Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Bachelor thesis

Perceptual Quality Assessment of Lip-Synchrony in Dubbing

Erfassung von wahrgenommener Qualität der Lippensynchronität von audiovisuell übersetztem Material

Christian Schuler

MIN-Fakultät
Fachbereich Informatik
Studiengang B.Sc. Informatik
Matrikelnummer 6449321
christianschuler8989@gmail.com

Erstgutachter: Dr. Timo Baumann
Zweitgutachter: Prof. Dr. Timo Gerkmann

Abgabe: 24.01.2022

"I wholeheartedly want to thank Anran and Dominik, for always supporting me and even indulging me after having talked way too much about this work already! I love you guys!"
– *Christian Schuler*

Abstract

Advances in network- and media-technology enable global distribution of video material in an ever greater expanding scope. This raises the need for high qualitative translations, not only of text but of audio and visual. There are different modes of audiovisual translation and the assessment of their quality, as perceived by human viewers, has gained more interest in the last decades. This work shows the intricacies of such investigations and how they stem from gaps of a seemingly fractured research field of audiovisual translation perception. This work tackles issues barely addressed by prior research endeavours and centres around spoken language in videos, and more precisely, focuses on the perceived quality of lip-synchrony. The conceptualisation and creation of a flexible crowd-based study with over 80 participants to assess subjective video quality will be followed by the statistical analysis of the gathered data. This analysis allows making valid statements in order to answer the research questions. Furthermore, it also enables to make a statement about the difference of quality perception by different genders, contrary to studies in the past.

Zusammenfassung

Fortschritte in der Netzwerk- und Medientechnologie ermöglichen die weltweite Verbreitung von Videomaterial in immer größer werdendem Umfang. Dies erhöht den Bedarf an qualitativ hochwertigen Übersetzungen. Dies gilt nicht nur für Text, sondern auch für Audio und Bild. Es gibt verschiedene Arten der audiovisuellen Übersetzung. Die Bewertung der von menschlichen Zuschauern wahrgenommene Qualität hat in den letzten Jahrzehnten zunehmend an Interesse gewonnen. Diese Arbeit zeigt die Feinheiten solcher Untersuchungen und wie sie aus den Lücken eines scheinbar zersplitterten Forschungsfeldes der audiovisuellen Übersetzungswahrnehmung entstehen.

Diese Arbeit befasst sich mit Problemen, die von früheren Forschungsbemühungen kaum angesprochen wurden. Sie konzentriert sich auf die gesprochene Sprache in Videos und auf die wahrgenommene Qualität der Lippensynchronität. Der Konzeption und Erstellung einer flexiblen Studie mit über 80 Teilnehmern zur Beurteilung der subjektiven Videoqualität folgt die statistische Auswertung der erhobenen Daten. Diese Analyse ermöglicht valide Aussagen zur Beantwortung der Forschungsfragen. Darüber hinaus ermöglichte sie, im Gegensatz zu früheren Studien, eine Aussage über sich unterscheidende Qualitätswahrnehmung verschiedener Geschlechter zu treffen.

Contents

1. Introduction

Audio visual translation (AVT) is the process of decoding video material from one language and re-encoding it into another. It has been gaining more relevance in the past decades due to technological advances and a rise in global demand (Sánchez-Mompeán, 2021). With an increasing number of people consuming video material (Cisco, 2020) after it was translated, the question of, which aspects influence the quality of the final product, becomes ever more pressing.

1.1. Motivation

Globalisation does its part in connecting different cultures and confronting people with materials in different languages. Many markets span the whole globe now and the use of the English language is no guarantee to deliver full accessibility, even for those, who are proficient in it. Using an audio-visual translation of video material helps to overcome this language barrier between the media and its consumers.

One of these AVT techniques, called dubbing, is voice actors replacing the dialogues as accurate as possible in accordance with the mouth articulation of the actors on screen, in e. g. movies, series and commercials.

Dubbing is the preferred method of AVT in many language communities (Häsk, 2009) and dubbed videos are easy to consume for children and the visually impaired. It is produced in professional studios involving voice actors, engineers and directors.

Translating the source material as accurately as possible into the target language enables it to reach more people but not necessarily connect with them. This is because there is more to the quality of a video than merely its content and language. One has to consider the cultural background of both the video the video creator and the video consumer. (Chaume Varela, 2004)

AVT (automatic or by human experts) has to make compromises between the importance of many features. Translation processes can be described as a method to convey the meaning of content. But often this can not be done word-by-word, because the translated text could be too long to be used as a subtitle, or the translated phrase could sound unnatural to the native speaker of the language. Also, a little girl talking on screen in the deep and manly voice of Morgan Freeman could immensely confuse the viewer.

As the focus of this work, lip syncing is regarded as the reproduction of timing, phrasing, and phonetic content of the original speech segments in the target language to match the lip movements of the original performers.

Considering the word “tartle” as a short example. It is a Scottish verb meaning “to hesitate while introducing someone due to having forgotten his or her name”. In many languages, it would be hard to explain or convey its meaning in the same amount of time it takes to speak the word “tartle”. Translating “tartle” as part of a scene in a movie while keeping the semantic meaning and the timing of the speech in accordance with the source material as well as keeping the audio in synchrony with the actors’ lip movements can be considered a non-trivial task. Sometimes, there simply is no way to simultaneously achieve all conflicting goals that are part of the translation process.

The quality of a video can be assessed via instrumental measures (by a machine) or perceptual measures (by a human) where the latter are usually considered to be more reliable but are also more costly to obtain. This is even worse for the evaluation of dubbing quality.

Agreed upon/standardised metrics (a hierarchy of features) could help establish evaluation frameworks and lead to an increase in quality/performance of the training of machines and human experts alike.

This thesis attempts to build a foundation to reach these goals.

1.2. Research question

The author of this work strives to ascertain how significant proper lip synchrony is for achieving a high dubbing quality and to what degree viewers accept sub-par lip synchrony while watching. The aim is therefore to answer the following research questions:

Q1:

“To what degree is the perceived quality of a video affected by Lip-Synchrony?”

Q2:

“Is there a threshold for the degree to which artefacts in Lip-Synchrony can be tolerated by a (lay-) viewer?”

Q2-a:

“Is there a difference in the effect size for different phonemes?”

Q2-b:

“Is there a difference in the effect size for different severity of modification?”

This work does not contain dubbed video material but focuses on video material in its original language and thus serves as a stepping stone to dubbing.

1.3. Outline

Before a question can be answered, it has to be (formulated, asked, and then) understood.

That is why Chapter 2 discusses the theoretical background and gives an introduction to the topic. We first examine language research in general and then in more detail at dubbing and quality receptions thereof.

In Chapter 3, we will see the current trend in this field of study reviewing some examples. We will also clarify why it has been and still is so difficult to find an answer to many of its questions. Also, we will analyse the trend of the current research.

Chapter 4 presents the established methods and approaches for assessing the perceived video quality. Like a carpenter picking the right wood or a painter choosing the right colour, it is important to decide which framework to base the study on, with which statistical methods to evaluate the results, and what kind of norms/ITU recommendations can serve as an orienting frame around the projects/thesis-work.

The Chapter 5 will provide insights into the test material used and its attributes. On the one hand, this chapter will show the capabilities of this work to answer the questions, and on the other hand, it will reveal the limitations to be kept in mind while evaluating and discussing the findings later.

Chapter 6 will show the conceptualisation and creation of a flexible study design. Here our considerations from Chapter 3 will bear fruit and we construct the array of methods for executing and evaluating the study.

Chapter 7 presents the study itself, explaining its different phases, the participants and their distribution and how we are going to analyse the data to extract new insights from them, involving the concept of “Tidy Data” and “Exploratory Data Analysis”.

In Chapter 8, the outcome of the study will be presented, guided by graphs and images. Insights regarding subjects’ rating tendencies will be revealed and discussed, based on statistical analysis and contextualises them with regards to the limitations of data and experiment.

Chapter 9 will lay out useful contributions of this work and follow up with an outlook on future advances in AVT reception research. Finally an attempt to answer the proposed research questions will be made.

2. Theory

This chapter focuses on the theoretical background and gives an introduction to the topic at hand: the perception of imperfect lip-synchronicity in audio-visual material. First, we look at language research and how language can be processed by machines. We will go over some of the related aspects of filmmaking before going into detail about what constitutes audiovisual translations (AVT), especially dubbing and the types of synchrony that are important in this context. Finally, there are different types of quality and also different methods to assess video quality reception by the viewers.

2.1. Spoken Language Research

Phones are all the different sounds we actually produce to communicate and the object of studies in phonetics. Phonology studies phonemes, which are abstract representations of groups of phones. The phones of a phoneme sound similar and can be replaced by each other without changing the meaning of the spoken word. If, on the other hand, a phoneme gets substituted by another phoneme, this would result in a different pronunciation and meaning of the spoken word.

Different languages use different phonemes so there is not necessarily an easy generalisation. Compared to the English language, the German language has 48 phonemes grouped in 8 diphthongs (aI, oI, ...), 16 monophthongs (a, o, ...) (7 long and 9 short)(see Table .2.1) and 24 consonants (p, f, ...) (see Table 2.2).

The face of a talking person takes on different shapes for different sounds uttered. These can be grouped together into so-called “visemes” (Shdaifat et al., 2001) and then worked with and analysed which is especially interesting for applications of visual speech recognition (Cappelletta & Harte, 2012), audio-visual speech synthesis (Aschenberner & Weiss, 2005), sign language (Schmidt et al., 2013) and lip-reading (H. Bear & Harvey, 2019)

There are many different ways to classify phonemes or even map them to visemes (Cappelletta & Harte, 2012; Aschenberner & Weiss, 2005) bringing with them their respective advantages and disadvantages (H. L. Bear & Harvey, 2017).

Looking at lip synchrony is looking at the visual aspect and the auditory aspect of the material at the same time. Beyond audio visual translation, taking into consideration both the auditory and visual dimensions is also important for more general experiments regarding audiovisual material. This was found in the analysis of twelve different experiments (Pinson, 2011), where the audio and video quality were equally important in the overall audiovisual quality.

		front		central		back	
Rounded		-RND	+RND	-RND	+RND	-RND	+RND
Received Pronunciation							
close	tense	/iː/					/uː/
	lax	/ɪ/					/ʊ/
close-mid	tense						
	lax						
open-mid	tense			/ɜː/			/ɔː/
	lax	/ɛ/					
open	tense					/ɑː/	
	lax	/æ/		/ʌ/			ɒ
German							
close	tense	/iː/	/yː/				/uː/
	lax	/ɪ/	/ʏ/				/ʊ/
close-mid	tense	/eː/	/øː/				/oː/
	lax		/œ/				
open-mid	tense	/ɛː/					
	lax	/ɛ/					/ɔ/
open	tense			/ɑː/			
	lax			/ɑ/			

Table 2.1.: Vowels of English and German. The vowels are classified according to their place (horizontal) and manner of articulation (vertical).
The in this work investigated phonemes are marked green.
Source: created by the author according to (Knapp, 2019).

The influence of vision on speech understanding is called McGurk effect (McGurk & MacDonald, 1976) when speech sounds are miscategorised based on a conflict between auditory cues in the stimulus and visual cues on the speaker's face. This shows the distinct multi-modal connection intertwining of the human senses while comprehending speech.

Buchan, Paré, and Munhall (2008) perform an eye-tracking study. They find that subjects show more fixations on the mouth in trials with different talkers than compared to when it is the same talker in every trial. Interesting was also the observed subjects' behaviour in the case of reduced intelligibility of speech (by added noise). In these cases the subjects adopted a vantage point that is more centralised on the face and lengthening the duration of their gaze fixations on the nose and the mouth instead of the eyes of the talker. So if bad sound makes people pay more attention to the lips, this raises two ideas:

1.: "Watching a video" and "Listening to a video" are inseparably intertwined.

2.: Maybe measuring the quality of intelligibility of different sections of a video can reveal the different levels of importance of intact lip-synchrony for different sections. ("In particular, low intelligibility would imply a need for high lip-synchrony of translations to remain understandable to the viewer.")

More recently the dubbing effect (Romero-Fresco, 2019) has been observed in viewers: This effect describes how viewers, who are used to consuming dubbed material (in this case from Spain), focus less on the speaker's lips without even realizing it while viewers not used to dubbing (in this case from England), pay way more attention to the lips and notice disturbed lip-synchrony. In an attempt to reproduce the findings the dubbing effect has not been observed to apply to voice-over (Flis & Sikorski, 2019).

Finally, there are many factors at play when considering lip-synchrony. The visibility of

	labial		apical				dorsal			
	bilabial	labio-dental	apico-dental	alveolar	post-alveolar	retroflex	palatal	velar	uvular	glottal
English										
Plosives	p b			t d				k g		
Fricatives		f v	θ ð	s z	ʃ ʒ					h
Affricates					tʃ dʒ					
Nasals	m			n				ŋ		
Laterals				l						
Glides	w				ɹ – ɻ		j			
Taps										
German										
Plosives	p b			t d				k g		
Fricatives		f v		s z	ʃ (ʒ)			ç – x – χ		h
Affricates	pf			ts	tʃ (dʒ)					
Nasals	m			n				ŋ		
Laterals				l						
Glides							j			
Taps									ʀ	

Table 2.2.: Consonants of English and German. The consonants are classified according to their place (horizontal) and manner of articulation (vertical).
The in this work investigated phonemes are marked green.
Source: created by the author according to (Knapp, 2019).

the speakers' lips (the angle of the face of the speaker/distance to the camera and therefore size on screen). Other actions happening on screen can distract the viewers' gaze.

2.2. Film Making and Video Attributes

The AVT process can and should be taken into consideration relatively early in the filmmaking process to make a film that is accessible in other languages and even for viewers with hearing or visual impairments. (Romero-Fresco, 2013)

There are many different shots used in filmmaking. Of these, the ones concerning "field size" are of special interest since they can drastically change how much of the speakers face is visible. Barsam and Mohanan (2010, pp. 232-235) distinguished 7 such shots: extreme long shot, long shot, medium-long shot, medium shot, medium close-up, close-up, extreme close-up. A low number of close-ups and extreme close-ups in many films and series could be due to the producers trying to avoid the issue of different synchronies(Koverienė & Čeidaitė, 2020). As written by Koverienė and Čeidaitė (2020, p.4):"Recently, however, the number of extreme close-up and close-up shots has increased considerably, as stated by (Romero-Fresco, 2019), and peaks at nearly 75 % for such TV series as EastEnders (BBC, 1985–1985) (as cited in Sánchez-Mompeán, 30, 2019)"

In the near future, this trend will unavoidably lead to a stronger focus on synchronies of AVT and lip-synchrony of dubbing in particular.

2.3. Audiovisual Translations (AVT)

The term translation in itself provides enough potential for research and controversy (Chaume, 2018b) and one linguistic definition can be: "translation is a type of language

mediation, socially serving to approximate a mediated bilingual communication to a common monolingual communication".(Sokolovsky, 2010, p.4)

The process of translation is a complex cognitive process that should not be underestimated. It consists of decoding a source text's (ST) meaning and re-encoding this meaning into the target language (target text (TT)), and requiring knowledge of not only the grammar, semantics, syntax, etc., but also of the culture corresponding to the respective communities that use both languages.(Arduini & Hodgson, 2007) Ignoring the different customs of very diverse cultures can lead to alienation or even confusion for the viewer.

With the addition of "audiovisual" the translation moves into new media like movies for which there are three well-established modes that are shown in Fig. 2.1. There is subtitling, adding a written translation (of speech but also of other elements like written text or street signs) to the screen and also voice-over, adding narration-like audio in the target language on top of the original audio. Then there is dubbing which can be understood as translation in combination with the processes of taking segmentation, inserting dubbing symbols, writing dialogues and emulating natural discourse as much as lip-sync.

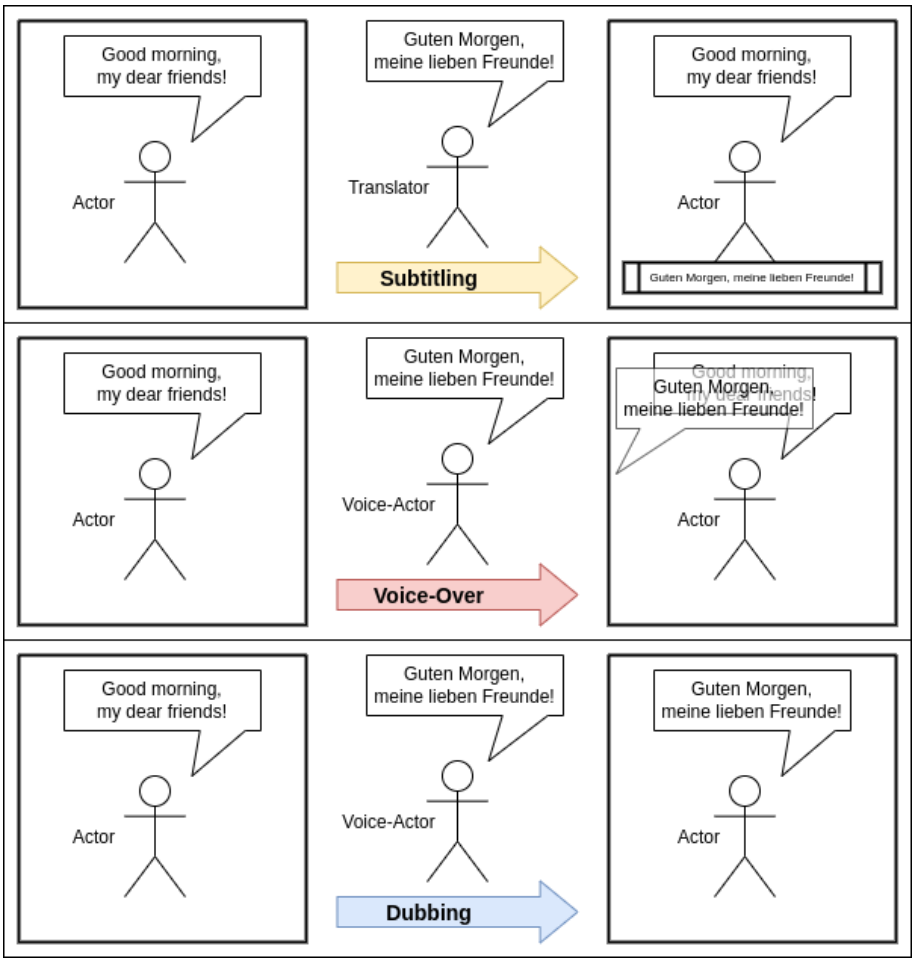


Figure 2.1.: Modes of audiovisual translation: Dubbing, revoicing the source speech into the target language; Subtitling, the addition of translated text on the screen; Voice-Over, translated speech in addition to the original sound.

A translator doing the first part requires a deep understanding of the source language, while a dialogue writer, doing the second part, needs to be fluid at the target language and have some understanding of the cultural background to be able to localise the material. Sometimes all of these steps are done by a single person. The main goal of the dialogue writer is to make the translation sound more natural and make it seem as if it was recorded in the target language, without the translation process being noticeable in the end. (Chaume, 2013)

Even though subtitling may be a more frugal variant of audiovisual translation, the advantages of dubbing in ensuring viewer engagement have been shown in viewer testing indicating that audiences are more likely to finish watching a series in case they started to view it with dubbed audio instead of subtitles. (Sánchez-Mompeán, 2021)

In many countries, dubbed content is the preferred mode of consumption of foreign media (as shown in Figure 2.2a[1] [2] [3]) and in more countries, it is at least extensively used for childrens' shows even with a differing main mode in their media landscape which in one way or another could be connected to the observation that younger viewers care less about lip-synchrony than older viewers (Huber & Kairys, 2021) or simply to the fact that children cannot read as fluently.

There is also the issue of media accessibility to consider. Besides subtitles for deaf and hard-of-hearing and audio descriptions for blind and partially sighted people, dubbing can be utilised to enable many elderly and the visually impaired to consume media and ease the partaking in the culture.(Romero-Fresco, 2018) The same applies to children, who can not yet read as fluently and dyslexic persons.

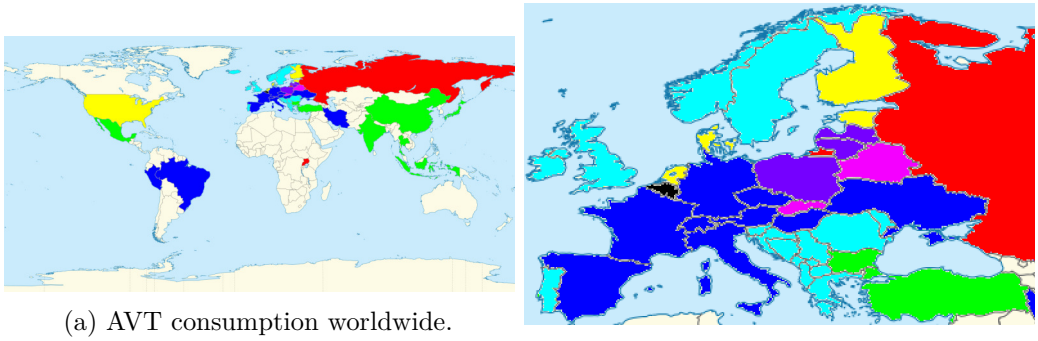
On the other hand, subtitles are shown to improve language skills (Borell, 2000) and might not even be as distracting as one might think (Perego et al., 2016). While some experiments showed no effect of the translation method on enjoyment (Wissmath et al., 2009), others found subtitles less enjoyable by viewers, not used to it, even though equally effective (Perego et al., 2016).

