the majority of AD tasks along with pre- and post-tests, it became feasible to measure the impact of AD practice on oral production skills (fluency, pronunciation, and intonation) in spontaneous speech.

In order to test and triangulate the results from the rubric applied to the assessment of oral skills, a similar procedure than the one used with the questionnaires was followed. Drawing from the researcher's teaching background, a detailed list of potential oral production features was meticulously compiled, focusing on common errors observed among Spanish speakers. In addition, rubrics from previous experiments (1 and 2), and examples of other rubrics used in AVT research, were revised, such as the ones Sánchez-Requena (2018) and Talaván and Lertola (2016) used for their studies. It had to be designed to be easily filled in by the evaluators. This time quantitative data, as well as qualitative data, was meant to be collected. Therefore, it needed to include variables that could be assessed numerically. Also, there was an open question to address any additional information that might not have appeared explicitly or might have been too specific to one particular participant. Again, only the strictly indispensable number of items were included. The rubric was also reviewed by the same observers that checked the questionnaires of this experiment. Table 2 represents a summary of the rubric employed to assess students' compulsory tasks, pre- and post-tests.

Pronunciation	**Intonation** **Naturalness**	**Fluency** **Hesitations** **Prolonged pauses**	**Fluency** **Continued speech** **Intelligibility** **Lexical knowledge** **Grammatical Knowledge** **Self-correction** **Pronunciation**
Vowels (e/o/u)	5 Excellent	4 Too many	5 Excellent
Dipthongs	4 Very good	3 Many	4 Very good
ae/ao/au/eo/eu	3 Good	2 Some	3 Good
ia/ie/io/iu/etc.	2 Average	1 Almost none	2 Average
Consonants	1 Low		1 Low
h/p/g/j/r/ñ			
b/v			
s/c			
t/d			
Orthographic confusion (que/qui, gue/gui)			

Table 2. Summary of rubric used for tasks, pre- and post-tests (source: author)

This rubric was used for both the AD tasks and the pre- and post-tests[1], since there was no reason to have different rubrics provided that the aim in both cases was to assess oral skills (fluency, pronunciation and intonation). Although in the case of the AD tasks, students had the chance to prepare their speeches, it should be noted that the objective of the study during this cycle was to find out if AD tasks had an impact on oral productive skills in spontaneous conversation. Hence, it was necessary to measure this impact on a number of oral features by comparing the tests before and after intervention.

The features analysed in this rubric were measured by a scale of 1 to 6 (1 = very poor, 2 = poor, 3 = adequate, 4 = good, 5 = very good, 6 = excellent). Scales of six instead of five values were selected following recommendations by Chyung *et al.* (2018). First, to avoid participants selecting middle options, which might cause the invalidation of certain questions. Second, to facilitate the statistical analysis of the variables. It is easier to separate the negative values from the positive ones by creating two isolated groups.

In the complete rubric, speed was assessed taking into account how fluent each participant was, and if speech was continuous or too slow with frequent pauses. Intonation was assessed in terms of how natural, that is near-native, the speech sounded. A general evaluation on pronunciation was also included, although appreciation of particular phonemes (consonants and vowels) were also incorporated in the following section of the rubric. The next feature, *easy to follow speech*, was related to the level of intelligibility of the participant's speech. These two last variables, *pronunciation* and *easy to follow speech,* attempted to assess overall pronunciation. For the second element, it is assumed that a listener will struggle to understand a speaker's speech for a number of reasons, such as lack of vocabulary and incorrect grammar (as mentioned earlier), but also due to unintelligible pronunciation. This design allowed the assessment of a number of features from a general point of view (pronunciation, intonation and stress), all of which affect the way speech is perceived by the evaluator. Finally, the last two variables of the rubric, *vocabulary* and *grammar,* were also selected to assess fluency, since deficiencies in vocabulary and grammar knowledge often cause dysfluency. This final feature was measured by assessing the level of wavering or prolonged pauses by the speaker: this was done by considering the number of pauses that the participant would make, since a broken speech is a clear sign of deficient fluency.

In terms of pronunciation, the teacher-researcher looked at vowels and groups of vowels where learners tend to struggle: the vowels /e/, /o/ and /u/ forming syllables, and articulation of diphthongs (/ai/, /ue/, /oi/, etc.). Regarding consonants, there was a group of consonants assessed individually (silent h, p, g, j, r, and ñ), which are the ones that tend to cause more problems to English speakers pronouncing Spanish words. In addition, the contrast between b/v, s/c and t/d was also assessed. In particular, special attention was devoted to the last pair (t/d), as there is a tendency by a large number of learners (in particular English speakers) to make these sounds alveolar instead of dental, because of their interference with English where these sounds tend to be the opposite. This general mispronunciation feature was explicitly explained, and samples taken from students ADs were assessed and worked on in class.

All these phonemes were selected based on the teacher-researcher extensive experience in FL teaching, in addition to the pronunciation studies by known scholars (Cala Carvajal, 1997; Ichaurralde, 2001; Mompeán-González, 2001). Finally, examining stress at word level and the mispronunciation of 'que/qui gue/gui' due to orthographic confusion were additional variables included in the rubrics for experiments 2 and 3. After the second experiment, the teacher-researcher noticed how some students, who had previously written their AD scripts,

[1] Both of them available online at: https://bit.ly/2SmUA1f and at: https://bit.ly/2GGhBec respectively.

did pronounce the 'u' in syllabic groups such as 'que/qui' and 'gue/gui'. This pronunciation error is common with level B1 students and it is due to orthographic confusion.-

5. Resources and procedures

For the audio description tasks, seven video clips were carefully selected. As stated by Garza (1994) and King (2002), this should be a rigorous task on the part of the teacher-researcher, clips need to be motivating and culturally relevant for the students (especially because they will be playing the same clip continuously in order to complete their tasks). Also, they need to be adapted to the language objectives of the course where they are integrated. The clips for the experiment were selected with the aim of covering a variety of topic areas, grammar structures, and semantic fields that were well-aligned to the course syllabus. All of them (see Table 3) lasted between 1 to 4 minutes approximately, following the recommendations by many authors (such as Garza, 1994; Rost, 2002; Tomalin, 1986; Stempleski, 1990) who agree on using short clips, preferably from 30 seconds to 3-4 minutes and with a maximum duration of 6 minutes. In order to facilitate video editing, none of the videos included dialogue scripts, only music soundtracks. Also, aspects such as the pace of the action, the mood of the characters and the setting were cautiously considered to reach the desired level of difficulty for each task. According to Burt (1999), the clarity of the message, rhythm, duration of the clip, and level of interest for the learners are key aspects to consider, all of which were also examined when selecting each clip.

Clips	Source	Duration
1 *The Hunger Games*	https://youtu.be/B8BD9txkGL4	0:56
2 *The Full Monty*	https://youtu.be/0MTuZoNvahI	1:07
3 *One Day on Earth…Cooking Tortilla de Patatas*	https://vimeo.com/56434999	1:21
4 *Maldición*	https://youtu.be/k-Z8xxygd2Y	1:45
5 *The Minions*	https://youtu.be/Q04KG7gVQtw	3:52
6 *Turismo Perú*	Clip no longer available online	2:23
7 *A short love story in stop motion.*	https://vimeo.com/877053	3:47

Table 3. Clips and their duration (main experiment, cycle 3) (source: author)

In the course of the experiment, tasks were scaffolded following Vygotsky's principles on learning: '[s]caffolding not only produces immediate results, but also instils the skills necessary for independent problem solving in the future' (Vygotsky, 1978, p. 57). Thus, tasks were organised with increasing level of difficulty and creativity involved to instigate both autonomous learning and creative problem-solving skills.

For recording purposes, all videos were divided into the number of members of each working team (usually three members per group). Thus, if a clip lasted three minutes, each student would record about one minute. The scripts were sent via Moodle, our university VLE, annotating participants' interventions. The aim of this request was twofold: first, to enable the teacher-researcher to provide individual feedback on fluency, pronunciation and intonation (when the tasks were compulsory), as summative assessment was involved; second, to allow her to analyse the data collected by learners' individual task performance.

As there were only ten sessions for the experiment to be run, it was decided not to include the technology teaching in class explicitly. However, some measures were considered to ensure that students were able to carry out the minimum audio or video editing requirements. Each team was created in such a way that at least one member was confident enough to use the

appropriate programs (students were asked about their computer skills in the first session to make this possible).

There were three types of tasks, for all of which students had to record ADs with their own voices. The first type (type I) included tasks where the English AD of a video clip had to be adapted into Spanish. This set served primarily as examples of best practice regarding AD creation. Students would deduce basic guidelines on AD and reflect on time constraints when carrying out their own. The second set of tasks (type II) involved the creation of an AD from scratch of a given video clip. Students had to write their own scripts by looking at the images on screen and by selecting relevant information. The final task (type III) went a few stages further: this time not only the AD was needed but also a short clip had to be created, so that students would become designers, producers, audio describers and actors, taking and supervising each other's roles.

All sessions apart from one (aiming at preparing students for the written exam) were devoted to the experiment. In each session, a new clip would be introduced, students would discuss potential difficulties, work with the ADs at home, submit their transcription and recordings to our university VLE, and upload their videos to the YouTube platform. Feedback would be provided in the following session where each new video was viewed and discussed in class. Table 4 shows a basic structure of each session. At the end of each one, there was a wrap up stage where students, with the help of the teacher, outlined the learning outcomes of each session and highlighted particular difficulties encountered.

Stages	Task	Description
1	Reflecting on videos watched with tasks completed	Assessing videos and clarifying queries Students watch the video clip they have worked with, and reflect on problems posed for the corresponding AD script
2	Watching the clip for next task for problem anticipation and procedure discussion	Familiarising with the clip and discussing potential problems they might pose
3	Wrapping up	Learning outcomes are summarised Expectations for next task are discussed

Table 4. Stages involved in each session (source: author)

As it can be seen in the above table, a similar structure was used for each lesson. The discussion in class gave the teacher-researcher the chance to take relevant observation notes. As Dörnyei (2007, p. 185) points out, '[t]he main merit of observational data is that it allows researchers to see directly what people do without having to rely on what they say they do.' Therefore, the observer in this experiment had the chance to recall aspects such as particular difficulties posed by the tasks, students who submitted or did not submit their tasks and the reasons provided, the way they sat in class, the relationship among students, the interaction among them, and general rapport and class dynamics. All of these notes were key to carry out the necessary amendments to subsequent stages of the experiment, such as time dedicated to the tasks, reinforcement of certain areas of the course syllabus, attendance in class, encouraging students who did not submit a task, understanding technical problems they faced and how to solve them, etc.

6. Quantitative analysis of results: Course task results and evaluators' rubric

As mentioned earlier, two tests were designed to examine the participants' improvement on oral productive skills. Furthermore, the teacher-researcher evaluated individually three of the

tasks that students had carried out during the course of the experiment to contrast results with the pre- and post-test results, which were also assessed by three additional evaluators.