2.4. Theoretical Frameworks and Quality Aspects of Synchrony

The following section is about synchronies usually associated with the field of AVT.

Fodor (1976) first introduced the term synchrony to the field of audiovisual studies, grouping them by phonetic synchrony, character synchrony and content synchrony. Whitman-Linsen (1992) further differentiated more finer-grained synchronies. Lip-synchrony as an adaptation of the translation to the characters' mouth movements, especially in close-ups and extreme close-ups, was added. More recently and finding much appeal in many research

[1] https://commons.wikimedia.org/wiki/File:Film-translation_standards_around_the_world.PNG
[2] https://en.wikipedia.org/wiki/Dubbing_(filmmaking)
[3] https://bigthink.com/strange-maps/dubbing-map/



(a) AVT consumption worldwide.

(b) AVT consumption in Europe.

Figure 2.2.: Blue: Countries using exclusively full-cast dubbing, both for films and TV series. Light Blue: Dubbing only for children; Otherwise solely subtitles. Green: Countries using occasionally full-cast dubbing otherwise solely subtitles. Yellow: Subtitling. Orange: Subtitling and Voice-Over. Red: Countries using usually one or just a couple of voice actors whereas the original soundtrack persists. Purple: Voice-Over and dubbing. Pink: Countries that occasionally produce their own dubbings but generally use dubbing versions of other countries since their languages are quite similar to each other and the audience is also able to understand it without any problems (Belarus and Slovakia). Black: Belgium: The Flemish-speaking region occasionally produces its own dialect dubbing versions, otherwise solely subtitles. The French-speaking region of Wallonia and the German-speaking region of East Belgium use exclusively full-cast dubbing, both for films and TV series. White: No information.
Source: created by the author according to [1] [2] [3]

circles, Chaume Varela (2004) set synchronisation to consist of the three types of synchrony, lip-sync, kinetic synchrony and isochrony while discarding the other types since they would not directly deal with translation operations.

Kinetic synchrony is maintaining the logical relation between the words and gestures of the characters that are usually known by the audience to avoid contradictions (Chaume, 2012).

Isochrony, the supposedly most obvious indicator of a poorly dubbed product, addresses the matching of the time between the original ST and the TT speech phrases and also pauses (Chaume, 2012).

There are many more kinds of synchronies like attentional synchrony (Smith & Henderson, 2008), describing viewers attention to scenes – differing for static or dynamic scenes.

However, of the many dimensions of synchrony (see Table 2.3), the work at hand will restrict itself on Chaume's notion of lip-synchrony as a focus for evaluations.

> Lip synchrony involves adapting the TT to the articulatory movements of the on-screen character or actor of the audiovisual product. While trying to achieve lip synchrony and creating the final translated script, the dialogue writer needs to take the actor's lip movements into account. (Chaume Varela, 2004, p.10)

Fodor (1976)	Phonetic				Character					Content
Whitman-Linsen (1992)	phonetic synchrony	syllable articulation synchrony	isochrony (length of utterance synchrony or gap synchrony)	kinetic synchrony (gesture and facial expression synchrony)	idiosyncratic vocal type	paralinguistic elements (tone, timbre, pitch of voice)	prosody (intonation, melody, tempo)	cultural variations	accents and dialects	(-Content)
Chaume (2004;2012)	Lip	Isochrony		Kinetic	(-Character)					(-Content)

Table 2.3.: Typologies of synchronies.
Source: created by the author according to (Koveriene, 2015).

2.5. Importance of Synchronies

While there are many aspects to the quality of dubbing and even more to the quality of videos in general, it is a consensus (Varela, 2007) that synchronies play an important role in the perceived quality while consuming video material.

While studying the impact of Italian dubbing on the viewers' immersive experience in an audience reception study, Raffi (2020) found that the group of English viewers were significantly more immersed in the viewing experience than the group of Italian viewers watching the dubbed version of a video excerpt from an episode of Game of Thrones. This study focused on the loss of linguistic typicality and socio-cultural elements detected in the Italian adaptation, diluting differences between characters and also mentioned the importance of the viewers' cultural background, familiarity with the television series and preferred language version for viewing Game of Thrones. But the aspect of lip-synchrony, which is shown to play some part in audience immersion, has not been taken into consideration.

As stated by Koverienė and Čeidaitė (2020) "However, contrarily to subtitling, there are relatively few in-depth and systemic inquiries as well as recent advances concerning lip synchrony and methodology of teaching proper lip synchrony, except for Istvan Fodor's seminal study Film Dubbing. Phonetic, Semiotic Esthetic and Psychological Aspects (1976)"

2.6. Measurements and Recommendations

Subjective video quality is video quality as experienced by humans. It is concerned with how video is perceived by a viewer (also called "observer" or "subject") and designates their opinion on a particular video sequence. It is related to the field of Quality of Experience. Measuring subjective video quality is necessary because objective quality assessment algorithms such as peak signal-to-noise ratio (PSNR) have been shown to correlate poorly with subjective ratings (Li & Bampis, 2017). Subjective ratings may also be used as ground truth to develop new algorithms (Min et al., 2020), which, still, will be limited by the quality of the subjective data used for it. Subjective video quality tests are psychophysical experiments in which a number of viewers rate a given set of stimuli. These tests are quite expensive in terms of time (preparation and running) and human

resources and must therefore be carefully designed.

In subjective video quality tests, typically, SRCs ("Sources", i.e. original video sequences) are treated with various conditions (HRCs for "Hypothetical Reference Circuits") to generate PVSs ("Processed Video Sequences") that can then be evaluated by humans. See also Figure 2.3 for a depiction of the process. The results of these evaluations can potentially be distilled into new insights of quality and useful/applicable metrics.

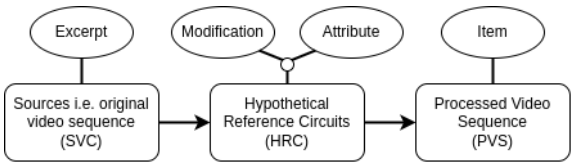


Figure 2.3.: SRCs are treated with HRCs to generate PVSs.

The work of dubbing experts is so multi-faceted that it is an enormous task to even break it down into digestible parts. Their expertise stems from many years of experience in this line of work and is connected to their subjective assessments. (Chaume, 2021)

To top it all off: there are many countries with different languages and cultures using, being used to, and preferring different modes of media-consumption, namely dubbing, subtitling, and voice-over (as shown in Figure 2.2a) which makes it even more difficult to extract rules and norms- and seemingly impossible to do the same for generally applicable ones for the whole globe/human population.

These differences might be the cause of strong and persistent fracturing of the research in this field of study.

An automated process of lip-sync evaluation enables even inexperienced users to assess the quality of a video regarding lip-synchrony. This in turn can lead to higher accessibility of this field of research.

Starting from chapter 4 a strong focus will be placed on a small set of phonemes, particularly bilabial consonants and open vowels, since Chaume Varela (2004) stated: "the translation should particularly respect the open vowels and bilabial and labio-dental consonants pronounced on screen."

3. Related Work

In this chapter, we will see in more detail and orient towards some examples how prior and current advances in the field of AVT studies look like and what they revealed. We will also clarify why it has been and still is so difficult to find definite answers for many of the questions investigated by researchers. The assessment of the perceptual quality of dubbing especially with regard to lip-synchrony can be considered to be an under-researched field (Chaume, 2013, 2016; Ranzato & Zanotti, 2019) which makes it difficult to clearly situate this work and to anchor it in established science may require a more than usual broader Related Work section.

For starters, we will look into the incorporation of the audience as important part of AVT processes. Then we will get some insights into the reception of video material in general followed by a overview of prior research in the reception of AVT to illustrate the necessity for more serious research with dubbing as focus. After that, we will see the division of opinions regarding the hierarchy of synchronies, especially lip-synchrony and how little has been done to answer this question. In the following sections we will see how different preferences of language groups seem to directly scatter this field of research, even further complicating cross-national research in this interdisciplinary field. At the end of this chapter, we will look at some state of the art solutions for dubbing.

3.1. Audience of AVT Reception

Even after years of research, the opinions of viewers seem to be mostly neglected, which are, after all, the target group and judges of quality (Gambier, 2019; Ameri et al., 2018).

As Ameri et al. (2018, p.4) stated: “Overall, despite the varying levels of attention in addressing AVT audience, ’much still needs to be done to contribute to a better insight into how different audiences make sense of audiovisual texts’ (Kruger, 2012, p. 67).”

3.2. Reception of Video Material

Other than the objective quality of a video, which can automatically be determined by machines the subjective quality of a video has to be measured/assessed by human viewers.

Often even more interesting is the whole package surrounding the video itself. As an example, in online streaming, the video could be of high quality but connection-related buffering would lower the quality of the entire viewing experience. To catch these aspects, the Quality of Experience (QoE) has been defined as “The degree of delight or annoyance

of the user of an application or service." by the International Telecommunication Union (ITU) in ITU-T P.10/G.100 (Vocabulary for performance, quality of service and quality of experience) [4]. It was early shown that the QoE involves many aspects and is sometimes less straightforward to assess and also that it is more an approximation of the actual user enjoyment since it usually contains sequences of up to 20 seconds, while watching video material in real-world applications are way longer. This leads to subjects being rather harsh in their assessment of short sequences compared to the level of artefacts they would tolerate in a feature-length film. (Reiter, 2011)

Subjective testing was shown to be a viable approach to measure QoE under different conditions and settings. Many studies (Seshadrinathan et al., 2010; Moorthy et al., 2012; Lin et al., 2015; Duanmu et al., 2016) in the past have been using very short video clips with a duration of around 10 seconds.

Unfortunately, these study conditions do not reflect typical video viewing conditions which, besides short-form videos like TikTok, contain much longer (minutes to hours) videos and therefore might be considered precise, but not particularly real, as shown in Figure 3.1.

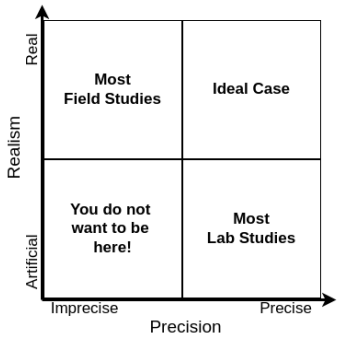


Figure 3.1.: Experiment design realism versus precision.
Source: created by the author according to (Janowski, 2019).

Therefore it is not feasible to analyse some long-term memory effects/factors affecting subjective QoE such as the recency effect (Hands & Avons, 2001).

3.3. History of AVT Reception Research

With just a few decades the field of research around AVT reception might be relatively young, however, a lot has happened in this short time. Previous research on dubbing reception encompasses (Herbst, 1997) investigating the reception of synchronization in dubbing. (Fuentes Luque, 2003) researched the reception of humour in dubbing and subtitling alike. (Peeters et al., 1988) examined the psychological perspective of the reception of subtitling and dubbing.

[4] https://www.itu.int/rec/T-REC-P.10-201711-I/en

(Perego, 2016) states that the empirical research of reception of AVT, though with a strong focus on psycholinguistic aspects of just subtitling, emerged as early as the 1980s with: (d'Ydewalle et al., 1985, 1987; Grillo & Kawin, 1988).

Besides some shining exceptions, the research in the reception of dubbing has been rather limited and overshadowed by investigations geared towards subtitling. There are the works of a group of Italian scholars on humour, culture-specific references in dubbing, and the performances of dubbing in general: (Antonini, 2007, 2008; Antonini & Chiaro, 2005, 2009; Bucaria, 2005, 2008; Chiaro, 2004, 2007).

Recently, there has also been a thriving line of enquiry in dubbing reception (Ameri et al., 2015; Denton & Ciampi, 2012; Giovanni & Fresco, 2019; Perego et al., 2015; Reyes Lozano, 2015) often in direct comparison with subtitling (Riniolo & Capuana, 2020; Koolstra et al., 2002).

Although (Ameri et al., 2018) reported an increase of reception studies in the area of AVT, there is still a considerable lack of work done on dubbing since as (Chiaro, 2014) put it, subtitling and accessibility have won the lion's share of AVT reception studies. The following should give a humble overview of some important studies and advances. First the heavily focused sector of subtitling: (Gottlieb, 1997; Widler, 2004; Antonini, 2005; Alves Veiga, 2006; Caffrey, 2008; Cavaliere, 2008; Tuominen, 2012; Perego et al., 2015, 2016; Giovanni, 2016) Some focused on voice-over: (Giovanni, 2012; Kizeweter, 2015; Szarkowska & Laskowska, 2015) And with a focus on audio description: (Chmiel & Mazur, 2012; Kruger, 2012; Szarkowska & Jankowska, 2012; Fryer & Freeman, 2013; Fryer et al., 2013; Ramos & Rojo, 2014; Ramos, 2015, 2016; Wilken & Kruger, 2016) Focused on subtitling for the deaf and hard of hearing: (Miquel Iriarte, 2014; Romero-Fresco, 2016; Szarkowska et al., 2016) Even on non-provessional subtitling: (Orrego-Carmona, 2014, 2016) Finally the localization of video games: (O'Hagan, 2009; Fernández Costales, 2016; Mangiron, 2016)

This shows a broad field of research with a history that is strongly tied to the translation and linguistic studies that could profit from more investigation into dubbing, for which many researchers call out (Ameri et al., 2018; Chaume, 2013).

3.4. Lip-Synchrony

One part of the perceived quality of dubbing, and the focus of this work, is the lip-synchrony for which there is still no clear consensus about its importance in the greater scheme of quality perception. The issue of lip-synchrony has been regarded by many scholars (Vöge, 1977; Delabastita, 1989; Whitman-Linsen, 1992; Zabalbeascoa, 1997; Barbe, 1996; Chaume Varela, 2004; Chaume, 2012) and even though not everyone regards it as important as done by (Fodor, 1976), "it remains one of the most challenging aspects of dubbing" (Koverienė & Čeidaitė, 2020, p.2).

One approach to gain more insights is to compare the different levels of lip-synchrony

between each other and ideally with additional artefacts/features. (Koverienė & Čeidaitė, 2020)

The addition of audiovisuals to the translation process already introduced new constraints (Mayoral et al., 1988) that have been conceptualised but require a finer-grained resolution of investigation to potentially reveal a threshold of degradation, (which presumably varies from viewer-group to viewer-group) at which the viewer of dubbed material will notice the discrepancy, therefore leading to a lower perceived quality of dubbing, inevitably lowering the quality of experience. With such knowledge integrated into the creation of the procedures of dubbing (Saboo & Baumann, 2019) and maybe of AVT in general, we could better orient ourselves toward results of higher quality.

For the time being, we have to make do with investigations of circumventing these constraints i.e. for machine translation only considering lip-synchrony while the speakers face is visible (Karakanta et al., 2020; Nayak et al., 2020)

3.5. Fractured AVT Research

In the early 2000s Frederic Chaume not only offered a definition of the three types of synchrony but also "tried to present a series of relevant translational factors that have to be taken into account in the analysis of synchronization. These factors have to be necessarily analysed in order to understand the existence of different norms in different audiovisual genres and audiovisual landscapes, as far as synchronization and its incidence on the translation are concerned."(Chaume Varela, 2004, p.17)

There have been endeavours to define the quality of dubbing, mostly by deciding on the features which it is built upon (Varela, 2007; Mazur & Chmiel, 2012).

There have been studies to gauge viewer preferences and attempts to comprehend a metric (Sotelo, 2015; Sinno & Bovik, 2018, 2019) or generate a better understanding of different dubbing features (Sánchez-Mompeán, 2020).

But unfortunately there are many hurdles and challenges in pursuing these possible norms:

Just Europe alone proves difficult with its many different approaches of AVT (see Figure 2.2b) for cross-national research to be done. There are many aspects that can manifest in different ways from nation to nation as observed in the case of the Dubbing Effect being detectable in a group of viewers used to dubbing from Spain (Romero-Fresco, 2019) but not for viewers more used to voice-over from Poland (Flis & Sikorski, 2019). Oftentimes reports speak of differences between nations, which is easier to convey, but a more precise notion would be to speak of different cultures and language groups, sometimes even independent of nations. For example, Finland shows films with both Finnish and Swedish subtitles, Belgium with a mix of Flemish-speaking, French-speaking and German-speaking viewers in one nation, all preferring different AVT solutions, or many English speaking nations sharing many characteristics.

These differences can be hard to overcome and not only the video viewing population but also the scientists are shaped by their experience and habits. Sometimes even to such a degree that they shun all the other modes of AVT than the one they grew up with.

A unified, i.e. global and guided by norms, research in this field might benefit many researchers (just the access to resources alone) but might be too impractical because of the varying character encountered in different cultural regions.

Similar to the topic of preferred modes for AVT per language group (as previously shown in Figure 2.2b), literature shows that the field of AVT research is also scattered and lacks cross-national approaches (Chaume, 2007).

There are also issues, observed in the dubbing industry, that play into the problematic of a fractured field. Just the practices of handling software for dubbing as an example: "This software is not available on the market; it is commissioned by a single studio and only used there. Moreover, translators do not have their own copy of the software; as a rule, they simply submit the translation and someone in the dubbing studio uses the software to segment it. Dubbing companies are afraid of industrial espionage; they fear that other companies may take advantage of their findings. Finally, little research on dubbing technology is being carried out in universities. This also obstructs the flow of information among dubbing companies." (Chaume, 2007, p.14)

Besides the issues between "the real world" of industry and "the ivory tower" of sciences, which both have an established position in the field of audiovisual translations, there also seem to be hurdles between the sciences themselves. Language sciences and computer sciences do not easily find together, possibly because the former attracts more artistically inclined people, while the latter gathers those of a practical and logical nature.

While this work and many from the here presented overview are geared towards the reception of dubbing, there are more aspects of dubbing to consider/keep in mind: As there is the potential of using dubbing for manipulation (Zanotti, 2013) and censorship (Mereu Keating, 2016) and politics (Giovanni, 2016).

3.6. State of the Art

Recent trends indicate an increased interest in the research of dubbing (Chaume, 2020b, 2020a) not at least through the contributions of Netflix investing in broader dubbing for their products (Riniolo & Capuana, 2020; Sánchez-Mompeán, 2021).

However, discrepancies between the standards used in the industry and advances in scientific research only get slowly dismantled, and a deep ravine persists between these groups of people, so far as to state "this is an open battle among schools of thought within academia, and also between academia and the industry" (Chaume, 2018a, p.17).

Cultural differences and varying preferences of lay-viewer make it difficult if not impossible to define a clear hierarchy of the features involved in achieving a high dubbing quality (Chaume, 2016; Perego, 2018; Pinson et al., 2012).

Despite many difficulties and the remaining gaps in the research of dubbing, it is undeniable that recent advances made remarkable progress, especially in automated dubbing by machines. With some state of the art solutions, at least to a certain degree, already being able to automatically translate videos from one language into another and synthesise the target language speech in the voice of the original speaker (Yang et al., 2020), or translating unconstrained talking face videos with significantly more accurate lip-synchronization than comparable models (Prajwal et al., 2020), by relying on new and more lip-sync-specific metrics.

4. Methods

This chapter will present us with already established methods and approaches to the assessment of quality reception. We will look at different aspects of studies and experiments in general. From there, we can detail how subjects and the choice of attributes of the items to be tested, or even the order in which they are presented can have a considerable impact on the outcome. Like a carpenter picking wood or a painter choosing the right colour: it is important to decide on which framework to base the study on, with which statistical methods to evaluate the results and what kind of norms can be used as orienting guidelines for this work.

4.1. Methodology

The main idea is to orient the processes of this work on already established norms and recommendations to meaningfully integrate it and its results into the present research.

As earlier elaborated, the research field of AVT is fractured enough to further complicate the task of pursuing a cross-national research approach (Perego, 2018), which could result in findings with wider inherited applicability, similar to recent endeavours in general (Rainer et al., 2015) and for subtitling (Perego et al., 2016).

This includes aspects like reproducibility and good documentation to make it possible for others to verify the findings.

But the nature of crowdsourced QoE assessment (Hoßfeld et al., 2014) runs contrary to in science often used laboratories and their well-controlled environment.

4.2. Experiment

Regarding testing methodology, this work is oriented toward already established testing methods in the field of audio-visual quality assessment.

For a broader application of the results of this work, it would be prudent to abide by established standards like ITU-BT.500

There have comparisons been made for many different testing methodologies (Pinson & Wolf, 2003; Shahid, 2014).

Not only the side of the researcher but also the behaviour of subjects in studies (Janowski & Pinson, 2015) have been investigated as much as the different approaches of preprocessing and analysis of the quality ratings (Kumcu et al., 2017).

4.3. Subjects and Objects of a Study

The results of experiments involving a study strongly depend on the people doing the study and what objects and in what manner these objects are supposed to be tested and rated. Studies indicated that subjects were not able to perfectly repeat scores and follow-up experiments came to the conclusion "First, subject scoring behaviour includes a random component that spans approximately half of the rating scale. Second, the sensitivity and accuracy of most subjective analyses can be improved if the subject scores are normalised by removing subject bias. Third, to some extent, multiple subjects can be replaced with a single subject who rates each sequence multiple times." (Janowski & Pinson, 2015, p.1) So even though subjective quality assessments are less accurate, they can be adjusted to a degree to still result in applicable/useful data.

There is an ongoing discussion in the QoE community as to whether a viewer's cultural, social, or economic background has a significant impact on the obtained subjective video quality results. A systematic study involving six laboratories in four countries found no statistically significant impact of the subject's language and culture/country of origin on video quality ratings. (Pinson et al., 2012)

Even though this study considered multiple facets of viewers backgrounds, it did, like so many similar studies, not focus on perceived lip-synchrony.