Figure 1 shows the average scores of the most representative tasks following the features assessed by the teacher-researcher. The conditions in which this evaluation took place were rather different to those of the pre- and post-test for two main reasons. First, the number of evaluators, a single one (the teacher-researcher) assessed the participants' speeches whereas in the case of the main tests there were four evaluators who assessed the participants' speeches before and after the intervention by using an ad-hoc rubric. Second, the AD tasks allowed students to write and prepare their speeches in advance whereas in the case of the main tests, participants were assessed in spontaneous speech.

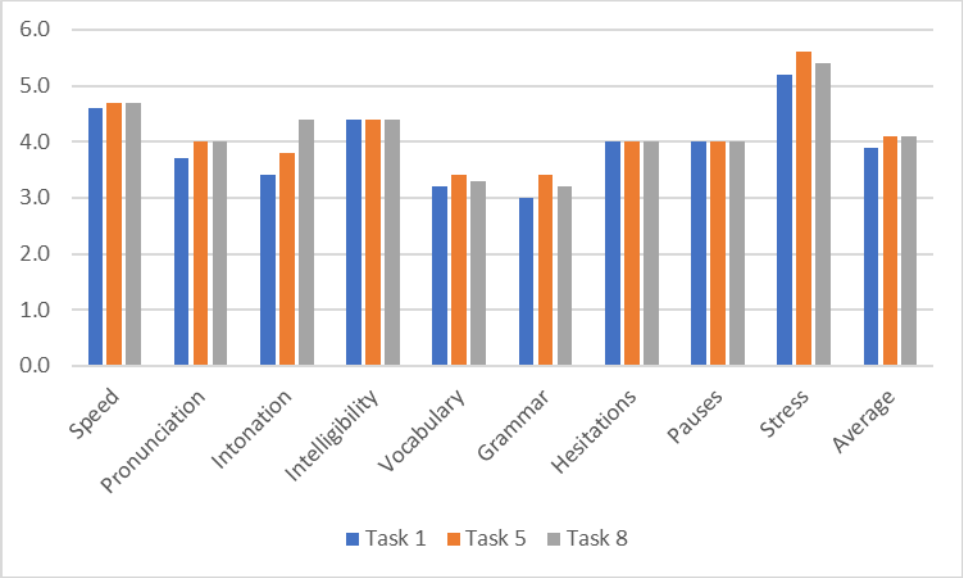

Figure 1. Average participants' scores across most representative tasks (experiment 3, cycle 3) (source: author)

As seen in Figure 1, the average improvement was not very significant, but one can appreciate a small increase from task 5. The features assessed for tasks 1, 5 and 8 respectively, in which an enhancement was most noticeable, were, in this order: intonation (3.4, 3.8 and 4.4), pronunciation (3.7, 4.0 and 4.0), stress (5.2, 5.6 and 5.4), grammar (3.0, 3.4 and 3.2), vocabulary (3.2, 3.4 and 3.3) and speed (4.6, 4.7 and 4.7). However, stress, grammar and vocabulary showed an increase in the rate of performance in task 5 rather than 8. These results could be explained by the fact that the final task was slightly different to all previous tasks, and learners had to create a video clip in addition to carrying out its corresponding AD. Thus, this task might have been more demanding from the participants' point of view, and their speech performances might have been slightly more careless.

The pre-test took place in the second week of the course, whereas the post-test was performed in the last one. Participants were recorded while responding to a series of questions that they had to elaborate themselves. Most features analysed showed relevant impact by the intervention, however less impact was found regarding specific sounds. All features were statistically analysed following the procedure below, as seen in Table 5.

Steps	Action	Objective
1	Convert nominal attributes into figures	To obtain quantitative data which allows pertinent calculations
2	Organise data in a coherent manner	To apply relevant tests
3	Plot the frequency of the pre-test ratings	To check if data sets have a normal distribution, so that either parametric or non-parametric tests can be applied
4	Apply T-Test (for parametric data)	To calculate the rate of the statistical significance per set of data
5	Select the features that have more increase in rating	To focus on the most relevant data obtained
6	Contrast the increase in rating across evaluators	To triangulate data

Table 5. Statistical procedure for data analysis (source: author)

In the data analysis process, Steps 1 and 2 involved preparing the data by converting nominal values into numerical ones, allowing for easier analysis. The third step included plotting the frequency of pre-test ratings to determine if the data had a normal distribution, which was confirmed. This enabled subsequent parametric tests.

Figure 2 illustrates the resulting analysis of the intonation attribute. If the pre-test had a normal distribution, one could infer that this was equally the case for the corresponding post-tests.

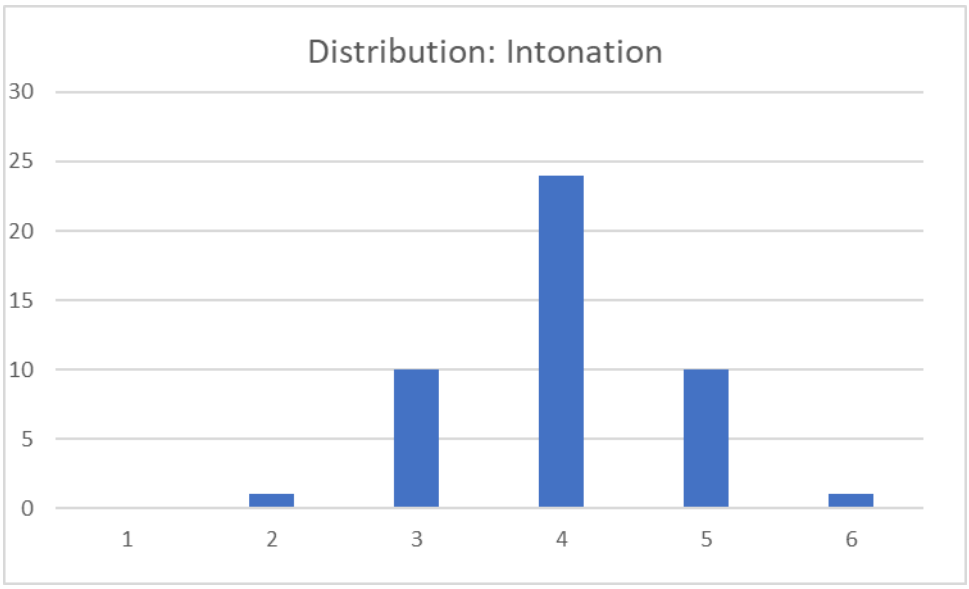

Figure 2. Intonation attribute showing a normal distribution (source: author)

It is important to note that although many attributes across the four evaluators were analysed for a normal distribution, not all of them showed such a perfect one as in Figure 2. However, the data analysed after pertinent calculations were performed showed enough evidence for a parametric test to be applied.

Having ensured that the data sets had a normal distribution, a parametric test (step 4) was applied. According to Dörnyei (2007), when comparing research designs that have two sets of scores obtained by the same group, or when the same participants are measured more than once (46 participants were the ones that carried out both the pre and post- tests), it is best to compute a paired-sample t-test (also known as 'matched t-test', 'matched-pairs t-test' and

'pairs t-test'). This test calculates whether the difference between two sets of scores reaches statistical significance.

Feature	Evaluator 1	Evaluator 2	Evaluator 3	Evaluator 4
Speed	0.00005%	0.05912%	0.00840%	0.00006%
Pronunciation	1.78928%	1.09207%	0.31826%	0.00000%
Intonation	0.00005%	1.91822%	0.00050%	0.00000%
Intelligibility	0.09559%	1.78928%	0.31826%	0.00009%
Vocabulary	0.02726%	20.94047%	0.05747%	0.00001%
Grammar	1.09207%	3.23111%	0.00021%	0.00002%
Hesitations	0.00002%	0.89548%	1.07893%	0.01641%
Pauses	0.00000%	42.02212%	42.02212%	0.00000%
Stress	0.00495%	0.18514%	0.00099%	0.00000%

Table 6. The two-tail (t-test) results across evaluators (source: author)

As seen in Table 6, the majority of findings show statistical significance, denoted by a threshold of $p<.05$ or 5% with a few exceptions that fail to meet this criterion. This is the case of some of the results given by evaluator 2 who gave a very low score to the *vocabulary* attribute, and so the average rate of increase is almost unnoticeable. The same is true when it comes to the *pauses* attribute; both evaluators 2 and 3 did not see much change in this attribute, so they both gave it a low score. These actions have translated into a lack of statistical significance on these specific areas. Thus, one could conclude that these two attributes discussed in relation to evaluators 2 or 3 did not allow to reject the Null Hypothesis for the reasons explained above.

Step 5 involved separating the data sets that exhibited a significant increase in ratings from those that did not. The aim was to analyse these data sets in a different way, focussing on the first set for a positive analysis of the results obtained that supported the successful impact of the intervention in learners' speech. The data sets that did not score a high increase in rating were examined from a remedial point of view in order to figure out possible causes that might justify this result. Thus, it would be necessary to decide whether the intervention had no impact or whether the design of the data collection tools and the tests themselves had not been appropriately carried out to assess certain features of oral production skills. This final step for data analysis aimed at contrasting the increase in rating among the four evaluators in order to triangulate the results obtained. Average rates of the different sets of data collected were calculated to improve the accuracy of the analysis of the results.

Table 7 depicts the average rate of change or increase given to each feature by each individual evaluator. These figures are the resulting calculations after comparing both sets of data, before and after the intervention. Average rates given to each feature across all evaluators are offered so that, by looking at this table, one can see the scores given to each feature individually and with all evaluators together as a group.

Feature	Evaluator 1	Evaluator 2	Evaluator 3	Evaluator 4	Average
Speed	0.4	0.3	0.4	0.6	0.43
Pronunciation	0.2	0.2	0.3	0.6	0.32
Intonation	0.4	0.2	0.4	0.8	0.45
Intelligibility	0.2	0.2	0.3	0.5	0.30
Vocabulary	0.3	0.1	0.5	0.5	0.34
Grammar	0.2	0.1	0.5	0.5	0.33
Hesitations	0.5	0.3	0.2	0.4	0.34
Pauses	0.0	0.0	0.0	0.6	0.16
Stress	0.3	0.2	0.5	0.7	0.42

Table 7. Average rate of increase in the features assessed across evaluators (source: author)

Table 7 lists the attributes that increased most and least depending on the evaluator. One should note that average figures that showed a greater increase appear in dark to lighter green; medium increase in darker to lighter brown, and the lowest increase is in red. It shows how intonation (0.45), speed (0.43) and stress (0.42) scored the highest rates, which means that they are the features in which the intervention had a maximum impact. It further validates that the frequency of prolonged pauses among participants, a negative aspect of learners' speech, did not decrease despite intervention, as indicated by its minimal increase in rating (0.16), which is the lowest among all the features analysed.

As depicted in Table 7, there is a notable discrepancy in the assessments provided by the four evaluators. Evaluators 1 and 3 reported similar results, while evaluator 2 reported lower scores and evaluator 4 reported higher scores compared to the former two. This disparity can be attributed to two combined hypotheses that address these shortcomings within the evaluation process, and it is crucial to consider the inter-rater reliability factor here.