"The total number of subjects appears to be the most important control variable. Our study indicates that 24 or more subjects should be used when in a controlled environment. It is recommended to increase to 35 subjects when using a public environment or a narrow range of audiovisual quality. Smaller numbers of subjects are suitable for pilot studies, to find trending." (Pinson et al., 2012, p.23)

"The topics of visual and audio quality assessment (QA) have been widely researched for decades, yet nearly all of this prior work has focused only on single-mode visual or audio signals. However, visual signals rarely are presented without accompanying audio, including heavy-bandwidth video streaming applications. Moreover, the distortions that may separately (or conjointly) afflict the visual and audio signals collectively shape user-perceived quality of experience (QoE)" (Min et al., 2020, p.1)

4.4. Assessment and Evaluation

Since the by the viewer perceived quality of audio and of video are important factors for the QoE, there exists a high demand for accurate measuring methods. The currently most accurate ones are subjective tests, presenting the viewers with video sequences and averaging their opinions. As of yet there are no unified standards for the assessment of video quality (Topiwala et al., 2020) but a great number of studies for assessing video quality have been carried out in the past that can help to orient this works processes: Some automated (Joskowicz et al., 2014; Sotelo, 2015), others combining the aspects of objective and subjective quality (Barman et al., 2018), and some requiring an available

sound laboratory (Sotelo et al., 2017). Recent crowd-sourced advances collected over 200000 ratings from almost 5000 different participants showing results that were consistent with the ones obtained in a controlled lab environment. (Sinno & Bovik, 2018, 2019)

Crowd-sourced approaches already found ample use in the past to gauge subjective and objective picture quality (Ghadiyaram & Bovik, 2015), speaker ranking in listening tests (Baumann, 2017) and even earlier in assessing internet video quality (Figuerola Salas et al., 2013) and serve as a reasonable basis for this work.

Regardless of the type of experiment setup, there are always a number of variables that can affect the outcome/results of a study i.e. the subjects environment or personal bias (Pinson et al., 2012).

4.5. (ITU) Recommendations

The International Telecommunication Union (ITU)[5] is the United Nations specialized agency for information and communication technologies and provides a vast array of recommendations for information technology solutions and evaluations thereof. Following internationally accepted and well-established guidelines in a certain sense, streamlines the research process and can lead to an easier replication of experiments.

Of the for this work sighted and viewed ITU-Recommendations (see Table 4.1 for an overview), the two most promising were ITU-R BS.1116-3 and ITU-R BS.1534-3. These provided the cornerstones for considerations while planning and executing the study.

Identifier	Name
ITU-R BT.500-14(10/2019)	Methodologies for the subjective assessment of the quality of television images
ABC/HR (ITU-R BS.1116-3)	Methods for the subjective assessment of small impairments in audio systems
ITU-R BS.1534-3	Method for the subjective assessment of intermediate quality level of audio systems
ITU-T P.910(04/2008)	Subjective video quality assessment methods for multimedia applications
ITU-T P.1204(01/2020)	Video quality assessment of streaming services over reliable transport for resolutions up to 4K
ITU-T P.913(03/2016)	Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment

Table 4.1.: ITU-Recommendations.

The scope of this work coupled with the still ongoing global pandemic made it near impossible to fully abide by the laid out recommendations as an example, inviting study participants to a sound laboratory. But in the spirit of comparability, the ITU-Recommendations, even though targeted at audio material, have been followed to the most parts and inspired many of the decisions regarding this study/project as a whole. In the ITU-Recommendations, it is stated, that there are no perfect and strict rules for any research-endeavour but every researcher has to decide on an on-case basis, not at least because of some recommendations partially contradicting each other.

The recommended Multi-Stimulus Test with Hidden Reference and Anchor (MUSHRA) is well established in audio coding research (Völker et al., 2015, 2018) and has already been used for video material (Mohammed & Färber, 2010).

[5] https://www.itu.int/en/Pages/default.aspx

But even MUSHRA is not without critics (Mendonça & Delikaris-Manias, 2018; Zieliński et al., 2007) and like so often in life there is no simple “fits-all”-solution for many problems and so final statistical analysis has to be accompanied by careful considerations.

While ITU-R BS.1534-3 recommends using test items of about “10, but not more than 12 seconds”, and “no more than 12 signals (9 systems under test, 1 hidden low anchor, 1 hidden mid anchor and 1 hidden reference)”, first tentative experiments lead to the conclusion, that the recommended durations for the items are way too high for the study at hand and that 12 videos are too many to reasonable cross-compare with each other. Therefore the duration of around 3 to 4 seconds and an amount of 7 (3 systems under test, 1 hidden low anchor, 1 hidden mid anchor and 1 hidden reference) will be the goal for the material of each trial in the following study.

4.6. Statistical Analysis

While MUSHRA testing is usually followed up by ANOVA for statistical analysis, more than two groups of unrelated, (at least partially) non normally distributed variables is a good basis for applying the alternative Kruskal-Wallis test. For more on statistical analysis tests, see Figure 4.1.

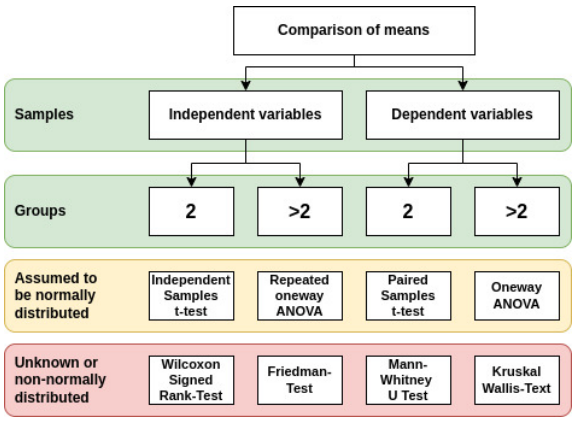


Figure 4.1.: Tests for statistical analysis and how to decide which one to use.
Source: Summarised by the author.

The early stages of the conceptualizing of the study should go hand in hand with an idea of how the resulting data could and should look like to make the analysis manageable. One established approach is the one of Tidy Data which also guided the work at hand. Wickham wrote “Like families, tidy datasets are all alike but every messy dataset is messy in its own way. Tidy datasets provide a standardized way to link the structure of a dataset (its physical layout) with its semantics (its meaning).” (Wickham, 2014, p.2)

More details regarding the statistical analysis using R can be found in the code and comments[6].

[6] https://github.com/christianschuler8989/dubbing-quality-lipsync

5. Material

This chapter will clarify the requirements of usable test material and why it is not as easy as picking them from a nearby tree. The Merkel Corpus as a source of video material will be showcased and we will see what the editing process entailed to get the video sequences that subsequently were used for the study. On the one hand, this shows the capabilities of this work regarding answering the proposed questions, and on the other, it will reveal the limitations to be kept in mind while evaluating and discussing the findings later on.

5.1. Material Attributes and the Merkel Corpus

There are plentiful resources in form of databases of audio, image and also video material[7] ready to be used in experiments and even though they recently reached a "large scale" (Cheng et al., 2020), there is a considerable lack of available video material with a focus on synchrony and distortions thereof. To this date, the author of this work was not able to find a dubbing – or lip-synchrony – database. Consequently, the material for the experiments of this work first had to be created, for which some considerations had to be made.

The video material should provide a clear view of the speaker's lips for which the angle of the speakers' face and the distance to the camera must be low enough, to ensure visibility of the speaker's lips on different screen sizes. Additional movement like in action movies makes the lip-synchrony detection impractical. The video should have a high fps and image resolution to make it easier to edit the sequences without introducing noticeable unwanted artefacts. These artefacts can reveal themselves as jumps of the speaker's face if the edit is in the middle of a short head movement or as cracking sounds in adjusted audio segments.

The Merkel Corpus(Saha et al., 2022) contains television-quality web-streamed data with former German chancellor Angela Merkel as speaker. This corpus provides a dataset of videos, corresponding transcripts, other metadata, and recently even forced alignment of utterances spoken by Angela Merkel in a controlled environment. There is a large array of different face angles (see Figure 5.1) and types of shots (see Figure 5.2) of the speaker available in this corpus.

In cases where there are multiple people on-screen or when the speaker has its back to the viewer(see Figure 5.3), it is necessary to decide the identity of the speaker and if the speaker's lips are even visible. (Nayak et al., 2020)

This kind of material had to be excluded from the current work and would not have delivered any insights towards the proposed research questions.

[7] https://stefan.winkler.site/resources.html

Figure 5.1.: Different face angles of speaker's face in the Merkel Corpus.

(a) Medium shot left

(b) Medium shot right

(c) Medium close-up left

(d) Medium close-up right

(e) Close-up left

(f) Close-up right

Figure 5.2.: Potential combinations of different shots and face angles in the Merkel Corpus.

The scope of this work made it necessary to exclude many of the possible attributes for inspection. Instead of many different face-angles, we only see the speaker in full-frontal (profile) shots. There is only material taken from the Merkel Corpus and therefore only one speaker, always speaking in the same language (German) instead of many speakers using varying languages and/or even actual dubbing. Of the 7 in (Barsam & Mohanan, 2010) categorised basic types of shots regarding a subjects closeness to the camera (used in cinematography), only two, the American Shot and the Medium Shot, got included in the material of this work (see Table 5.4).

5.2. Exploratory Pre-Study-Experiments

In literature it has been stated, that bilabial consonants, labio-dental consonants and open vowels are the most visible parts of a person's speech.

(a) Medium shot plus another person.

(b) Back of speaker plus another person.

Figure 5.3.: Examples of special cases found in the Merkel Corpus.

(a) Medium shot 1

(b) Medium shot 2

(c) Medium close-up 1

(d) Medium close-up 2

(e) Close-up 1

(f) Close-up 2

Figure 5.4.: Examples of different shots found in the Merkel Corpus. (a) and (b) have been used in this work.

Therefore these will be the focus of the first exploratory experiments, which started by modifying the vowel durations inside an audio file.

5.3. Creation of Video Material for Testing

There are different strategies and many considerations to take into account while selecting test material for a study, but the limited scope of this work deemed it reasonable to drastically limit this selection. As earlier laid out in Chapter 2, (Chaume Varela, 2004) recommended open vowels, bilabial consonants and labio-dental consonants for this kind of investigation. From these the open vowel “a” and the bilabial consonant “p” have been included in this work. By applying a healthy dose of intuition while exploring the provided data from the Merkel-Corpus semi-manually, viable video sequences had to be found. For this, a script detected occurrences of the interesting target phonemes based on the generated TextGrid files and extracted these as short video sequences. These extracted sequences were first checked for the distance of the speaker and the angle of their face to get a reasonable selection, before their number was narrowed down by excluding those where the speaker was moving too much.

Therefore the data of the Merkel Corpus was viewed/searched for occurrences of these with some limiting parameters in mind as the face of the speaker should be clearly visible and the distance should be as short as possible. Also, the neighbouring phonemes of the ones in focus were considered to have some reasonable variety. Additionally, the decision was made to narrow the selection down to sequences taken from a single podcast for easier comparability. Even though the Merkel Corpus provides a considerable array of different video types (resolution, angle of the face, ...) this would also introduce new variables for the following/final data analysis. These picked/found sources i.e. original video sequences (SRCs) then had to be modified using different types/combinations of modifications and attributes, called Hypothetical Reference Circuits (HRCs) which would result in Processed Video Sequences (PVSs) (see Figure 2.3) which then would be used for testing in the study.

In the following shifting by a single frame equals shifting by 0.04 seconds, based on the used videos all having 25 fps. For a depiction of measurement conversion see Figure 5.5.

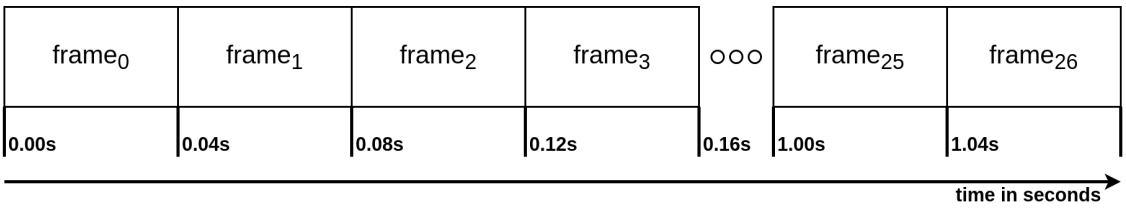
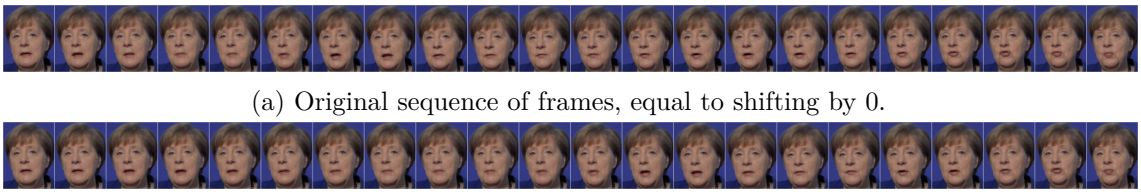


Figure 5.5.: Measurements for modifications of audio and video.

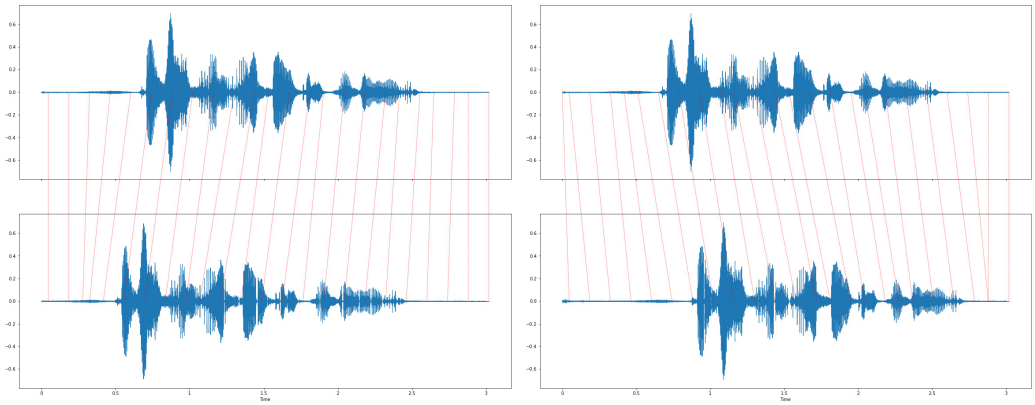
Shifting (repositioning) the target-phoneme to left by 0 frames (= 0 * 0.04 seconds = 0.00 seconds) in the visual-dimension results in no change/the original video sequence Figure 5.6a.



(a) Original sequence of frames, equal to shifting by 0.

(b) Sequence of frames visually shifted to the right.

Figure 5.6.: Shifting of visual.



(a) Shifting (repositioning) of the target-phoneme to the left by 6 frames (= 6 * 0.04 seconds = 0.24 seconds) in the audio-dimension. (Audio shifted left - AudioLeft - aL)

(b) Shifting (repositioning) of the target-phoneme to the right by 6 frames (= 6 * 0.04 seconds = 0.24 seconds) in the audio-dimension. (Audio shifted right - AudioRight - aR)

Figure 5.7.: Shifting of audio.

While shifting (repositioning) the target-phoneme to right by 6 frames (= 6 * 0.04 seconds = 0.24 seconds) in the visual-dimension results in a new video sequence Figure 5.6b which after re-combining with the original audio results in an item with partially impaired lip-synchrony.

A depiction of shifting audio to the left can be seen in Figure 5.7a and analogous for shifting to the right in Figure 5.7b.

These impairments are strongest at the location of the specified phoneme, weaker in the close proximity to it and non-existent in the rest of the video sequence. Modification of very short video sequences necessitated the addition of a temporary buffer around the target phoneme to not disrupt the editing process. Even with this buffer the effect of disrupted lip-synchrony can under certain circumstances kind of flow over the ends of the video sequence and introduce noticeably different video sequences based on their start or end.

The shifting of audio and visual into different directions results in the strongest possible disruption of lip-synchrony. By shifting the audio to the left and the visual to the right, the distance from the targeted phonemes visual representation to its audio component increases two-fold. In the following, these modifications can be identified by “iL” for a shift to the left and “iR” for a shift to the right.

Shifting of audio and visual into the same direction results in video sequences with their lip-synchrony relatively intact. Even if differences in the functionality between the audio-shift and the visual-shift can create small degradations in lip-synchrony and the audio-modification alone leads to clearly perceivable artefacts, the resulting item should end up less preferable than the unmodified reference but still superior to the lower-anchor. In the following, these modifications can be identified by "uL" for a shift to the left and "uR" for a shift to the right.

Simply relocating audio or visual segments of a video sequence is too strong of a disruption of the flow/continuity of the media and further adjustments were needed. The following explanations work analogous for left and right respectively.

To relocate the visual frames of a specific phoneme in one direction makes it necessary to replace some of the frames on this side and add additional frames on the other side of the moved phoneme. For a phoneme being moved to the left one might simply squeeze it in between already existing frames of other phonemes. But for the final video, this means that the right neighbour gets pushed further to the right and this moves the next phoneme and so on until two phonemes meet each other, where earlier the target phoneme was acting as a buffer between them. This results in clear cuts and unrealistic jumps of the actor and their actions in the video. This has also the effect of every phoneme located between the target phonemes starting position and end position being moved and creating an array of lip-synchrony distortions. To prevent this from happening, or at least reduce this effect, every second frame in front of the target phoneme gets removed from the sequence (how many depends on the length of the final item and should be decided by trial and error on an on case basis) smoothing the transition Figure 5.8. With the material used in the study of 25 fps, this approach resulted in acceptable video sequences and could likely achieve vastly better results with higher fps material. But removing frames from the sequence shortens the visual part of the video and introduces new artefacts between the original and the final position of the target phoneme. Therefore it is necessary to introduce new frames to the sequence on which the final video will be composed of. Just copying every second frame starting from the first frame to the right of the target phoneme proved to be clunky and lead to an alternative approach. Every new frame is composed of 50% of each of its final neighbours to get a smoother transition. Again, a higher fps might improve the results of this method or might even make it redundant. This whole process works analogously for the left and the right side of the target phoneme.

A similar problem arose in the editing of the audio position of phonemes. Just moving the audio would lead to overlapping speech or an entirely asynchronous video. To prevent this it was necessary to modify the duration of the surrounding phonemes, similar to the visual modifications, by shortening the ones positioned in the direction of shift/relocation and lengthening the ones in opposite direction. While compressing just omits information, the process of stretching is more complicated and has to bridge unknown parts of the audio with new approximated information. Compressing and stretching the audio of the

surrounding phonemes by a specific factor (again, found by trial and error) reduced the severity of introduced unwanted artefacts and disruptions of lip-synchrony.

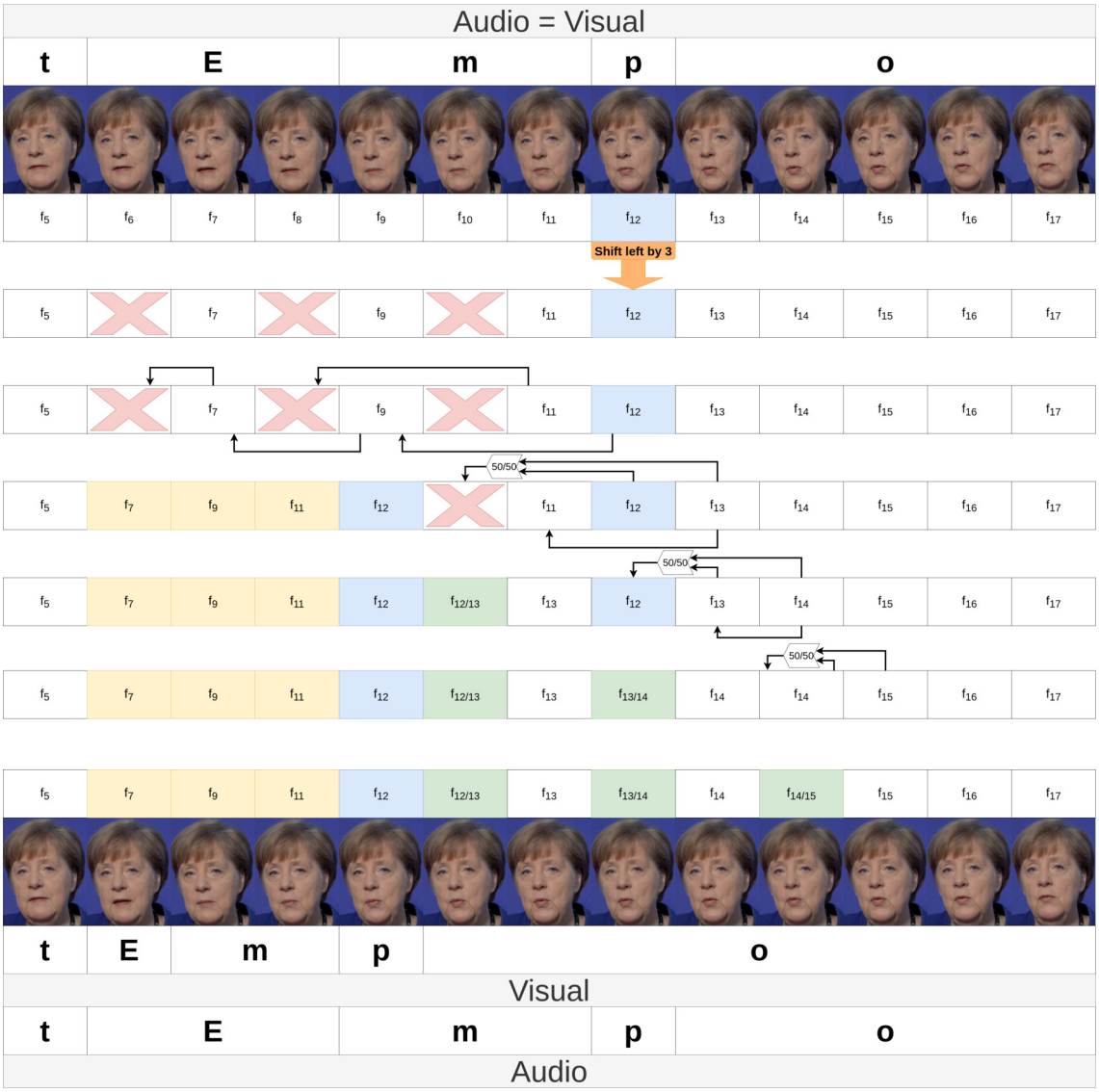


Figure 5.8.: Shifting specific phonemes' visual location in time shown with actual corpus-material.