Firstly, it is possible that the rubric used for assessment lacked sufficient information in terms of performance criteria. Some authors such as Cox *et al.* (2015) and Dawson (2015) emphasise the importance of incorporating these additional items. Instead, the rubric included a graded list of values for assessment, leading to differing interpretations among evaluators. Secondly, this resulted in evaluators 2 and 4 establishing their baseline assessments from different starting points. For instance, evaluator 2's performance ratings oscillated between 3.8 and 4.0, while evaluator 4's oscillation ranged from 2.9 to 3.5. In contrast, evaluators 1 and 3 maintained a narrower oscillation level of 3.9 to 4.2. This divergence in the interpretation of the rubric's criteria and their respective starting points played a key role in the varying assessment outcomes, emphasising the importance of addressing the inter-rater reliability factor when evaluating performance. It could be argued that the evaluators' professional and research expertise were somewhat assumed, given their status as experienced teachers, who in addition, had previously assessed a similar experiment involving dubbing, which shared similar characteristics with the present one.

In the following Figures, 3 to 6, one can clearly appreciate what the increase rate per attribute, and by each evaluator's perception of potential improvement were. Each feature is represented with a different colour. These figures are useful because they add information on the evaluators' appreciation of change, taking into account the scale at which they started when assessing the pre-test and what was the scale at which they finished when evaluating participants' post-tests.

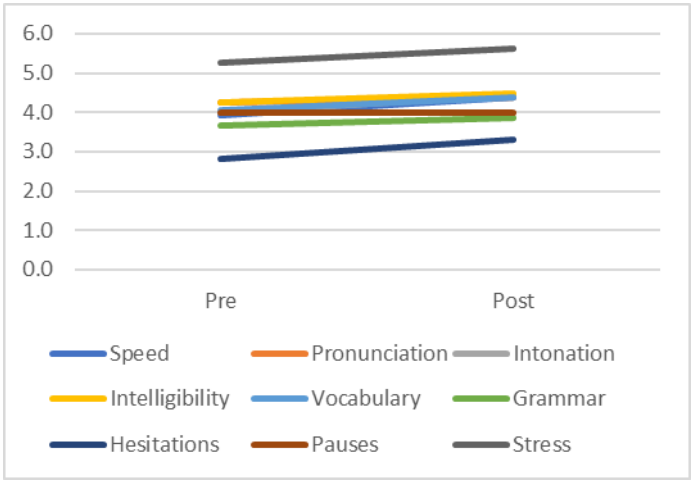

Figure 3. Pre- to post-scores for evaluator 1 (source: author)

Looking at Figure 3, one can see how evaluator 1 valued a positive increase in all features except for the pauses attribute. Most features (from a scale of 0.0 to 6.0) started just below the 4.00 rate and they improved above that same average scale. However, the *hesitations* feature was the one that started and finished at a lower scale, from just below 3.00 to just above 3.00, whereas the starting rate for stress was much higher, just above 5.00 and went to just below the 6.00 rate.

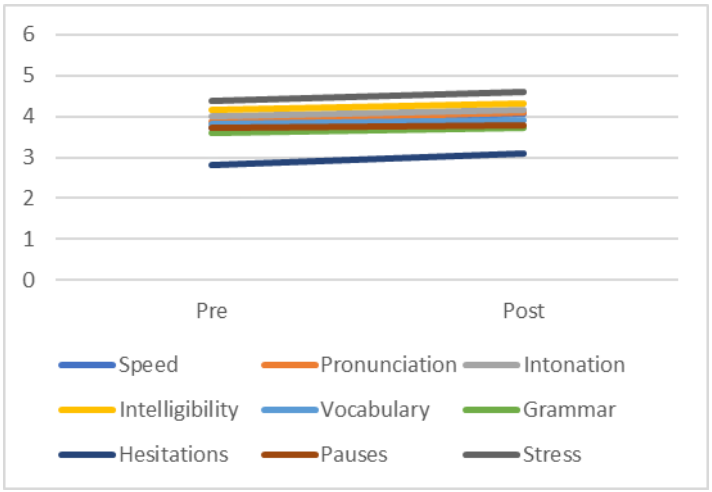

Figure 4. Pre- to post-scores for evaluator 2 (source: author)

In terms of starting and finishing points for pre- and post-tests, evaluators 2 (Figure 4) and 3 (Figure 5) bear some resemblance, as most features are assessed from below and above the scale of 4.00. However, both evaluators started scoring from a noticeably lower point than in the case of the rest of the features examined when assessing the number of hesitations for the pre-tests (just below the scale of 3.00 for evaluator 2, and just above this same scale for evaluator 3).

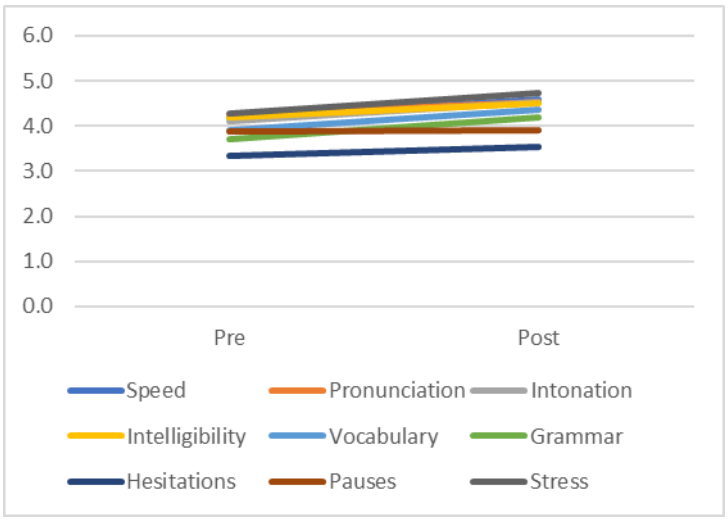

Figure 5. Pre- to post-scores for evaluator 3 (source: author)

Looking at Figures 5 and 6, another similarity in the way in which evaluators 3 and 4 assessed the tests is that both scored a consistent change from the pre- to the post-tests across all features with the exception of speech pauses. As previously mentioned, none of them perceived much of an evolution from one test to the next when assessing this particular feature.

As illustrated in Figure 6, evaluator 4 is the one that scored the highest improvement across all features from one test to the next. However, she is also the only one that tended to give the lowest scores to most features in both the participants' pre- and post-tests. Her starting scores went from 2.20 to a 3.30 rate, but also, they finished in a generally lower rate (2.50-4.00) than all other three evaluators, whose score tendency was above 4.00.

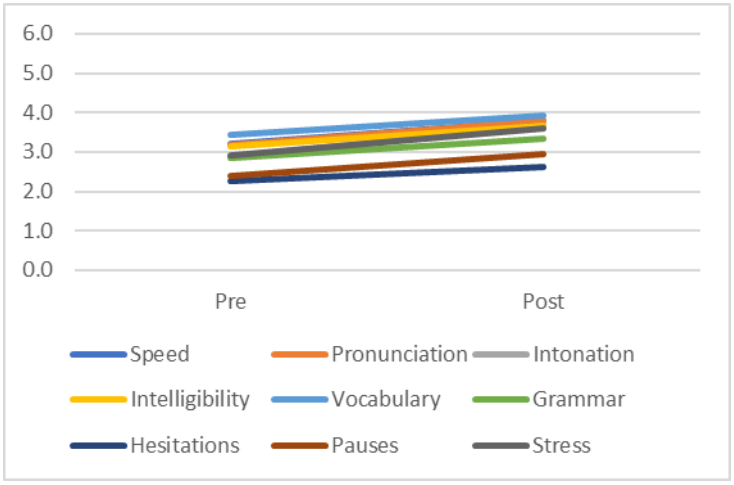

Figure 6. Pre- to post-scores for evaluator 4 (source: author)

Figure 7 illustrates in a very simple and efficient manner the evaluators' scores for each of the tests, their starting and finishing rates and the differences among evaluators. It provides a summary of the pre- and post-tests across all attributes for each evaluator in different colour lines.

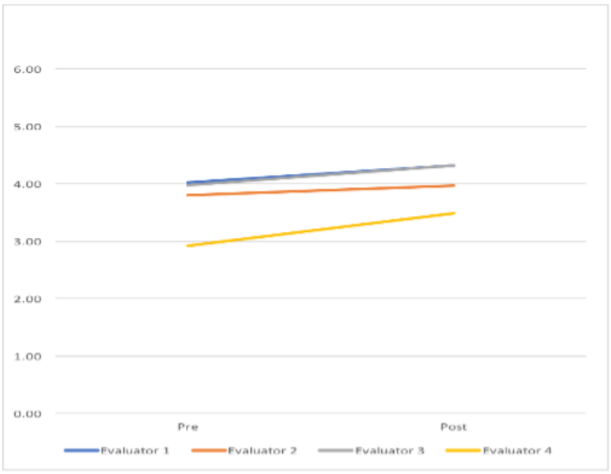

Figure 7. Summary pre- to post-tests across all attributes for each evaluator (source: author)

Figure 7 also shows how all evaluators' lines are ascending, which means that the intervention had a successful impact on learners' speech, but it is noticeably more horizontal in the case of evaluator 2's perceptions. Besides, the graph highlights how evaluators 1 and 3 assessed an average rate for improvement starting and finishing in a rather similar scale for both tests across all participants.

7. Conclusion

In most studies focusing on oral production skills based on DAD practice, attempts to quantify the actual improvement of these skills were rare, with the exception of Talavan & Lertola's study (2016). They emphasised that these findings should be interpreted within their particular context and cannot be generalised. Similarly, Navarrete's study (2018) faced limitations due to a small participant pool but highlighted learners' positive attitudes toward DAD practice and its potential for future investigation.

Statistical analysis of the data gathered from the evaluators' rubrics in this study confirms our hypothesis: DAD practice has a significant impact on oral production skills (fluency, pronunciation and intonation) in spontaneous speech. Most features showed a noticeable improvement, albeit with only minimal reduction in prolonged pauses. The experiment also highlights the need for a more comprehensive rubric to avoid differences in how evaluators interpret results and to establish a shared baseline, thereby enhancing evaluation consistency and reliability.

In light of the context presented above and the analysis of the obtained results, it is plausible to assert that oral proficiency in spontaneous speech can be effectively fostered through DAD's practice. The empirical evidence gathered suggests that the incorporation of DAD practice into language learning environments yields positive outcomes. Nonetheless, it is imperative to acknowledge that while the results demonstrate promise, further research is needed to ascertain the robustness and generalizability of these findings across diverse linguistic and educational contexts.