The complete process of material editing Figure A.1 and the used data structure Figure 5.9 can be retraced/comprehended on [8].

[8] https://github.com/christianschuler8989/dubbing-quality-lipsync

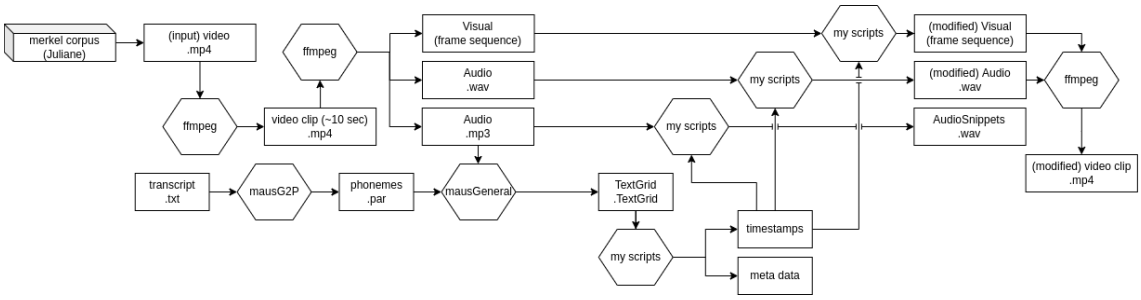


Figure 5.9.: Pipeline of video material editing.

6. Implementation and Approach

Now it is time to get acquainted with the underlying framework for crowd-based testing BeagleJS[9]. Before we look at the study itself we have to mind the parts a decent study is made of and get to know some of the vocabularies to prevent misunderstandings. This chapter ends with a clarification of how the overall study was conceptualised in many separate parts to enable a dynamic execution despite the uncertainty about participation numbers.

6.1. Setup of Test Framework for Crowd-based Testing

To reach as many people as possible and even enable cross-nation-wide participation it needed an online accessible solution for the study implementation. This was accomplished by building on the already existing framework for creating browser-based listening tests BeagleJS (Kraft & Zölzer, 2014) (browser-based evaluation of the audio quality and comparative listening environment). BeagleJS is purely based on open web standards like HTML5 and Javascript enabling any modern web browser to run the tests and at the moment supports ABX and MUSHRA style testing.

Details of the implementation can be found in the code which is publicly available at [10].

6.2. A Study

In the spirit of the ITU-Recommendations, the following will use the term “listening panel” for the group of participants even though it was not a test only involving hearing sounds but also visuals. All the participants (each called subject) make up the listening panel of the study responsible for producing the data to be analysed. For this to happen every subject goes through a separate session made up of several trials to be finished. Every trial consists of a small set of items, in general, called excerpts (in this case short video sequences of around 3 seconds duration) that have to be evaluated according to certain criteria. As can be seen in Fig. 6.1 these sets of items for every trial contain the not to be rated unmodified reference, three different versions produced by the system under test alongside three different anchors. One of these, the High-Anchor, is in actuality just a copy of the unmodified reference, the next one, the Middle-Anchor, is a slightly modified version, which should mostly still be synchronous, and finally the Lower-Anchor, a strongly

[9] https://github.com/HSU-ANT/beaqlejs
[10] https://github.com/christianschuler8989/dubbing-quality-lipsync

modified version with barely any intact lip-synchrony. To finish a trial a subject has to give a grade, according to a given scale from 0 to 100, to every single presented item, after having watched and compared them between each other.

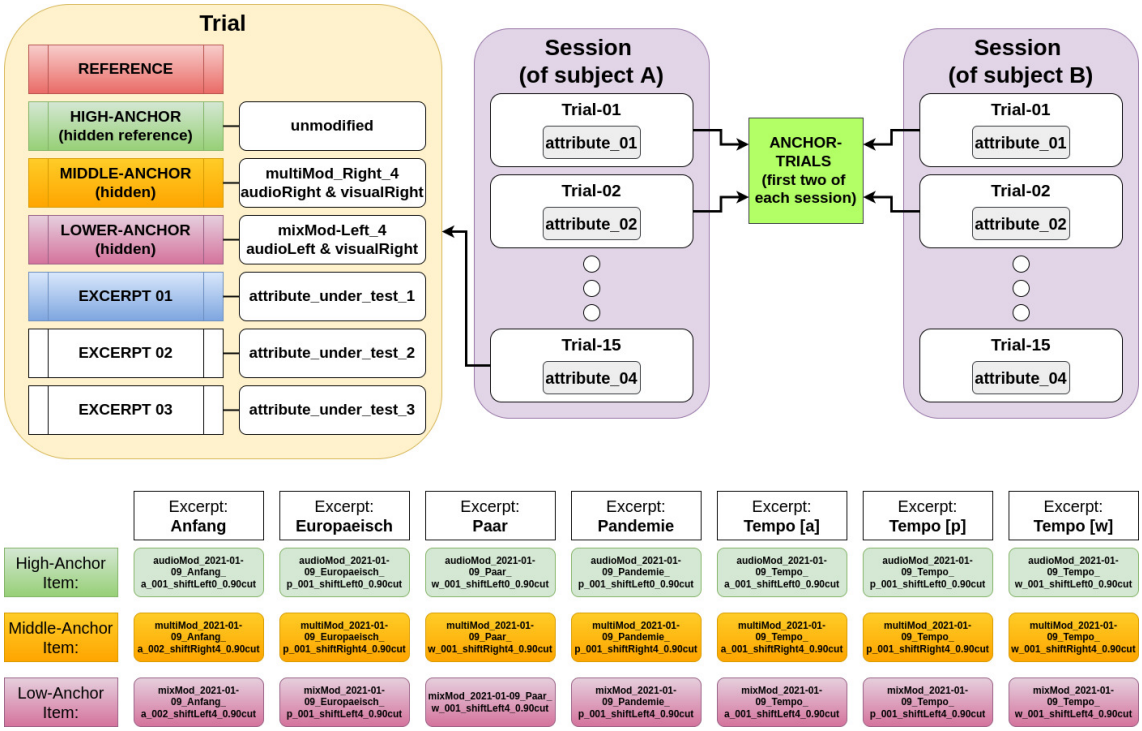


Figure 6.1.: The concept of trial structure.

All of these provided ratings make up the grades Fig. 6.2 that can then be used for statistical evaluation of the different systems under test. In this context, the term object refers to the combination of chosen phoneme (a, p, or pause) and the applied mode of modification (shifting audio left or right, shifting visual left or right, or combinations thereof). (In the later following statistical analysis part the term object refers to just the chosen phoneme, while the modification is called modification at that point) These objects are created in different levels of severity of modification (1 frame, 2 frames, ...), called an attribute, and also for a selection of different excerpts, which are the video sequences used in this study.

Subjects can be evaluated for their ability to detect disrupted lip-synchrony in the excerpts by their rating of the different anchors. The High-Anchor (hidden reference) should always be rated as the best, the Middle-Anchor as worse than the High-Anchor and the Lower-Anchor even worse than anything else. (Since it is created by shifting the audio 4 to the left and the visual 4 to the right and therefore having the strongest disruption of lip-synchrony, while every other combination is being shifted by 3 or less and the single modifications only by up to 6, but never 8.)

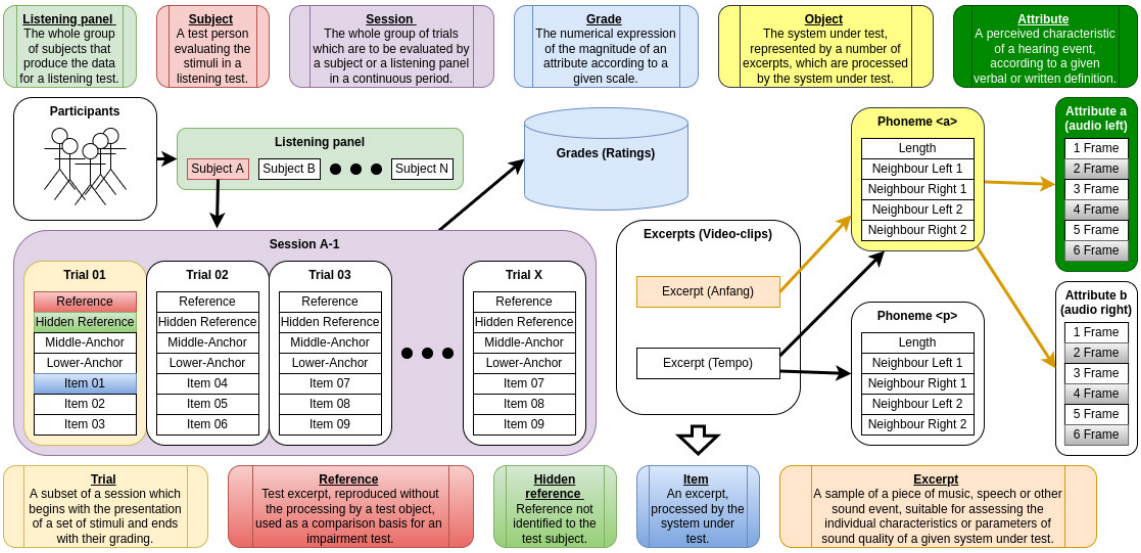


Figure 6.2.: The structure of a study in general terms.

6.3. A Study in Blue

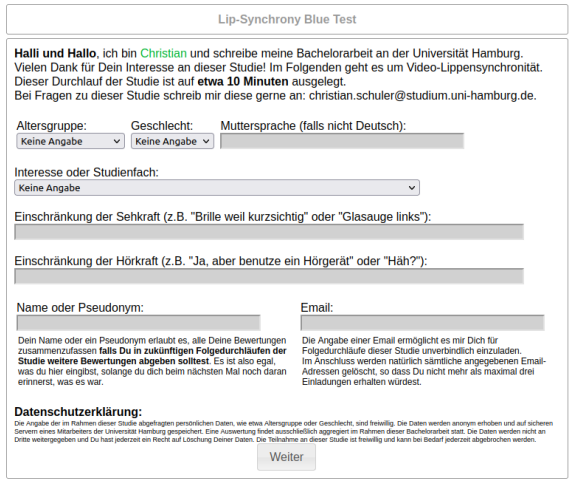
Since it was at the beginning quite unpredictable what number of participants would be reached for this experiment, some considerations had to be made. Acquiring many ratings for just a few test items would enable a more precise analysis. But the undecidedness on which features were the most important would make it necessary to include a vast array of different test items. Therefore this work applies a study in multiple parts to deal with the limitations: The first part of the study includes vastly varying test items for the subjects to rate, has affectionately been named 'Blue', and is also recognisable by this colour in the corresponding Fig.6.5, 6.3, 6.4. Once a certain amount of results have been reached, different test items (first the "Yellow" ones, then the "Red" ones, and so forth) should be presented to the participants of the study. This modular approach enabled a very flexible execution of the study, easily adjusting to low or high numbers of reached participants.

Just like the anchors inside of each trial, providing a means to evaluate the ability of each subject to evaluate the lip-synchrony degradations, there have been specific trials at the beginning of each session to serve as between-session-anchors. Based on these anchor trials it is possible to compare the rating tendencies of all the subjects, unrelated to in which part of the study they participated. This way made it unnecessary for every subject to rate every item and therefore a smaller number of subjects was needed.

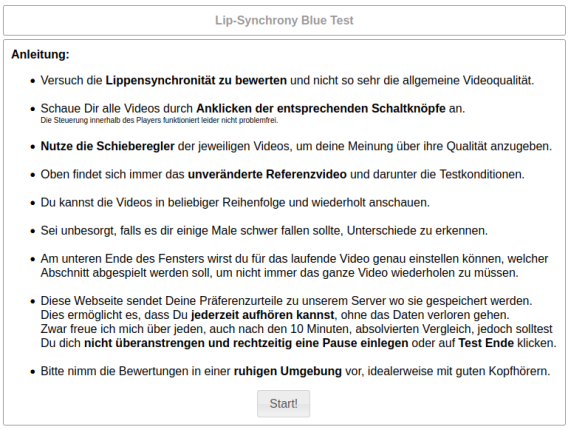
This is how the study presented itself to the participants Fig. 6.1a and where their metadata was recorded.

The next part was the page with instructions Fig. 6.1b explaining how to properly proceed while testing/comparing the video material.

Here we see the layout of a single trial inside of the study at the beginning right after opening Fig. 6.2a and at the end of rating by the subject Fig. 6.2b.



(a) Introduction at the beginning of every study session.



(b) Instructions following the introduction.

Table 6.1.: Start of a study session.

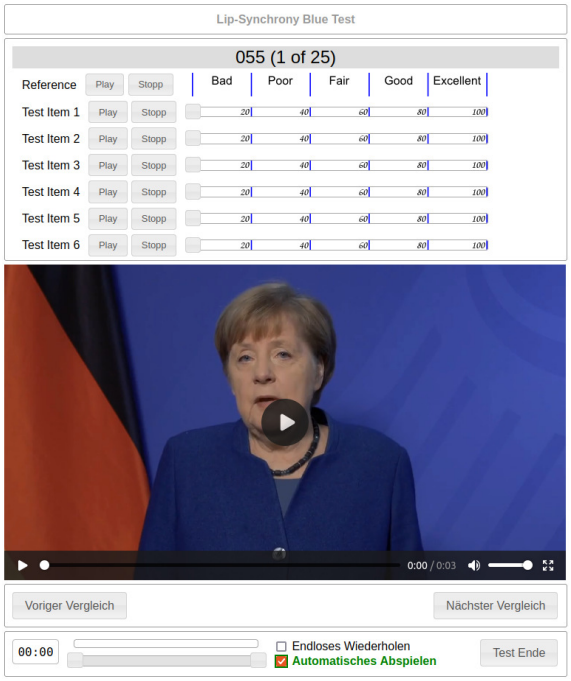
6.4. A Study Dynamically Designed

Since it was hard, in the beginning, to gauge which aspects exactly should be tested and also how many participants for the study could be found- the study got structured in several runs to be most flexible.

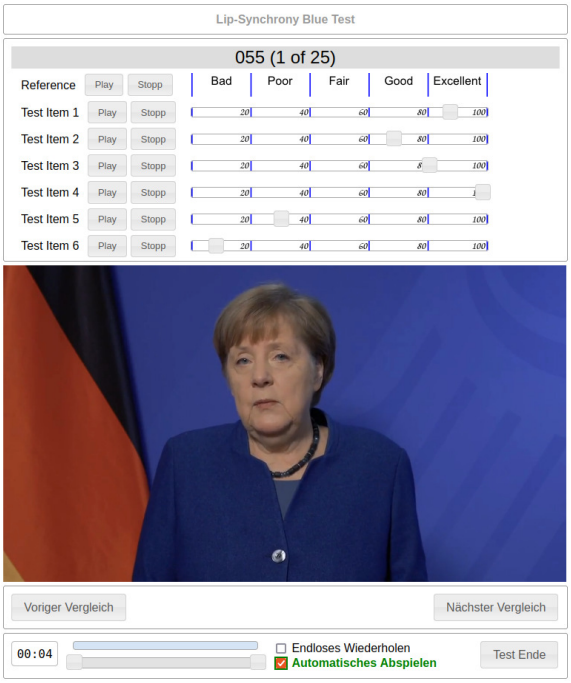
On the one side, there is the pursuit to find the relevant aspects to modify for resulting in a change of perceived quality in lip-synchrony. For this, it might seem reasonable to create all thinkable combinations and then let them be tested. But on the other side, we require data that can be statistically analysed which is hardly possible if every type of modification has only been seen and rated by one or two subjects. The recommendations provided in the literature range from 4 to 40 and more subjects per item.

Unfortunately, it turned out that not many people made themselves available to spend extended periods of time watching Angela Merkel, repeating the same phrases again and again, to evaluate the lip-sync quality of the video sequences. Consequently, the acquired ratings were used to calculate a general rating tendency for the combined listening panel which consisted of the mean rank of each test item. Based on this now every subject could be evaluated how close their rating corresponded to the general rating of each test item they encountered in the study. The subjects with the lowest divergence from the general rating tendency were contacted and two could be convinced to finish the remaining parts, marked in different colours Fig.6.5, 6.3, 6.4, of the study. Naturally, the ratings of only two subjects can not properly be statistically processed, but only serve as an approximation of what the general rating tendency of the listening panel might have looked like. Based on these results a new subset of trials has been put together (called “Pink” and specially marked, since it consists of trials from all other colours) in hope of shedding some more light on observed conspicuousness.

The final study comprises 210 unique trials of which two have been used as anchors in



(a) Beginning of a trial inside the study session.



(b) Possible rating at the end of finishing a trial.

Table 6.2.: Layout of a trial.

each of the 10 different runs/parts. Every trial contained three pre-specified anchors and three items corresponding to one type of modification, where the audio or the visual got shifted to one direction Fig. 6.3, where the audio and the visual got shifted into the same or opposite directions Fig. 6.4, and even a few where items of different modifications or items of different objects (phonemes) got mixed together Fig. 6.5.

6.5. Final Distribution of the Test Material

On the one hand it could be of advantage to create a great number of unique test items to prevent some of the effects observed in the past. Because a possible variation of talker identity has been shown to result in viewer evincing varying behaviour(Buchan et al., 2008). And to include or not include repeated scenes in an evaluation experiment can have an effect on the outcome of subjective video quality assessment as investigated by (Janowski, 2019). This can, at least in part, be confirmed by the author's experience. The two anchor trials, used in every session of the entire study were constructed from the same excerpt ("Tempo") and had to be the very first trials of each session, to ensure that these anchoring trials would always be rated under as close as possible circumstances for later comparison. If they would have been randomly ordered with the rest of the trials, they could have been rated in the end of a long session, with the subject already being rather exhausted. Or, even more likely, the subject might have stopped and ended the session, before even encountering one of the anchor trials. But since all the remaining trials were randomly

Identifier	Spoken text	Length
Anfang	Von der ich am Anfang gesprochen habe	3 s.
Impfangebot	Jedem der es das moechte - ein Impfangebot machen koennen	4 s.
Zusammen	lassen Sie uns desshalb zusammen weiter das tun - was noetig ist	4 s.
Tempo	werden es mehr - das Tempo wird zunehmen	3 s.
Pandemie	gefaehrliche zweite Welle der Pandemie - in der unser Land	3 s.
Paar	Langsamer Start - ein paar hunderttausend sind geimpft	3 s.
Europaeisch	Stoffe nicht national - sondern europaeisch organisiert haben	4 s.

Table 6.3.: Selected excerpts for creation of test material.

ordered and also contained a number of items built from the same excerpt as the anchor trials, it would occasionally happen, that a subject had to rate the same excerpt, even though with different combinations of object, modification, and attribute, for five times in a row, right at the beginning of the session. This was the most glaring of the, through the comment functionality of the testing framework, but also verbally in person, received feedback and cause for a lot of annoyance for many participants.

Contrary to this goes the attempt to include as few variables as possible into the analysis. Even though this work should be perceived as an exploratory experiment, too many factors would blur the result's accuracy. Some information of the excerpts, which the items of the study were created from, can be seen in Table 6.3.

The distribution of test material combined over all parts of the study in regard to object-excerpt combinations (see Figure 6.6a) and object-modification combinations (see Figure 6.6b).

The distribution of test material that were part of the study part named "Blue" in regard to object-excerpt combinations Fig.6.7a and object-modification combinations Fig.6.7b.

The distribution of test material that were part of the study part named "Pink" in regard to object-excerpt combinations Fig.6.8a and object-modification combinations Fig.6.8b.

Figure 6.3.: The structure of the study sessions recognisable by colour: All trials containing audio or visual shifts.
Name of excerpt followed by the object (phoneme) being modified. On top is the applied type of modification noted while the numbers indicate the severity of modification. The hexagons show the internally used TrialID.

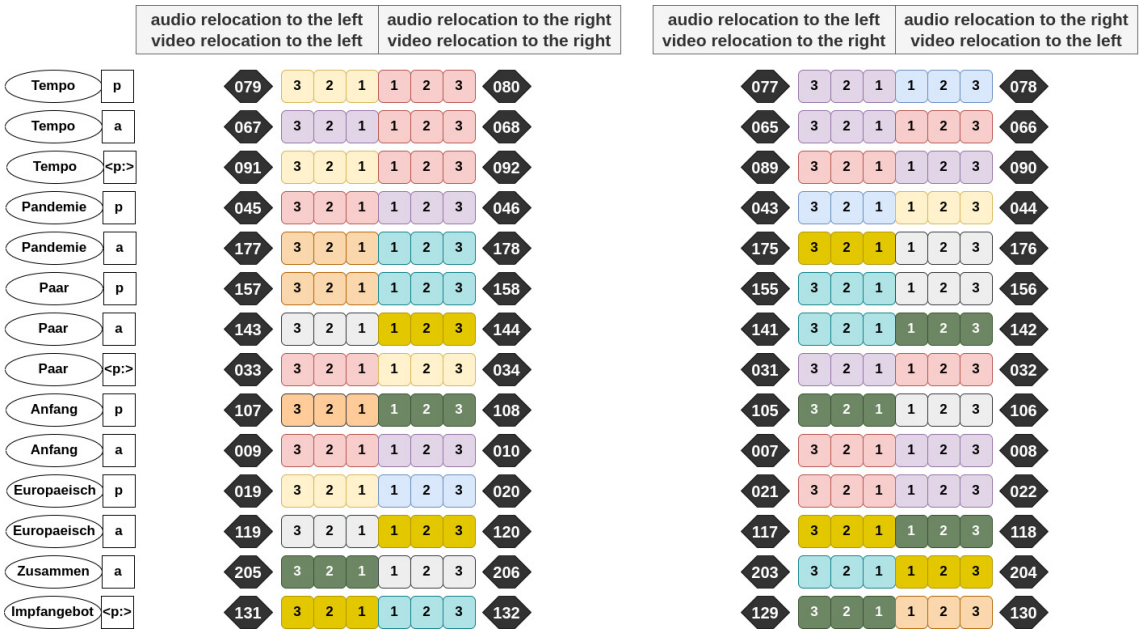


Figure 6.4.: The structure of the study sessions recognisable by colour: All trials containing audio and visual shifts combined.
Name of excerpt followed by the object (phoneme) being modified. On top is the applied type of modification noted while the numbers indicate the severity of modification. The hexagons show the internally used TrialID.

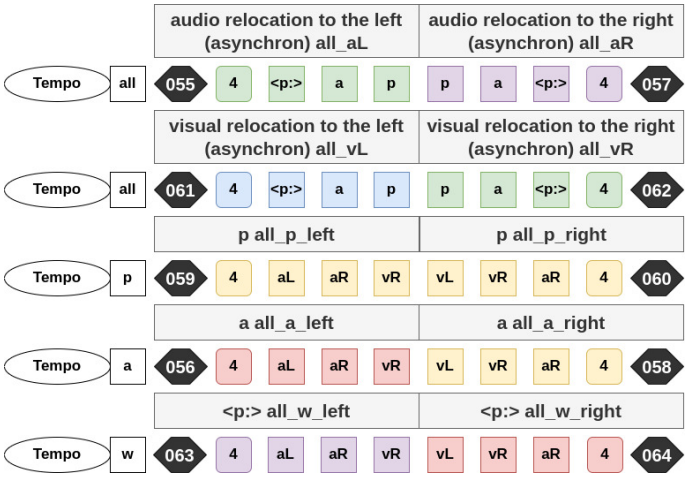
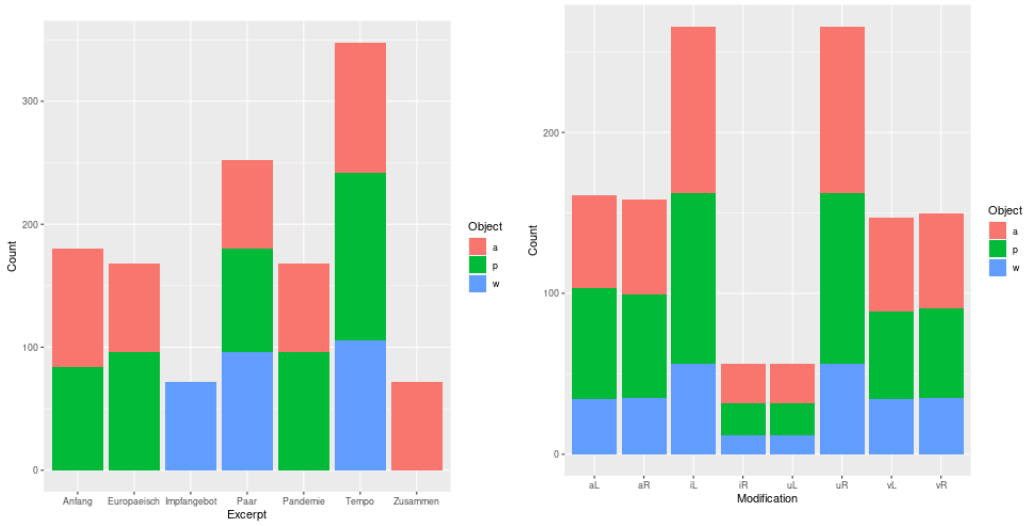


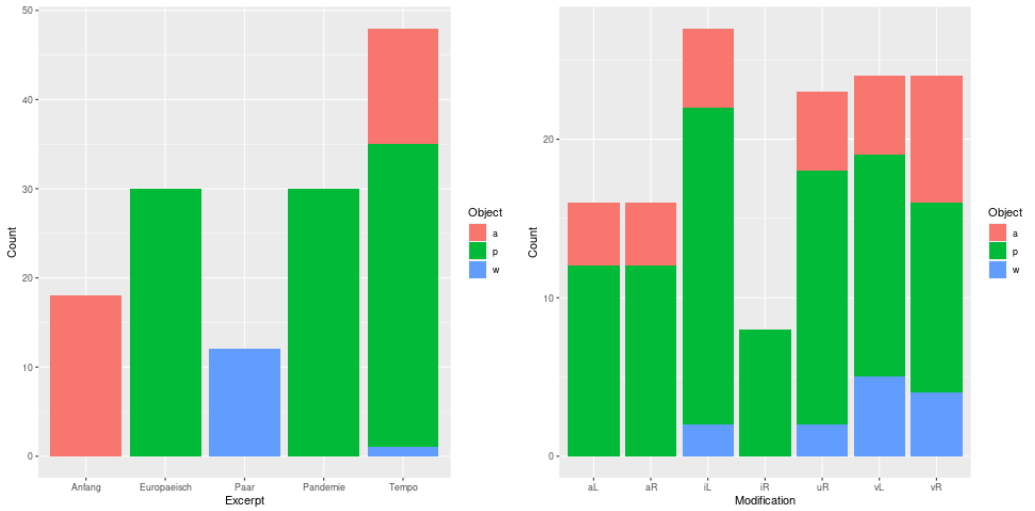
Figure 6.5.: The structure of the study sessions recognisable by colour: All trials containing mixed combinations of audio and visual shifts or phoneme occurrences inside.
Name of excerpt followed by the object (phoneme) being modified. On top is the applied type of modification noted while the numbers indicate the severity of modification. The hexagons show the internally used TrialID.



(a) Distribution of tested objects to excerpts.

(b) Distribution of tested objects to modifications.

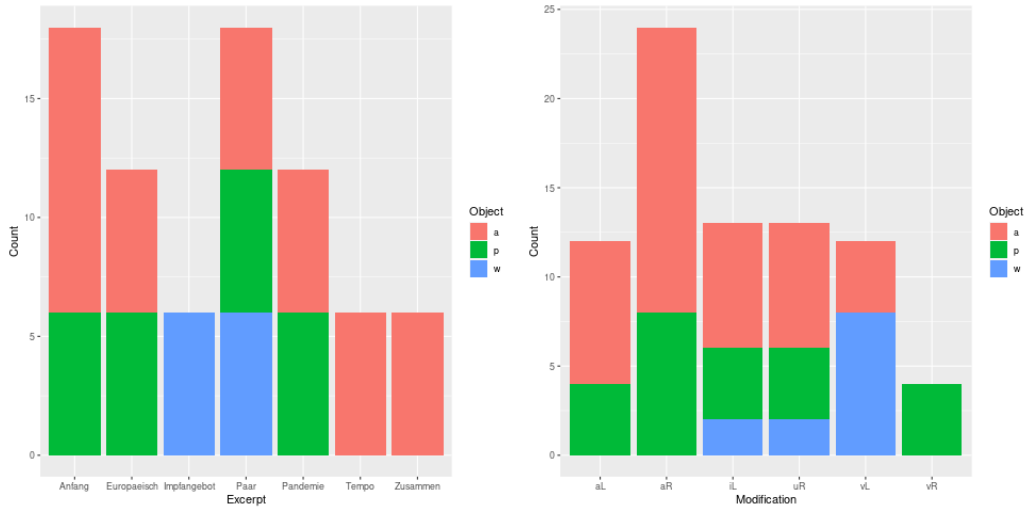
Figure 6.6.: Test material used in all parts of the study combined.



(a) Distribution of tested objects to excerpts, as included in study part 'Blue'.

(b) Distribution of tested objects to modifications, as included in study part 'Blue'.

Figure 6.7.: Test material used in the study part called 'Blue'.



(a) Distribution of tested objects to excerpts, as included in study part 'Pink'.

(b) Distribution of tested objects to modifications, as included in study part 'Pink'.

Figure 6.8.: Test material used in the study part called 'Pink'.

7. Experiments and Evaluation

This chapter examines the study and the observations that could be gathered. This will have more of a quantitative character while the next chapter will go into a more qualitative analysis of the data. For now, we will occupy ourselves with the initial cleaning of the data and the first exploratory data analysis. Based on these, further runs of the study were carried out to arrive at the final rating data. Finally, all the data had to be preprocessed to be ready for the statistical analysis.

7.1. Data Cleaning

Oriented towards the principles laid out by Wickham (2014), the preprocessing of the data was straightforward to implement. First, using a python script, the more than 1500 text files containing the subject's ratings in a JSON format were transformed into a single csv file, where each column was a variable name and each row a unique observation. Now every value in the table belonged to a clearly assigned variable and observation.

All the study runs resulted in 9234 single observations provided by 83 participants.

The names of the subjects were anonymised to protect their privacy. A runtime value of over 300000 ms (5 minutes) for a single trial, consisting of only 6 items to be rated, was an indication for a subject taking a longer break and therefore got reduced to 300000 for sake of reasonable plotting.

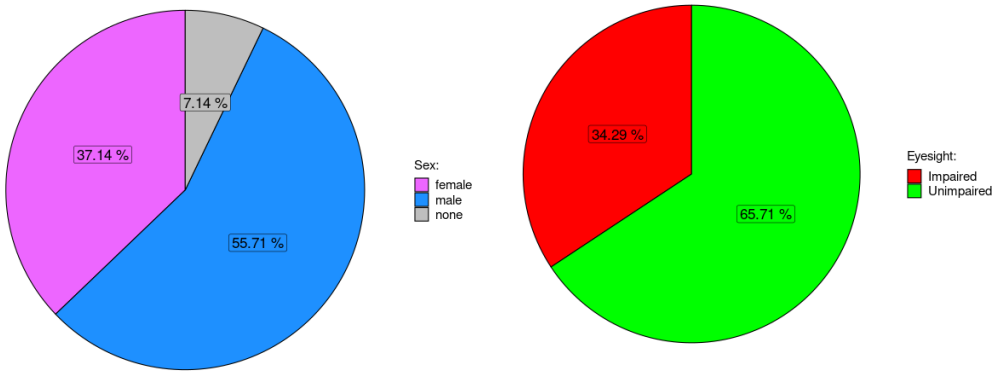
Next, the data was further reduced by dropping invalid results. Observations with a too low runtime got excluded. Since all used excerpts were of 3 or 4 seconds length, it is unlikely that a subject could properly evaluate all 6 video sequences of a trial in less than 15 seconds and rather just randomly clicked through, or skipped over some of them.

The remaining number of observations got reduced to 7908, provided by 76 participants.

If a trial contains the same grade 4 times, the subject was not able to distinguish differences between any of the anchors and at least one of the actual PVSs, or perceived all PVSs and at least one of the anchors to be the same. These were excluded during the data cleaning process.

This lowered the number of valid observations to 7494, and the number of participants to 70.

Part of the cleaning process was also to exclude all duplicates. These duplicates could have been caused by reasons like technical difficulties leading to a participant having to reload or repeat a started session which would lead to the server receiving multiple



(a) Distribution of participant's gender.

(b) Distribution of participants eyesight impairments.

Figure 7.1.: Participant's distribution of demographic information.

observations from the same participant for the same PVSs. This would then skew the following analysis of the ratings and had to be prevented.

The number of unique observations, considered in the final analysis was 6930, provided by 70 participants.

For the time being, the few noticeable outliers have been kept, since it was not obvious, what underlying reason could have caused them and removing outliers for pure convenience's sake was simply not an option.

7.2. Demographic Information

65 out of the 70 participants are native German speakers, 2 Bengali, 1 Hindi, 1 Portuguese and 1 Russian (see Table 7.2). Even though later analysis revealed some statistically significant differences based on the subject's native language, non-German subjects were not excluded from the analysis. This can be justified by the fact that these differences appear between many different languages and not primarily between German and others.

Of the 70 participants, 24 (34.29%) reported having some kind of, even if corrected, eyesight impairment, almost all of them being short-sighted as can be seen in Figure 7.1b.

No one reported impaired hearing.

The subject's interests can be seen in Table 7.3 and show a majority of natural sciences (67.14%) with just a few interested in social sciences (10%) and barely anything else. Naturally, this complicates if not even prevents investigations towards dubbing reception based on field of study, as done by Ameri et al. (2018).

The distribution of the subject's gender can be seen in Figure 7.1a and additionally grouped by age in Table 7.1.

Gender to Age	10 - 19	20 - 29	30 - 39	40 - 49	50 - 59	no info
female	3	17	6	0	0	0
male	3	24	6	4	2	0
no info	0	1	2	0	0	2

Table 7.1.: Distribution of participants gender and age.

Language	Count
German	65
Bengali	2
Hindi	1
Portuguese	1
Russian	1

Table 7.2.: Distribution of participants native language.

7.3. Data Distribution

In the study part called “Blue”, the subjects finished 13.32 out of 25 trials on average, contributing 79.94 out of 150 item ratings, which were saved in form of observations

In the study part called “Pink”, the subjects finished 8 out of 15 trials on average, contributing 48 out of 90 item ratings, which were saved in form of observations.

All, after the data cleaning process remaining subjects, further called the listening panel, of the combined parts of the study, finished on average 16.5 out of 210 trials, contributing 99 out of 1260 item ratings, saved in form of observations.

As already explained in Section 6.4, a subset of participants that were close to the rating tendency of the entire listening panel, went through all remaining parts of the study. This happened in the pursuit of detecting areas of interest, inside the remaining test data. These were then included in the last, in the remaining time still reasonably executable, study part, called “Pink” (for a depiction of this process see Figure 7.2).

As shown in Table A.7, a subset of the participants can represent the rating tendency of the whole group regarding the first part of the study (“Blue”), and Table A.8 shows the same conclusion for the complete study.

Since these few subjects finished a number of different sessions, they also had to rate the Anchor-Trials on multiple occasions. This can be used to not only analyse their intra-subject consistency but also to compare their rating tendencies for the items of these specific trials with that of the remaining listening panel as done in Figure 7.3.

7.4. Exploratory Data Analysis

Through visualisation (as seen in Figure 7.4a and Figure 7.4b), the expectation of a lower rating for an increased level of modification seems to hold true in general. However, in a few cases, something peculiar was observed: At certain places, an increase of modification

Identifier	Interest or study subject	Count
natur	Natural sciences (mathematics, computer science, physics, chemistry, geology, biology, ...)	47
geist	Humanities (history, language, literature, philosophy, ethics, religion, ...)	1
kunst	Art studies (music, art, theater, photography, art history, ...)	0
sozio	Social sciences (psychology, education, sociology, politics, law, economics, ...)	7
agrar	Agricultural sciences (agriculture and forestry, fishing, animal breeding, veterinary medicine, ...)	0
gesun	Health sciences (medicine, pharmacy, medical biotechnology, ...)	1
techn	Technical sciences (mechanical engineering, civil engineering, electrical engineering, ...)	1
nonw	No info	13

Table 7.3.: Distribution of participant's interest or study subject.

strength suddenly lead to better ratings, before, after a further increase of strength, going down again (as can be seen in Figure 7.5b.

In many cases, the distribution plots for grade and for ranks looked very similar to the ones in Figure 7.5. But in some other cases, the plotting of distributions resulted in different images and insights like in Figure 7.6 and Figure 7.7. The relatively high ratings in Figure 7.6a and Figure 7.7a, especially for the shifting of audio to the right (aR) may be caused by this modification barely being perceivable. This could be explained by the delay of audio being easier forgiven by the subjects than the delay of the video since sound waves travel slower than light and therefore the movements of a speaker get registered a tiny bit later than the sound which people are used to from the real world.

But once the graph shows the rating like in Figure 7.6b and Figure 7.7b, it becomes clear that, even though less harshly graded by the subjects, the modifications with their introduced artefacts, still lead to the corresponding PVS being of less, by the members of the listening panel perceivable, quality.

Additionally, this juxtaposition indicates that even though the tendency of subjects to give higher or lower grades in the range of 0 to 100 vastly varies, this does not affect the ranking of the compared PVSs so much as it blurs the differences/lines of perceived quality a bit.

The results for grades given by subjects can serve to find a threshold of acceptable-degradation of quality.

On the other hand, the results structured for ranks given by subjects can be used to find differences between the types of modification and their effect on the quality.

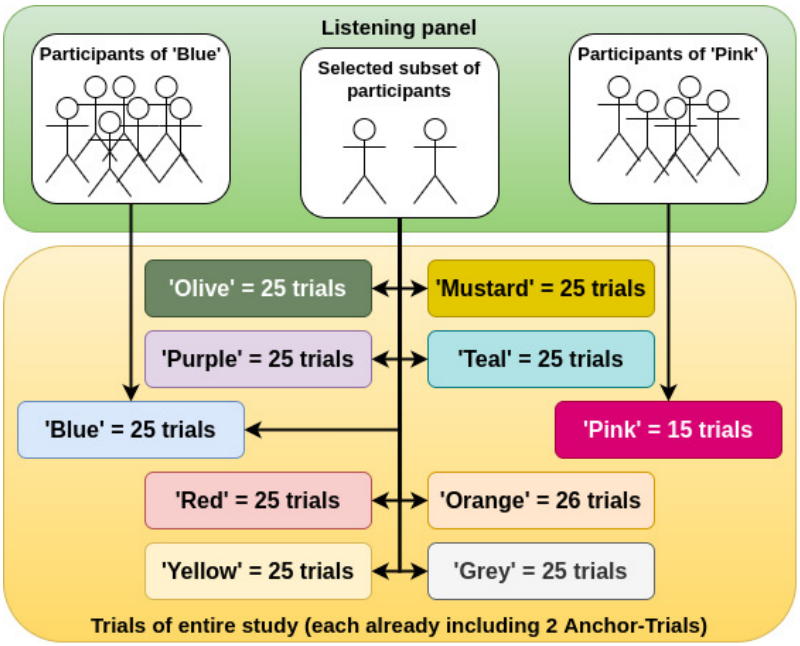
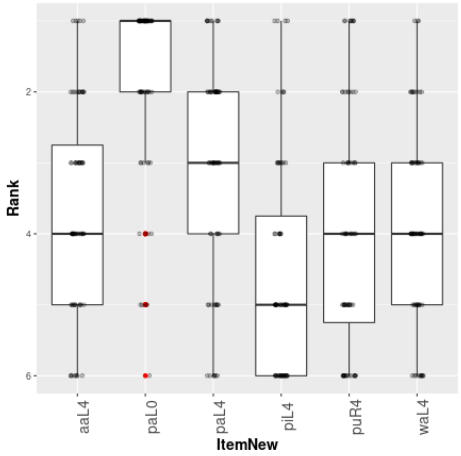
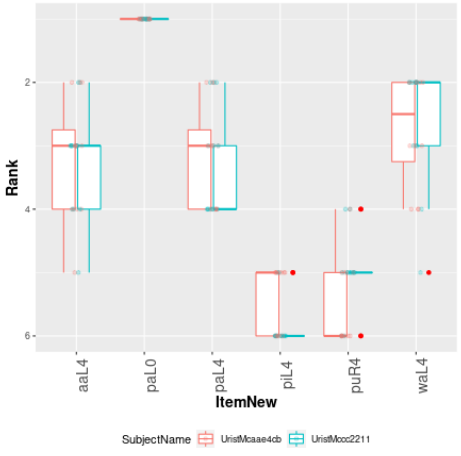


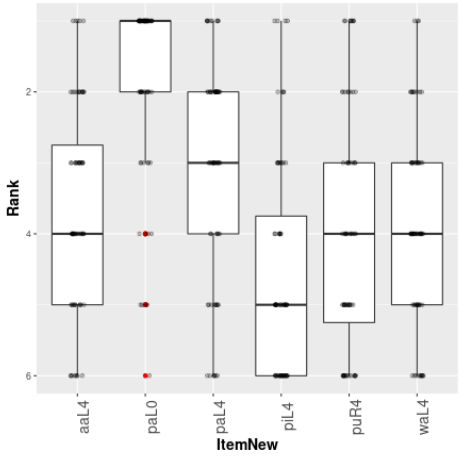
Figure 7.2.: Visualisation of the applied dynamic study design's eventual execution. On the top: the subsets of participants over the course of the entire study. At the bottom: all sessions of the entire study and the number of included trials.



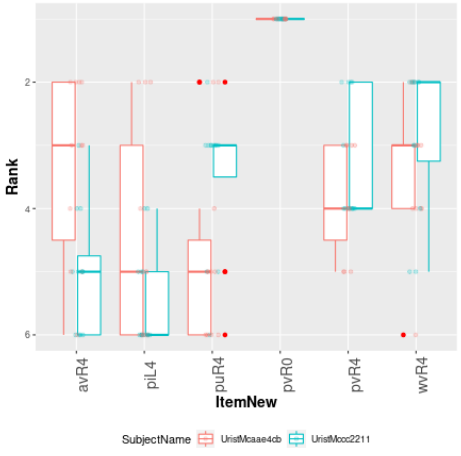
(a) Ranks per item for Anchor-Trial 055 by listening panel.