8. References

Bausells-Espín, A. (2022). Audio description as a pedagogical tool in the foreign language classroom: An analysis of student perceptions of difficulty, usefulness and learning progress. *Journal of Audiovisual Translation*, *5*(2), 152–175. https://doi.org/10.47476/jat.v5i2.2022.208

Brown, D. J. (2008). Research methods for applied linguistics: Scope, characteristics, and standards. In A. Davies & C. Elder (Eds.), *The handbook of applied linguistics* (pp. 476–500). Blackwell Publishing. https://doi.org/10.1002/9780470757000.ch19

Burt, M. (1999). Using videos with adult English language learners. *ERIC Digest*, ED 434539. https://eric.ed.gov/?id=ED434539

Cala Carvajal, R. (1997). Dos sistemas cara a cara. In F. Moreno, M. Gil, & K. Alonso (Eds.), *ASELE. Proceedings of the VIII Congreso Internacional de la ASELE. (The 8th International Conference of ASELE)*. *Alcalá de Henares, 17th-20th September 1997* (pp. 221–226). Universidad de Alcalá.

Calduch, C., & Talaván, N. (2018). Traducción audiovisual y aprendizaje del español como L2: el uso de la audiodescripción. *Journal of Spanish Language Teaching, 4*(2), 168–180. https://doi.org/10.1080/23247797.2017.1407173

Cenni, I., & Izzo, G. (2016). Audiodescrizione nella classe di italiano L2. Un esperimento didattico. *Incontri*, *31*(2), 45–60. https://doi.org/10.18352/incontri.10169

Chyung, S. Y., Swanson, I., Roberts, K., & Hankinson, A. (2018). Evidence-based survey design: The use of continuous rating scales in surveys. *Performance Improvement (International Society for Performance Improvement), 57*(5), 38–48. https://doi.org/10.1002/pfi.21763

Council of Europe (2020). *Common European framework of reference for languages: Learning, teaching, assessment. Companion volume with new descriptors*. Council of Europe. https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/16809ea0d4

Cox, G. C., Morrison, J.& Brathwaite, B. (2015). An assessment tool to guide students and markers. *1st International Conference on Higher Education Advances, HEAd'15.* Universitat Politècnica de València. *http://dx.doi.org/10.4995/HEAd15.2015.414*

Dawson, P. (2015). Assessment rubrics: Towards clearer and more replicable design, research and practice. *Assessment & Evaluation in Higher Education*, *42*(3), 347–360. https://doi.org/10.1080/02602938.2015.1111294

Díaz-Cintas, J. (1995). El subtitulado como técnica docente. *Vida Hispánica*, 12, 10–14.

Díaz-Cintas, J. (1997). Un ejemplo de explotación de los medios audiovisuales en la didáctica de lenguas extranjeras. In M. Cuéllar (Ed.), *Las nuevas tecnologías integradas en la programación didáctica de lenguas extranjeras* (pp. 181–191). Universitat de Valéncia.

Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies.* Oxford University Press.

Fryer, L. (2016). *An introduction to audio description: A practical guide*. Routledge.

Garza, T. (1994). Beyond MTV: Music videos as foreign language text. *Journal of the Imagination in Language Learning, 2*, 106–110.

Herrero, C., & Escobar, M. (2018). A pedagogical model for integrating film education and audio description in foreign language acquisition. *Translation and Translanguaging in Multilingual Contexts, 4*(1), 30–54.

Ibáñez Moreno, A., & Vermeulen, A. (2013). Audio description as a tool to improve lexical and phraseological competence in foreign language learning. In D. Tsagari & G. Floros (Eds.), *Translation in language teaching and assessment* (pp. 45–61). Cambridge Scholars Publishing.

Ibáñez Moreno, A., & Vermeulen, A. (2014). La audiodescripción como recurso didáctico en el aula de ELE para promover el desarrollo integrado de competencias. In R. Orozco (Ed.), *New directions in Hispanic linguistics* (pp. 263–292). Cambridge Scholars Press.

Ibáñez Moreno, A., & Vermeulen, A. (2015). Using VISP (VIdeos for SPeaking), a mobile app based on audio description, to promote English language learning among Spanish students: A case study. *Procedia -Social and Behavioral Sciences*, *178*, 132–138. https://doi.org/10.1016/j.sbspro.2015.03.169

Ibáñez Moreno, A., & Vermeulen, A. (2016). VISP: A MALL-based app using audio description techniques to improve B1 EFL students' oral competence. In E. Martin-Monje, I. Elorza, & B. García-Riaza (Eds.), *Technology-enhanced language learning for specialized domains: Practical applications and mobility*. (pp. 266–276). Routledge.

Ibáñez Moreno, A., & Vermeulen, A. (2017). The ARDELE Project: Controlled empirical research on audio description as a didactic tool to improve (meta)linguistic competence in foreign language teaching and learning. In J. Díaz-Cintas & K. Nikolić (Eds.), *Fast-forwarding with audiovisual translation* (pp. 195–211). Multilingual Matters. https://doi.org/10.21832/9781783099375-014

Ibáñez Moreno, A., & Vermeulen, A. (2021). A comparative analysis of a mobile app to practise oral skills: In classroom or self-directed use? *Journal of Universal Computer Science, 27*(5), 472–484. https://doi.org/10.3897/jucs.67032

Ichaurralde, C. (2001). *Los sonidos del español: ejercicios de pronunciación con grabaciones*. Mira Editores.

Jakobson, R. (1959). On linguistic aspects of translation. In R. A. (Ed.), *Brower on translation* (pp. 232–239). Harvard University Press.

King, J. (2002). Using DVD feature films in the EFL classroom. *Computer Assisted Language Learning, 15*(5), 509–523.