(b) Ranks per item for Anchor-Trial 055 by two subjects.

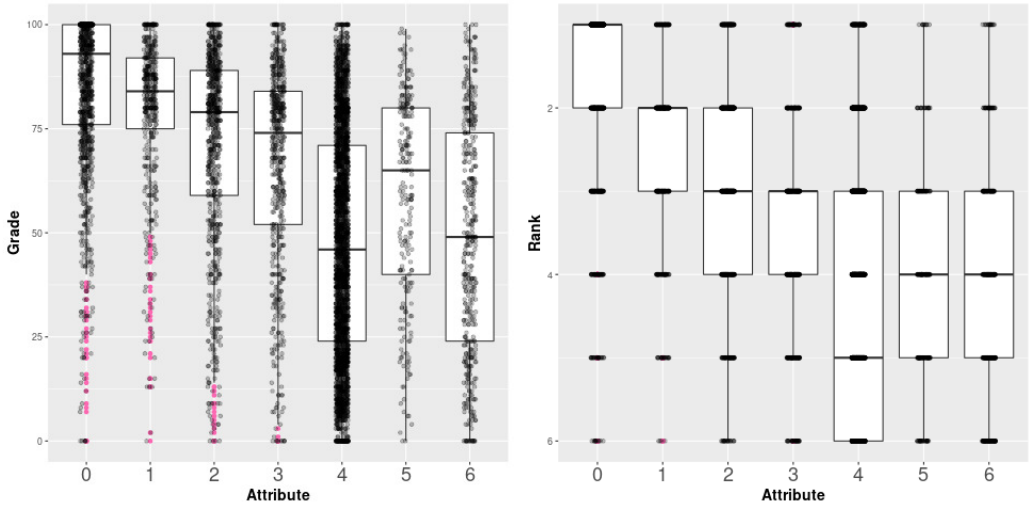


(c) Ranks per item for Anchor-Trial 062 by listening panel.



(d) Ranks per item for Anchor-Trial 062 by two subjects.

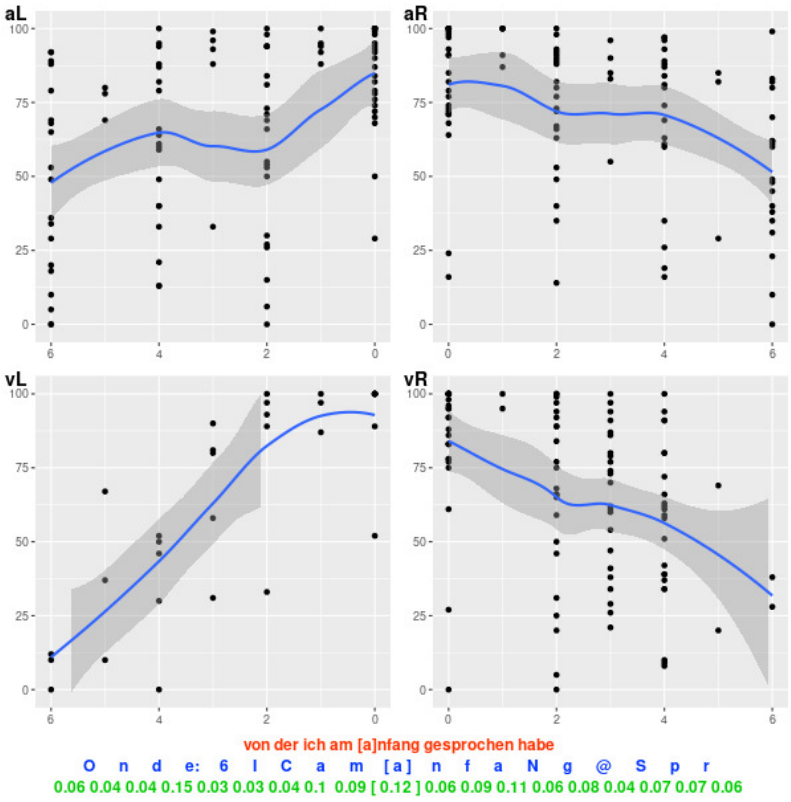
Figure 7.3.: Rating tendencies of the listening panel for the two Anchor-Trials.



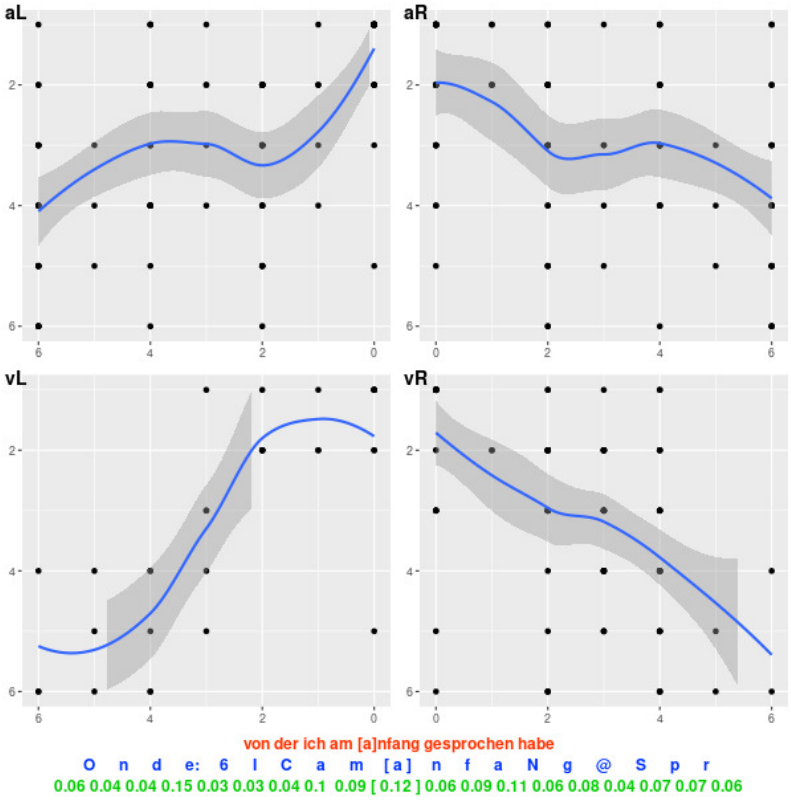
(a) Grades per attribute.

(b) Ranks per attribute.

Figure 7.4.: Rating grouped by attribute (strength of modification) of listening panel.

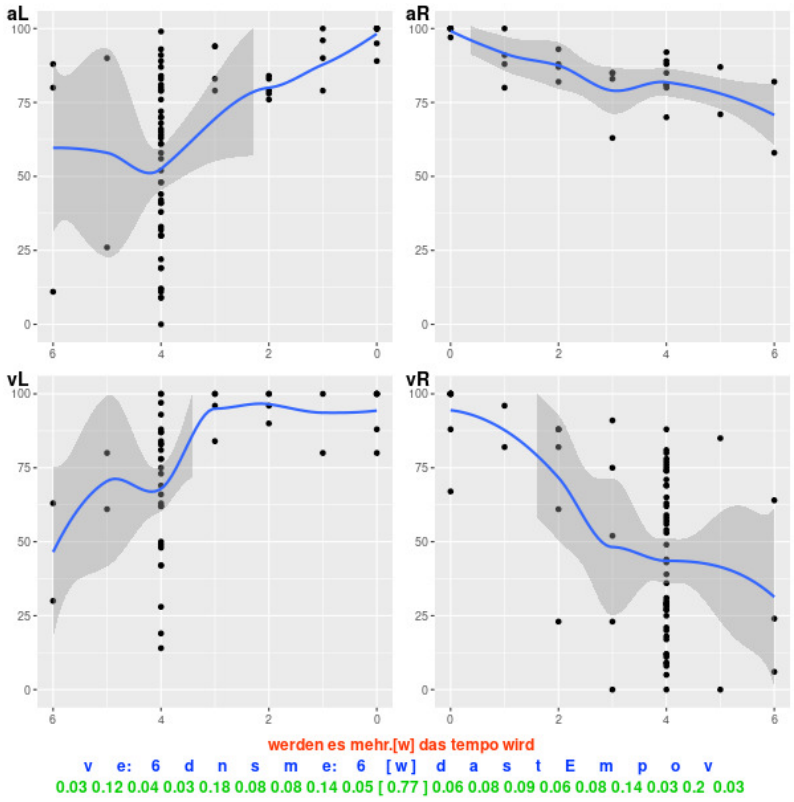


(a) Anfang-a for grade.

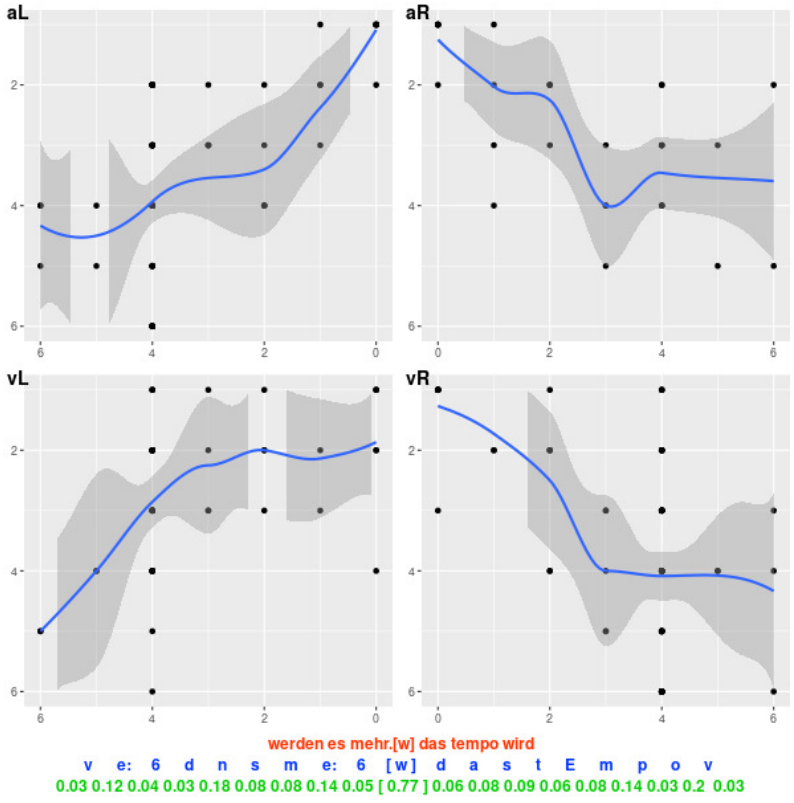


(b) Anfang-a for rank.

Figure 7.5.: Distribution of ratings (black dots) for Excerpt Anfang with phoneme a. aL: audio shifted to the left; aR: audio shifted to the right; vL: visual shifted to the left, vR: visual shifted to the right.

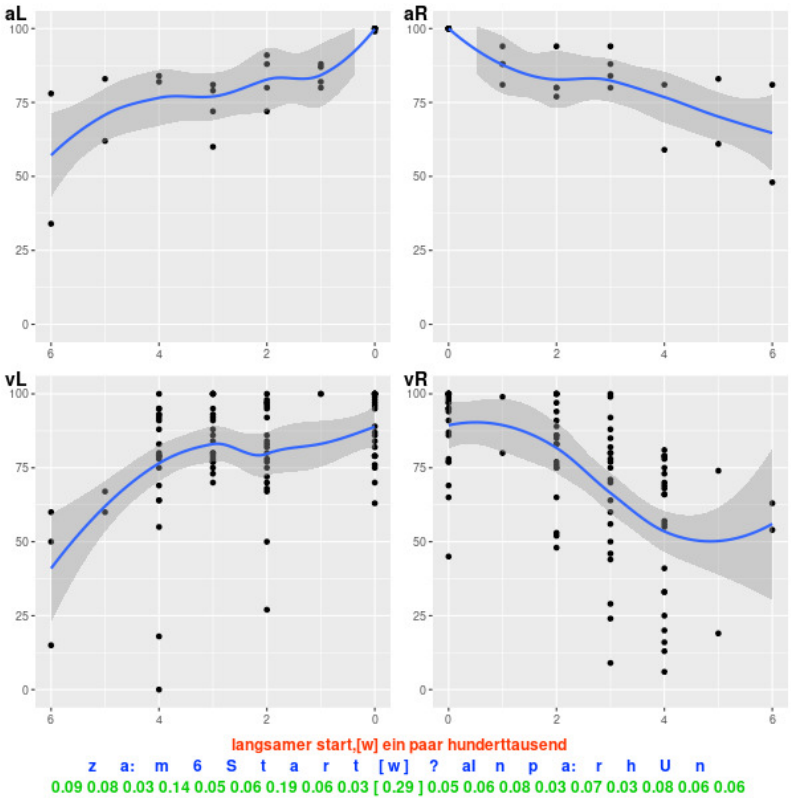


(a) Tempo-w for grade.

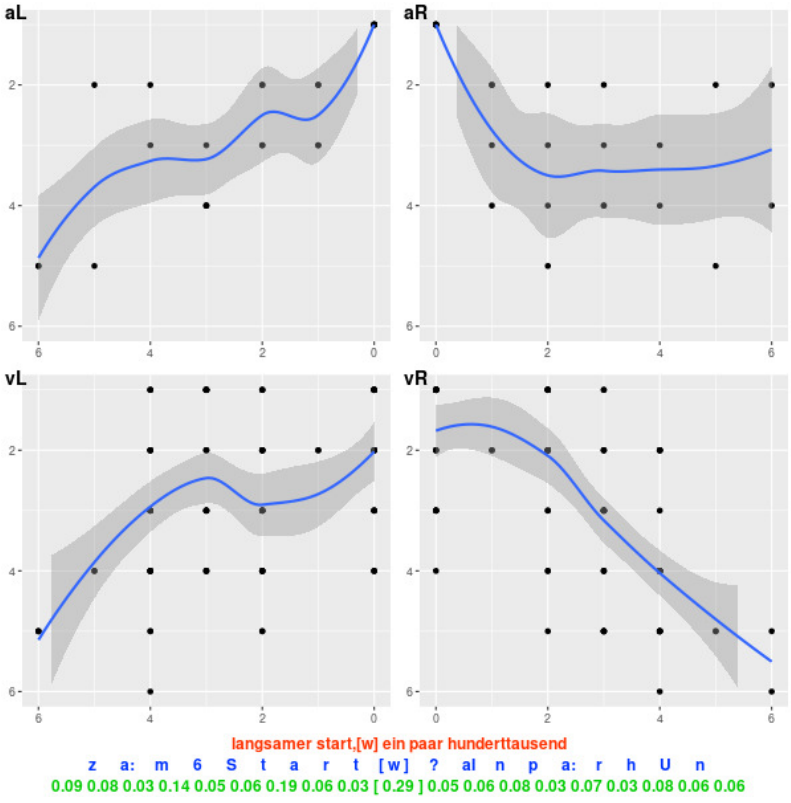


(b) Tempo-w for rank.

Figure 7.6.: Distribution of ratings (black dots) for Excerpt Tempo with a short pause (depicted as w in the plot).
aL: audio shifted to the left; aR: audio shifted to the right;
vL: visual shifted to the left, vR: visual shifted to the right.



(a) Paar-w for grade.



(b) Paar-w for rank.

Figure 7.7.: Distribution of ratings (black dots) for Excerpt Paar with a short pause (depicted as w in the plot).
aL: audio shifted to the left; aR: audio shifted to the right;
vL: visual shifted to the left, vR: visual shifted to the right.

8. Results and Discussion

For the sake of brevity and in sight of this work's limitations and exploratory nature, a subset of observations and implications will be showcased in the following sections. Additional data and results can be reviewed in the following Appendix and more will be made available online[11].

8.1. Data Analysis

As an entry point to the data analysis, we examine the hidden anchors' performances. If, for example, most of the subjects would have rated the High-Anchor, which is the unmodified reference item, as having a very low quality, then this would indicate a critical problem for the entire experiment. The same is true for the other anchors, and to a certain degree, for the remaining items. Although some deviation in the item ranking, could just be an indicator for possible effects as investigated in this work.

As can be seen in Figure 8.1, the subjects were able to rank the High-Anchor (HA) in 65.38% of the trials as the highest item and recognised the Low-Anchor (LA) in 57.76% of the trials as the item with the lowest quality. Also, the Middle-Anchor (MA) is predominantly in the middle-field to lower end of the ranking scale, which corresponds with the expected outcome based on the applied modifications in the editing process. Not only the ranking for the anchors, but also for the actual items under test correspond to the degree with which they have been modified, at least when it is generalised over all instances. Therefore an item only modified with an attribute value of 1 (corresponding to a shift of a single frame, or 0.04 seconds), tends to be ranked higher, while higher valued attributes result in a tendency to be ranked lower. This observation indicates, that the subjects of the study were able to differentiate between even the most minute of differences in quality, based on their perception. For more details and exact values, refer to Table 8.1.

8.2. ANOVA

Datanovia provides an excellent introduction and tutorial[12] to the ANOVA statistical analysis which is quoted below in lack of better words from this author: "The ANOVA test makes the following assumptions about the data:

[11] https://github.com/christianschuler8989/dubbing-quality-lipsync
[12] https://www.datanovia.com/en/lessons/anova-in-r/

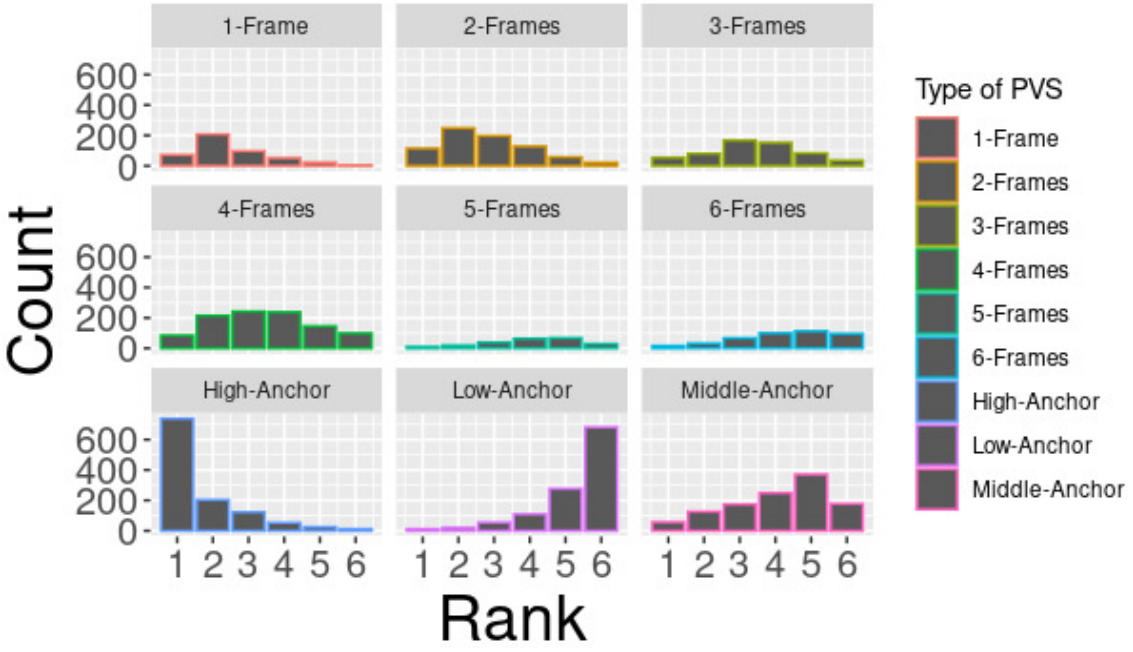


Figure 8.1.: Distribution of ranking done by listening panel.

1.: Independence of the observations. Each subject should belong to only one group. There is no relationship between the observations in each group. Having repeated measures for the same participants is not allowed.

2.: No significant outliers in any cell of the design

3.: Normality. the data for each design cell should be approximately normally distributed.

4.: Homogeneity of variances. The variance of the outcome variable should be equal in every cell of the design."

First tests done on the from the study received data, revealed, that no ANOVA could be applied to it. Here, for example, shown in regard to grades, the applied Shapiro-Wilk test of normality (statistic = 0.910 and p-value = 8.51^{-44}) confirmed, that the normality assumption can mostly not be met. This is visualised in the plot of the model's residuals in Figure 8.2a, as also done separately for each group in Figure 8.2b.

8.3. Kruskal-Wallis Test

In the spirit of an exploratory study, the statistical analysis tests have been run for various levels of granularity. Additionally, this was in parallel done once for the raw grade value given by the subjects and once for the rank, determined by the order of preference inside of each trial.

For Objects (the phonemes a and p but also w standing for pauses): There were statistically significant differences between object groups regarding the grades as assessed

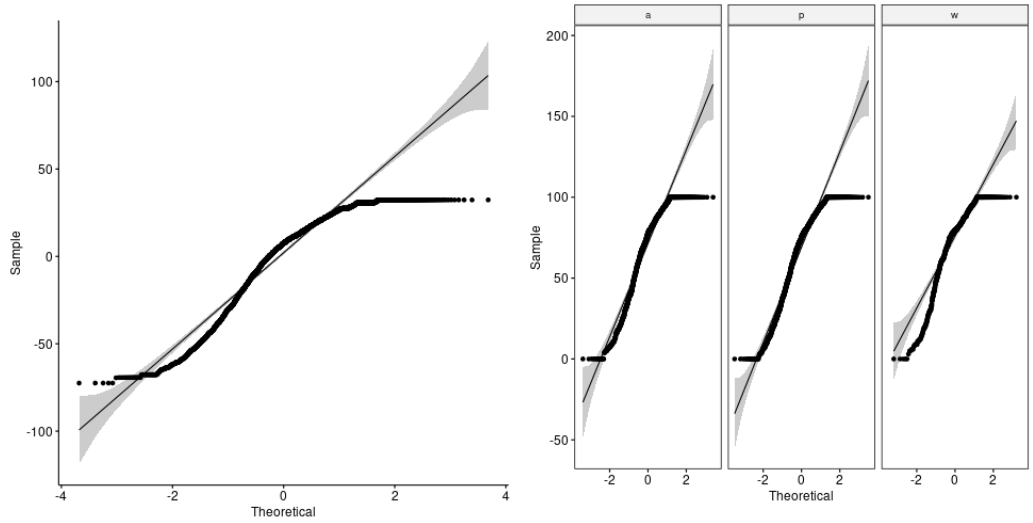
Rank	Count	Percent	PVS	Rank	Count	Percent	PVS	Rank	Count	Percent	PVS
1	738	63.9%	HA	1	72	15.89%	1-Frame	1	85	8.3%	4-Frames
2	204	17.66%	HA	2	207	45.7%	1-Frame	2	215	21%	4-Frames
3	123	10.65%	HA	3	97	21.41%	1-Frame	3	242	23.63%	4-Frames
4	53	4.59%	HA	4	51	11.26%	1-Frame	4	238	23.24%	4-Frames
5	25	2.16%	HA	5	22	4.86%	1-Frame	5	145	14.16%	4-Frames
6	12	1.04%	HA	6	4	0.88%	1-Frame	6	99	9.67%	4-Frames
1	57	4.94%	MA	1	114	14.9%	2-Frames	1	10	4.37%	5-Frames
2	126	10.91%	MA	2	248	32.42%	2-Frames	2	19	8.3%	5-Frames
3	174	15.06%	MA	3	197	25.75%	2-Frames	3	39	17.03%	5-Frames
4	249	21.56%	MA	4	128	16.73%	2-Frames	4	64	27.95%	5-Frames
5	371	32.12%	MA	5	57	7.45%	2-Frames	5	68	29.69%	5-Frames
6	178	15.41%	MA	6	21	2.75%	2-Frames	6	29	12.66%	5-Frames
1	102	1.04%	LA	1	53	9.3%	3-Frames	1	16	3.77%	6-Frames
2	20	1.73%	LA	2	80	14.04%	3-Frames	2	32	7.55%	6-Frames
3	55	4.76%	LA	3	167	29.3%	3-Frames	3	67	15.8%	6-Frames
4	109	9.44%	LA	4	152	26.67%	3-Frames	4	100	23.58%	6-Frames
5	277	23.98%	LA	5	83	14.56%	3-Frames	5	113	26.65%	6-Frames
6	682	59.05%	LA	6	35	6.14%	3-Frames	6	96	22.64%	6-Frames

Table 8.1.: Distribution of ranking done by listening panel for different PVSs.
HA: High-Anchor, MA: Middle-Anchor, LA: Low-Anchor

using the Kruskal-Wallis test ($p = 0.0000735$). Pairwise Wilcoxon test between groups (see Figure 8.3) showed that the difference was significant between a and p ($p = 0.042$), and between p and w ($p = 0.0000576$). Although the Kruskal-Wallis test ($p = 0.0305$) indicated a statistically significant difference between object groups regarding the ranks, neither a pairwise Wilcoxon test nor a pairwise Dunn test showed any significance.

For Excerpts (the used video sequences by name): There were statistically significant differences between excerpt groups regarding the grades as assessed using the Kruskal-Wallis test ($p = 1.53^{-21}$). Pairwise Wilcoxon test between groups (see Figure 8.4) showed that the difference was significant between “Anfang” and “Paar” ($p = 0.000273$), between “Anfang” and “Tempo” ($p = 0.001$), between “Europaeisch” and “Paar” ($p = 0.00000758$), between “Europaeisch” and “Tempo” ($p = 0.003$), between “Impfangebot” and “Pandemie” ($p = 0.013$), between “Impfangebot” and “Tempo” ($p = 0.0000279$), between “Paar” and “Pandemie” ($p = 2.02^{-8}$), between “Paar” and “Tempo” ($p = 7.79^{-20}$), and also between “Tempo” and “Zusammen” ($p = 0.001$). There was also a statistically significant difference between object groups regarding the ranks as assessed using the Kruskal-Wallis test ($p = 0.0379$). But pairwise Wilcoxon test between groups could only reveal a significant difference between “Paar” and “Pandemie” ($p = 0.037$).

For Modifications (the types of editing like audio, visual and combinations): There were statistically significant differences between modification groups regarding the grades as assessed using the Kruskal-Wallis test ($p = 2.31^{-15}$). Pairwise Wilcoxon test between groups (see Figure 8.5) showed that the difference was significant between aL and vR ($p = 0.00000941$), between aR and vR ($p = 6.46^{-12}$), between iR and vR ($p = 0.0000112$), between uL and vR ($p = 0.00000159$), an between vL and vR ($p = 3.74^{-9}$). There were no statistically significant differences between modification groups regarding the ranks as



(a) Check normality assumption by analyzing the model residuals for Grades.

(b) Check normality assumption by analyzing the model residuals of each group for Grades.

Figure 8.2.: Participant's distribution of demographic information.

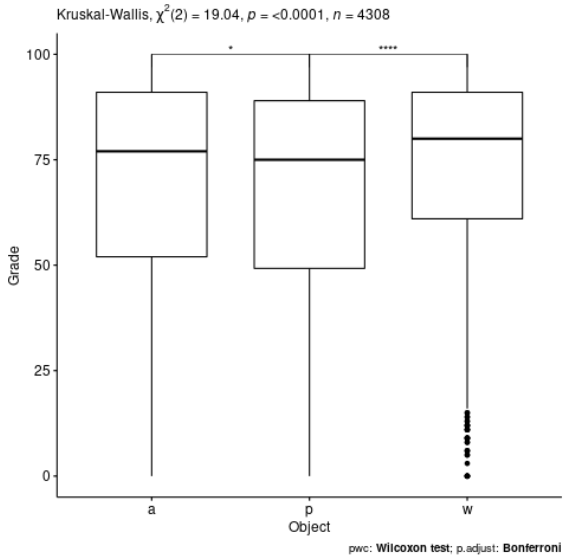


Figure 8.3.: Statistical analysis of the objects, being phonemes a and p, and w standing for pauses.

assessed using the Kruskal-Wallis test (p = 0.0841).

For Attributes (the severity of introduced change in frames, or 0.04 seconds): As to be expected, the difference between attribute groups was statistically significant in regard to grade ($p = 1.37^{-211}$) and also for ranks ($p = 4.14^{-304}$). Interestingly, however, were those groups showing the lowest p-value for significance: For the grades given by subjects, these were between 2 and 3 (p = 0.054), 4 and 5 (p = 1.0), and 5 and 6 (p = 0.000104) and for ranks 1 and 2 (p = 0.061), 3 and 4 (p = 1.0), and 5 and 6 (p = 0.714), again, using the Kruskal-Wallis test (see Figure8.6). These observations can reasonably be brought back

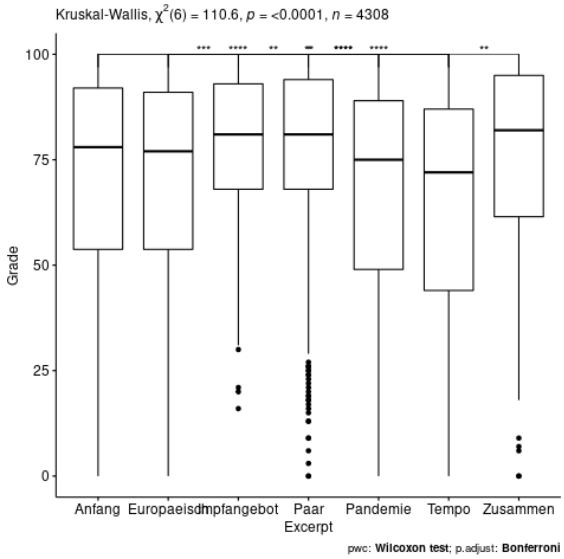


Figure 8.4.: Statistical analysis of the excerpts, being the different video sequences.

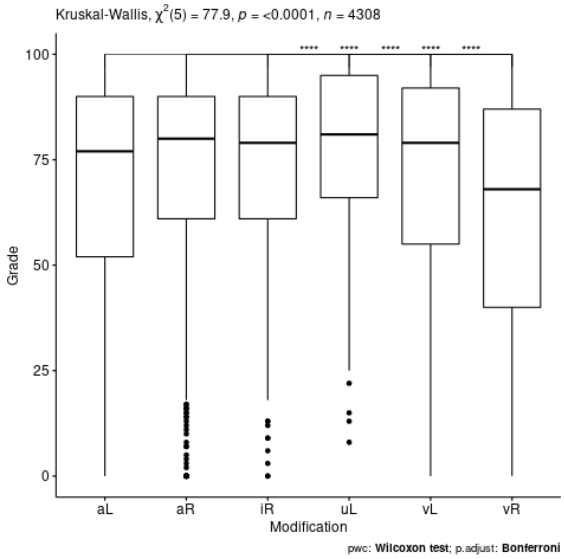


Figure 8.5.: Statistical analysis of the modifications, being aL and aR for audio shifted left and right, vL and vR for visual shifted left and right, uL and uR for audio and visual shifted in the same direction, and iL and iR for audio and visual shifted in opposite directions.

to the setting of the study itself. A large part of trials contained either modifications of attribute size 2, 4 and 6 or 1, 3 and 5 (or, less frequently, 1, 2 and 3 or 2, 3 and 4), but never 4 and 5 or 5 and 6 together.

For subject's gender (as stated by the subjects to be male, female or none) There were statistically significant differences between groups of subject's sex regarding the grades as assessed using the Kruskal-Wallis test ($p = 4.7^{-29}$), but none in regard to ranks (p =

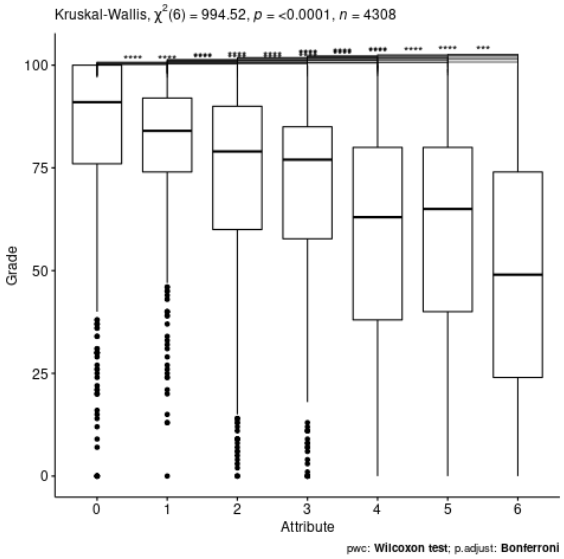


Figure 8.6.: Statistical analysis of the attributes, indicating how strong the modification of a video sequence was.

0.323). Pairwise Wilcoxon test between groups (see Figure 8.7) showed that the difference was significant between female and male ($p = 1.77^{-28}$), and between male and none ($p = 0.000354$), and also, but to a lesser degree, between female and none ($p = 0.047$), which was even not significant for the Dunn test ($p = 0.182$). Although the Kruskal-Wallis test ($p = 0.0305$) indicated a statistically significant difference between object groups regarding the ranks, neither a pairwise Wilcoxon test nor a pairwise Dunn test showed any significance.

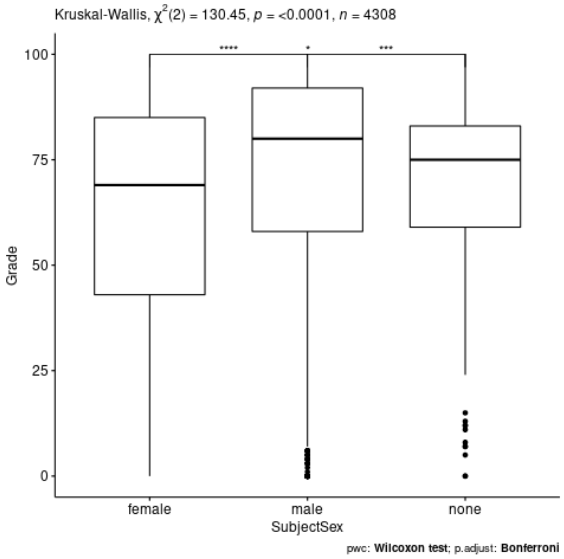


Figure 8.7.: Statistical analysis of the subject's sex.

For subject's interests (like fields of study as explained in Table 7.3) There were statistically significant differences between groups of subject's interests regarding the grades

as assessed using the Kruskal-Wallis test ($p = 6.62^{-80}$), and also in regard to ranks ($p = 0.0147$). Pairwise Wilcoxon test between groups (see Figure 8.8) showed that the difference was significant between "geist" and "gesun" ($p = 6.09e\text{-}4$), between "geist" and "natur" ($p = 0.00000408$), between "gesun" and "natur" ($p = 0.00000222$), between "gesun" and "none" ($p = 0.019$), between "gesun" and "sozio" ($p = 0.001$), between "gesun" and "techn" ($p = 0.00016$), between "natur" and "none" ($p = 4.84^{-73}$), between "natur" and "sozio" ($p = 6.99^{-9}$), between "none" and "sozio" ($p = 0.001$), and between "none" and "techn" ($p = 0.000214$).For ranks the pairwise Wilcoxon test only showed a significant difference between "natur" and "none" with ($p = 0.041$).

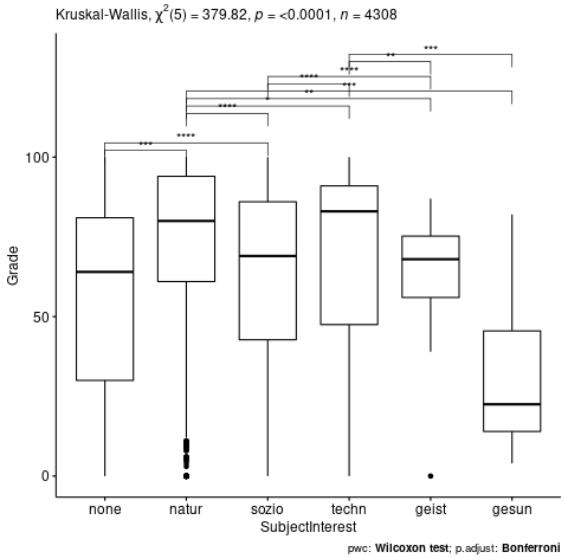


Figure 8.8.: Statistical analysis of the subject's interests.

For subject's eyesight impairments There were statistically significant differences between groups (see Figure 8.9) of subject's eyesight regarding the grades as assessed using the Kruskal-Wallis test ($p = 4.74^{-23}$), but none in regard to ranks ($p = 0.446$).

For general types of modification There were statistically significant differences between types of modification groups regarding the grades as assessed using the Kruskal-Wallis test ($p = 8.11^{-7}$), but none in regard to ranks ($p = 0.112$). Pairwise Wilcoxon test between groups (see Figure 8.10) showed that the difference was significant between audio and visual ($p = 0.000304$), between mixed and visual ($p = 0.012$), and between multi and visual ($p = 0.000443$). This was to be expected since the methods used in this work introduced considerable additional artefacts while editing audio segments, while the modification of visual aspects turned out to be comparatively smooth. However, a look at the summary statistics Table 8.2 reveals a lower median and mean for visual compared to all other types of modification, which is the opposite of the author's expectation. At the very least, this indicates, how great the importance of chosen types for modification are while investigating and doing experiments concerning perceived lip-synchrony. This could also mean, that

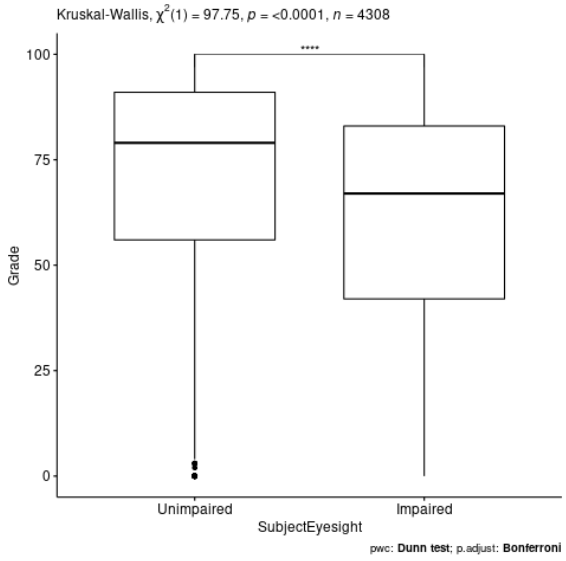


Figure 8.9.: Statistical analysis of the subject's eyesight.

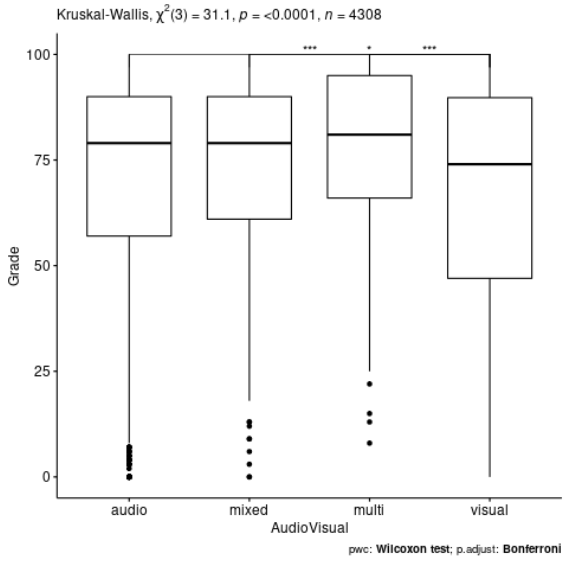


Figure 8.10.: Statistical analysis of different types of modification, being audio and visual, but also multi for audio and visual in the same direction and mix for audio and visual in opposite directions.

the visual modifications were less smooth than perceived by the author and as pre-study experiments indicated.

8.4. Discussion of the Findings

From the outset this work was of an exploratory nature, as made necessary by gaps in the research of AVT reception. After exploring, which aspects to investigate, would be the

AudioVisual	variable	n	min	max	median	iqr	mean	sd	se	ci
audio	Grade	2002	0	100	79	33	70.5	25.7	0.574	1.13
mixed	Grade	252	0	100	79	29	72.8	23.8	1.50	2.95
multi	Grade	132	8	100	81	29	76.7	21.7	1.89	3.74
visual	Grade	1922	0	100	74	42.8	66.6	27.6	0.63	1.24

Table 8.2.: Summary Statistics for general types of modification's grades.

most promising one and executing experiments, including a human viewer based study and statistical data analysis, a vast number of observations have gathered. This collection of data, made available online[13] for future investigations, provides a number of possible approaches.

Other than in a dubbing reception study with 477 Iranian adults (207 female), where it was concluded that "All in all, it is safe to claim that dubbing reception was not affected by the factor of gender." (Ameri et al., 2018, p.10), the analysis of this work indicates a statistically significant difference in regard of grading (compare Figure 8.7) with a harsher grading by female subjects than by the other two subject groups, male and none (see Table 8.3).

SubjectSex	variable	n	min	max	median	iqr	mean	sd	se	ci
female	Grade	1248	0	100	69	42	62.7	27.3	0.773	1.52
male	Grade	2832	0	100	80	34	71.9	25.9	0.487	0.955
none	Grade	228	0	100	75	24	68.7	22.0	1.46	2.87

Table 8.3.: Summary Statistics for grades based on subject's gender.

Though the study's instructions (as shown in Figure 6.1b) clearly stated, that study participants should base their rating on perceived lip-synchrony, a lot of the received feedback, verbally and through comments, indicates otherwise. Subjects repeatedly reported, that introduced audio artefacts, as side effect of the modifications, would make it really easy to decide, which of the test items would be the ones that have been modified the most. Contrary to the subjects' self-assessments, the results clearly show worse ratings for visual modifications than for the audio modifications (as previously shown in Table 8.2). Especially intriguing is the observation that the visual modifications ended up with an even lower ratings' mean than the combined modification type, which, as previously laid out, introduces much stronger disturbances of lip-synchrony. In conjunction with this is the fact, that subjects with unimpaired eyesight tended to rate higher than visually impaired subjects. Assuming the visual modifications actually were the most damaging, wouldn't it be fair to expect this ratio to be the opposite?

On the other hand, it is possible that many of the subjects, having reported to have no impaired eyesight, actually would benefit by using a visual aid (e.g. glasses), while the others took means to correct their known impairments, by using glasses and consequently have a better ability to detect visual inconsistencies.

[13] https://github.com/christianschuler8989/dubbing-quality-lipsync

Since almost all of the participants were German native speakers, the results of this study could be caused by their prior consumption of dubbed material and a habituation to impaired lip-synchrony, as explored by (Romero-Fresco, 2019).

Solely based on the presented results it can be concluded, that lip-synchrony has less of an effect on the perceived video quality, than has initially been suspected by the author of this work. Keep in mind, that the study was carried out without supervising the participants. To name just a few of the causes, results may be due to usage of screens with varying sizes, different headphones or surrounding conditions like a noisy background.

8.5. Challenges and Limitations

The chosen dataset can be seen as a blessing and as a curse in regards of limitation. Considering the constraints of this work, regarding time, resources, and expertise make it abundantly clear that only a few select aspects of lip-synchrony, which in itself is just one of the many aspects of dubbing, could be investigated. Using video material testing only a single speaker talking in merely one language meant that many aspects of lip-synchrony could not be included in the experiments. However, these limitations would prove crippling for future endeavours in investigating similar questions.