Lertola, J. (2019). *Audiovisual translation in the foreign language classroom: Applications in the teaching of English and other foreign languages*. Research-publishing.net. https://doi.org/10.14705/rpnet.2019.27.9782490057252

Mompeán-González, J. A. (2001). A comparison between English and Spanish subjects' typicality ratings in phoneme categories: A first report. *International Journal of English Studies, 1*(1), 115–156. https://revistas.um.es/ijes/article/view/47641

Navarrete, M. (2018). The use of audio description in foreign language education. A preliminary approach. *Translation and Translanguaging in Multilingual Contexts*, *4*(1), 129–150. https://doi.org/10.1075/ttmc.00007.nav

Navarrete, M. (2020). *Active audio description as a didactic resource to improve oral skills in foreign language learning.* [Unpublished doctoral dissertation, Universidad Nacional de Educación a Distancia].

Navarrete, M. (2022). La audiodescripción como actividad mediadora en el aula de lenguas. In A. Sánchez Cuadrado (Ed.), *Mediación en el aprendizaje de lenguas. Estrategias y recursos* (pp. 41–66). Edelsa: Anaya ELE.

Navarrete, M. (2023). Training the trainer: The art of audio describing in language lessons. *Translation and Translanguaging in Multilingual Contexts 9*, (2), 216–238. https://doi.org/10.1075/ttmc.00109.nav

Navarrete, M., & Bolaños García-Escribano, A. (2022). An action-oriented approach to didactic audio description in foreign language education. *Revista de Lenguas para Fines Específicos, 28*(2), 103–120. https://doi.org/10.20420/rlfe.2022.556

Rost, M. (2002). *Teaching and researching listening. Applied linguistics in action.* Longman.

Salway, A. (2007). A corpus-based analysis of audio description. In J. Díaz-Cintas, P. Orero & Remael, A. (Eds.), *Media for all: Subtitling for the deaf, audio description, and sign language* (pp. 151–174). Amsterdam: Rodopi.

Sánchez-Requena, A. (2018). *Audiovisual translation in foreign language education: the use of intralingual dubbing to improve speed, intonation pronunciation in spontaneous speech*. [Doctoral dissertation, Manchester Metropolitan University]. Open. https://e-space.mmu.ac.uk/620483/

Schaeffer-Lacroix, E. (2020). Integrating corpus-based audio description tasks into an intermediate-level German course. *International Journal of Applied Linguistics, 31*, 173–192. https://doi.org/10.1111/ijal.12294.

Stempleski, S. (1990). Teaching communication skills with authentic video. In S. Stempleski & B. Tomalin (Eds.), *Video in second language teaching: Using, selecting, and producing video for the classroom* (pp. 7–24). Teachers of English to Speakers of Other Languages (TESOL).

Talaván, N. (2006a). Using subtitles to enhance foreign language education. *Porta linguarum*, 6, 41–52. https://www.ugr.es/~portalin/articulos/PL_numero6/talavan.pdf

Talaván, N. (2006b). The technique of subtitling for business English communication. *Revista de Lenguas para Fines Específicos*, *12*, 313–346. https://ojsspdc.ulpgc.es/ojs/index.php/LFE/article/view/170

Talaván, N. (2020). The didactic value of AVT in foreign language education. In L. Bogucki & M. Deckert (Eds.), *The Palgrave handbook of audiovisual translation and media accessibility* (pp. 567–592). Palgrave Macmillan.

Talaván, N., & Lertola, J. (2016). Active audiodescription to promote speaking skills in online environments. *Sintagma*, 28, 59–74. https://doi.org/10.21001/sintagma.2016.28.04

Talaván, N., Lertola, J., & Ibáñez, A. (2022). Audio description and subtitling for the deaf and hard of hearing media accessibility in foreign language learning. *Translation and Translanguaging in Multilingual Contexts, 8*(1), 1–29 https:/doi.org/10.1075/ttmc.00082.tal

Tomalin, B. (1986). *Video, TV and radio in the English class: An Introductory Guide*. Macmillan.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Walczak, A., & Fryer, L. (2017). Creative description: The impact of audio description style on presence in visually impaired audiences. *British Journal of Visual Impairment, 35*(1), 6–17. https://doi.org/10.1177/0264619616661603

Filmography

Balda, K., & Soret, J. (2015). The Minions: Mini-Movie – Competition. [online] Available from: https://youtu.be/Q04KG7gVQtw

Cattaneo, P. (1997). The Full Monty. [online] https://youtu.be/0MTuZoNvahI

Navarra, L. (2012). One Day on Earth… Cooking Tortilla de Patatas. [online] Available from: https://vimeo.com/56434999

Ross, G. (2012). The Hunger Games. [online] Available from: https://youtu.be/B8BD9txkGL4

Vektor, J., & Payá, C. (2012). Maldición [online] Available from: https://www.youtube.com/watch?v=k-Z8xxygd2Y&feature=youtu.be
A short love story in stop motion (uploaded by Lascano, C.) [online] Available from: https://vimeo.com/877053
Turismo Perú. (clip no longer available available)

Marga Navarrete
University College London
School of European Languages, Culture and Society
Dept Spanish, Portuguese and Latin American Studies / Centre for Translation Studies
Gower Street
London WC1E 6BT
United Kingdom

m.navarrete@ucl.ac.uk

Biography: Dr Marga Navarrete is an Associate Professor (Teaching) and a Spanish Language Coordinator at University College London, UK, where she also teaches Spanish, translation and localisation at both undergraduate and postgraduate levels. Her research focuses on the impact of audiovisual translation (AVT) practice on language learners' competence and teacher training. She has taken part in a number of AVT research studies on language learning, including the ClipFlair and TRADILEX projects, where she has been designing AVT tasks and disseminating lessons learnt.