The limitation of only working with video material of 25 fps meant that all selected scenes had to be at a time where the person on screen was moving as little as possible. Because even slight gestures like the tilt of the head would turn into very obvious jumps in the video sequence after just removing one or two frames. So to be able to edit the visual representation of specific phonemes in the video sequence required a certain type of material. This, in turn, lead to a long and tedious selection process, that could only partially be supported by automating the code of the author.

Irrespective of the prior discussed limitations, video editing in pursuit of affecting lip-synchrony proved to be a herculean challenge in itself. As previously examined in Chapter 5, achieving modifications of just the audio, or just the visual of a video sequence can be accomplished with relatively little effort. There exist many tools available for free, exactly for these processes. But introducing a modification of just a short section, like a single phoneme of around 0.06 seconds length, without disrupting the entire video sequence, is something, that can not be solved by the author for the lack of out-of-the-box solutions. This work's approach took a long time to conceptualise and successfully implement, based on the author's prior expertise and knowledge. And still the introduction of undesired artefacts, especially for sound modifications, could not be prevented. Following this, the experiments and the final analysis require considerations in this regard.

This work's limited time frame in combination with its exploratory character, introduced the challenge of cultivating a manageable scope. There had to be a balance between merely exploring the bare minimum of a single phoneme with a few modifications on the one hand, and including many, if not all of the phoneme groups and a vast number of modifications

and many more variables like the subjects' demographic information. The first would be easy to handle but also risk not leading to any useful observations and insights. The latter promises to uncover many interesting aspects, but quickly bury the investigation in a sea of data.

9. Future Work and Conclusion

Every insight has been shown by now, thus the research questions can be answered. Before we have a look at them, this chapter lays out the contributions to future research that have been made during this work. This mainly includes the testing framework. Besides that, an outlook for further research is presented.

9.1. Useful Contributions

The usage of an easily online accessible testing framework enabled the execution of a human viewer based perception study in the midst of a global pandemic. The used and adjusted framework can easily be used in future advances and even be modified with relative ease.

The design of a flexible study, structured to be adjusted to the number of participants and the use of anchoring test items inside of trials and anchor trials inside each session, made this work less dependent on the participant's count being very large. Similar approaches suffer gravely from low participation numbers. Considering the voluntary and uncompensated nature of this work, and to be expected difficulties in acquiring a high number of participants, made these prior considerations pay out in the end.

As mentioned and partially laid out, the fact of this field of research, lip-synchrony in dubbing, perception of AVT in general, being so fractured, can serve as another angle for future experiments and investigations. Possibly a collaboratively approach combining multiple disciplines could path the way to more connected advances, new insights and maybe even applicable norms and conventions in scientific research and practical industry alike.

9.2. Outlook

For an exhaustive investigation, it would be prudent to include a finer level of detail and also a broader array of dimensions (see Figure 9.1) in the testing. A finer level of detail by not only including a single open vowel and a single bilabial consonant but different examples of the same phoneme group. And with a higher fps it could be feasible to modify shorter than 0.04 second long shifts, or even edit video material with more movement and action than was used in this work. A broader array of dimensions refers to including more of the phoneme groups than just the two in this work. Also, the lengths of the used phonemes and the type and length of their neighbours could lead to interesting insights as

much as possible additions from the subject's side and if and how personal criteria of the participants affect the ratings.

A possible layout for future investigation, ideally with considerably more participants, could resemble the following:

A: Different cultural groups from (British, Scottish, American, ...)
B: Different language groups (English, Chinese, German, ...)
C: Different phoneme-groups (bilabial consonants, open vowels, ...)
D: Different (reasonable) combinations of neighbouring phonemes (pap, pam, ...)
E: Different types of cinematic shot (extreme close-up, close-up, ...)
F: Different angles of the speakers face (frontal, profile, ...)
G: Different speaker (male, female, young, old, ...)
H: Different kinetic circumstances of the speaker (sitting still or running)
I: Different genres (documentary, action movie, drama series, ...)
J: Different forms of media consumption (cinema, tv-screen, tablet, smartphone, ...)

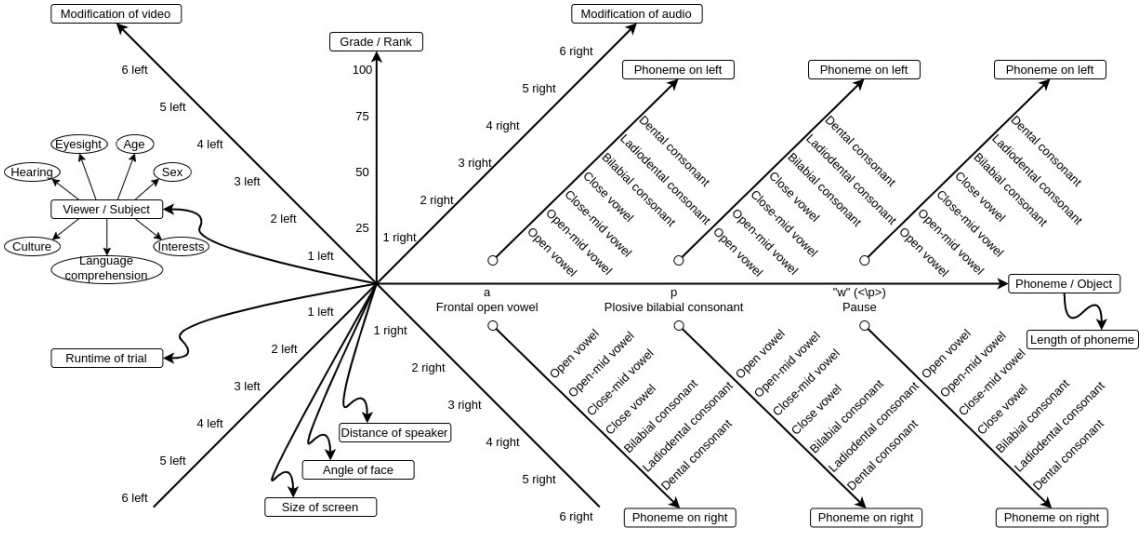


Figure 9.1.: Dimensions of analysis.

9.3. Summary

After the introduction, we familiarised ourselves with the topics and specific terms important for work in the field of AVT reception research. Chapter 2 also gave insights into the intricacies of lip-synchrony, how its quality can be assessed and how it is situated in audiovisual translations. In Chapter 3, we saw the history of AVT reception research up to recent advances laid out. Shifts in new directions, like the viewer moving more into the focus of the investigation. But also many gaps, that are only slowly being addressed, likely caused by the field of AVT research being as fractured as it is, and major differences in preferred AVT modes between many language groups. Chapter 4 provided general

considerations and methodologies for quality assessments. This was followed by details of subjective quality assessment, using human-based studies and Crowd-based approaches. Closing with a short outlook on managing tidy data and statistical analysis. As part of Chapter 5, and in agreement with the required attributes of video material, the Merkel Corpus was presented. Based on this the creation of suitable test material was exemplified. All of this is guided by initial exploratory pre-testing. In Chapter 6 a framework for Crowd-based testing with a finishing setup was presented. Followed by the conceptualisation and creation of a very flexible study. Closing with an overview of the material included in the study. Tying into the study's completion, Chapter 7 covered the initial cleaning of the received data in form of participants observations. The demographic information of the participants and the distribution of the gathered data was shown as well as some selected parts of the exploratory data analysis. Chapter 8 guided us through an extensive statistical data analysis, revealing many insights into the subject's rating tendencies. This was then followed by a discussion of selected findings and explanations of encountered challenges and limitations. In Chapter 9 the useful contributions are laid out. With the following outlook on further investigations, possible future advances for research in AVT reception were posited. This brings us to the final part of this work, the research questions from the beginning.

9.4. In Pursuit of Knowledge

This works character was of such an exploratory nature, that even the initial statement of the research questions has been peppered with a dose of uncertainty. As elucidated in Section 3.5 as part of Chapter 3, the content of this work, aiming to fill gaps in the current research of AVT reception, has been built on sometimes rather vague implications (lip-synchrony is important, yes. But how important?) and instructions (these phonemes are to be investigated, yes. But how exactly and in what manner?).

In the following, the in Chapter 1 stated questions will be answered with the, through the study gained knowledge.

"To what degree is the perceived quality of a video affected by Lip-Synchrony?"

Test items, created by modifying solely the visual components of the video sequence, received comparably worse quality ratings than all other types of modification (as can be seen in Table 8.2). This sole observation indicates, that the investigated degree of lip-synchrony on the perceived quality of a video ought to be quite considerable. Because other than the disruption of proper lip-synchrony, the other test items also contained unwanted artefacts, lowering the sound quality, which arguably should result in even lower ratings.

"Is there a threshold for the degree to which artefacts in Lip-Synchrony can be tolerated by (lay-) viewer?"

Working under the assumption, that the experiments actually assessed the subject's

perceived video quality under different levels of lip synchrony and not just simply different levels of video quality, it can be concluded, that such a threshold, if it exists, has to be of a very low value. Since, as can be seen in the Tables A.1, A.2, A.3, A.4, A.5, even the tiniest modifications, resulted in a drop in item ratings, subjects' prove to be extremely sensitive to introduced artefacts. There are some noticeable exceptions, like in Figures A.1c, A.1d, where the results clearly show, that subjects had difficulties, distinguishing the High-Anchor (unmodified hidden reference video sequence) from the next best test item. This indicates an introduced change of quality, barely detectable by viewers. Similar observations can be made for multiple test settings, but no definite principle could be extracted by the author at the time of writing.

"Is there a difference in the effect size for different phonemes?"

They found differences between the bilabial consonant "p" and short "pause" in regard to grading, were displayed as statistically very significant, while the open vowel "a" showed only some significance in the pairwise comparison with "p", as previously shown in Figure 8.3. These very minor, especially compared to some other observed, differences would spontaneously indicate no difference in the investigated effect size. However, it is to consider that only very few select phonemes could be included in the experiments of this work. These were exactly the ones, past research deemed essential for viewers detection of lip-synchrony and therefore can be expected to show similar results under equal experiment conditions.

To answer this question more exhaustively, would require the inclusion of, ideally, many more, but at least, very different phonemes into the investigation.

"Is there a difference in the effect size for different severity of modification?"

The answer to this research, and more of a fall-back question, can clearly be given by consulting Figure 8.6 and has to be a yes. This confirmation might be carelessly dismissed if it would not so ideally serve to conclude the subject's ability to detect a greater level of modifications of the test items. Therefore it can be taken for granted, that the results at hand are at the very least, not based on random chance.

After answering the proposed research questions, there are still many more to be investigated. One such possible question, a combination of the prior two, that could be answered with the given data set would be:

"Is there a difference in the effect size for different combinations of phoneme and type of modification?".

The data set produced in this work can yield more information to potentially gain further insights into the question above and so much more.

Appendices

A. Appendix

A.1. Data Structure

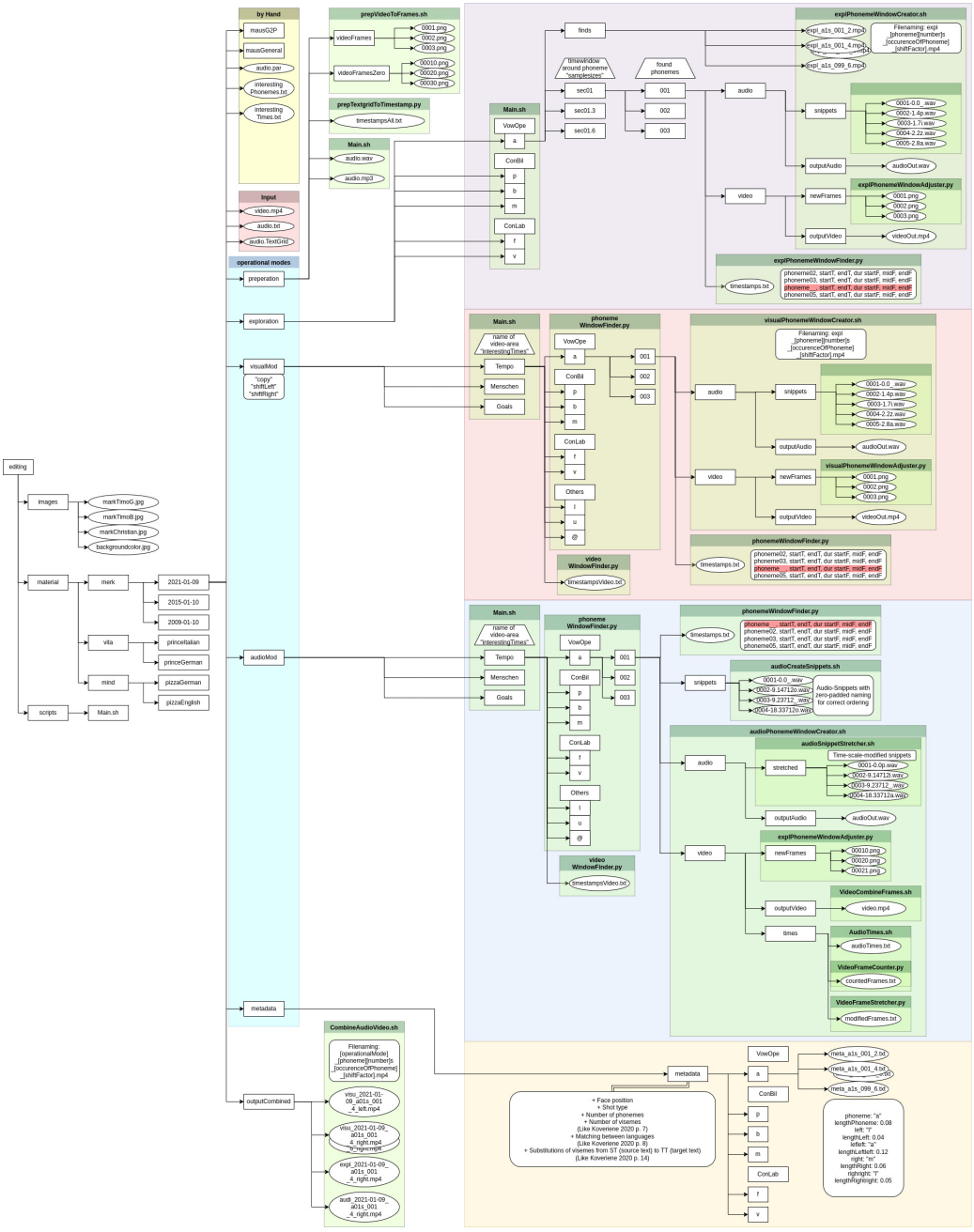
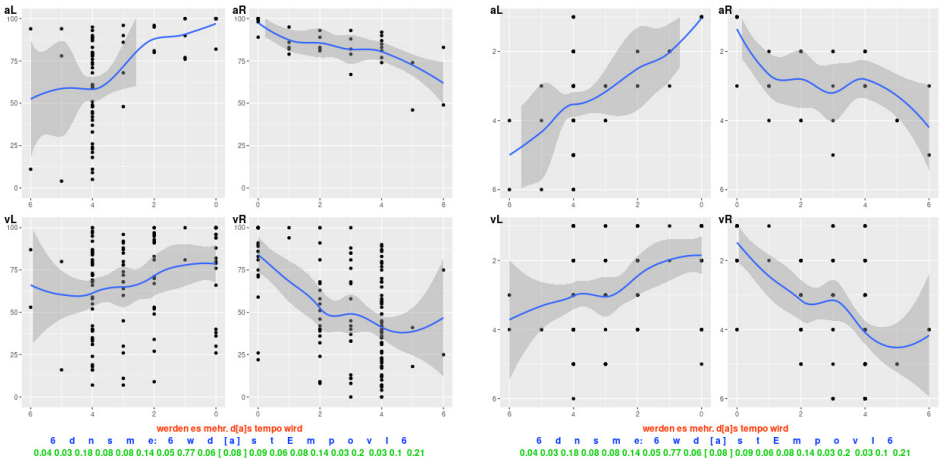


Figure A.1.: High-Level overview of datastructure for material-editing-process of this thesis.

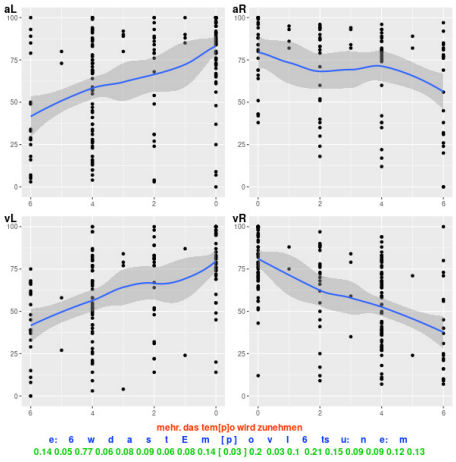
A.2. Complete Exploratory Data Analysis

A.3. Analysis of Listening Panel

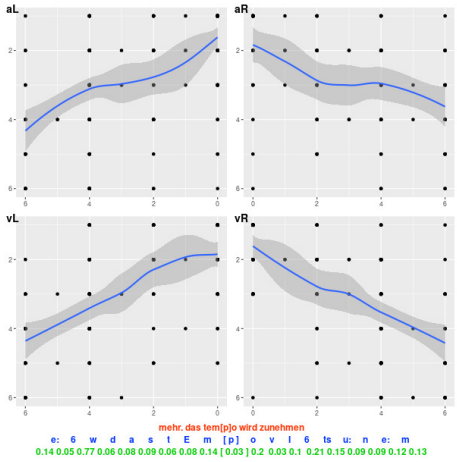


(a) Tempo-a for grade

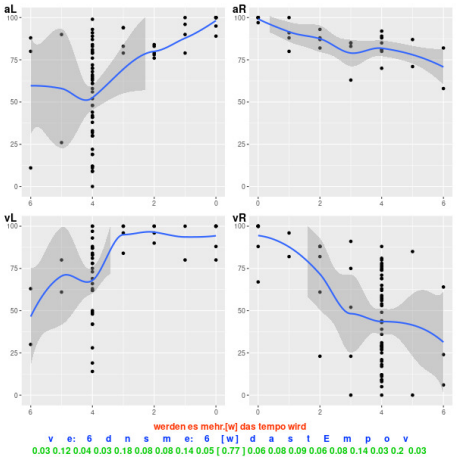
(b) Tempo-a for rank



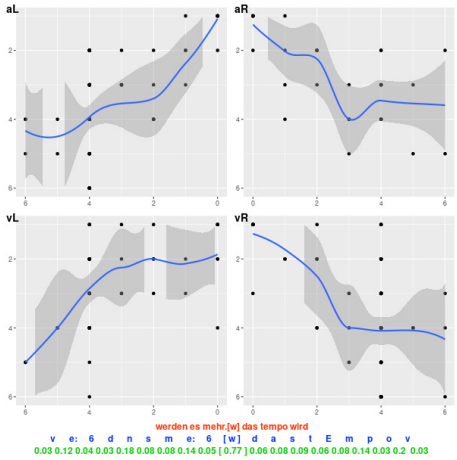
(c) Tempo-p for grade



(d) Tempo-p for rank

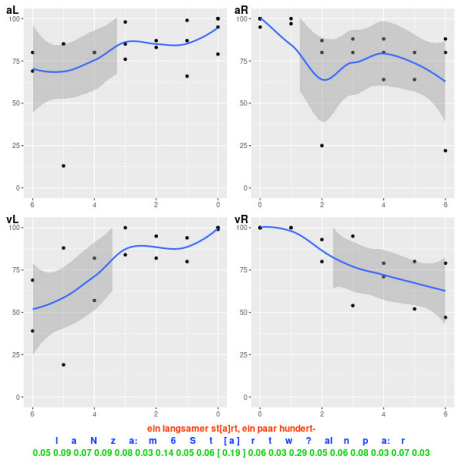


(e) Tempo-w for grade

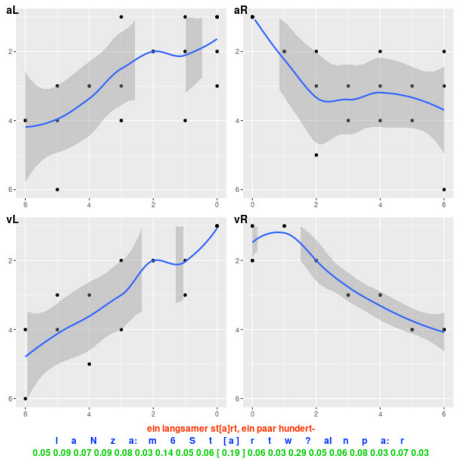


(f) Tempo-w for rank

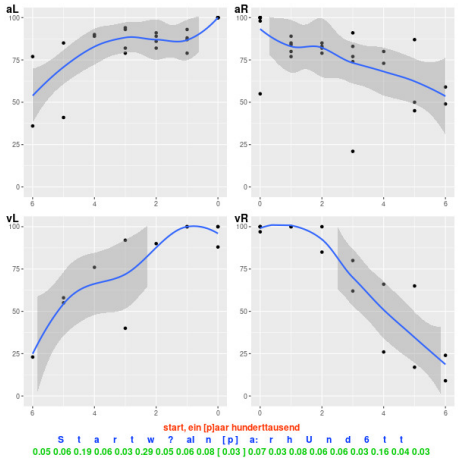
Table A.1.: Overview of exploratory analysis for excerpt Tempo.
aL: audio shifted to the left; aR: audio shifted to the right;
vL: visual shifted to the left, vR: visual shifted to the right.



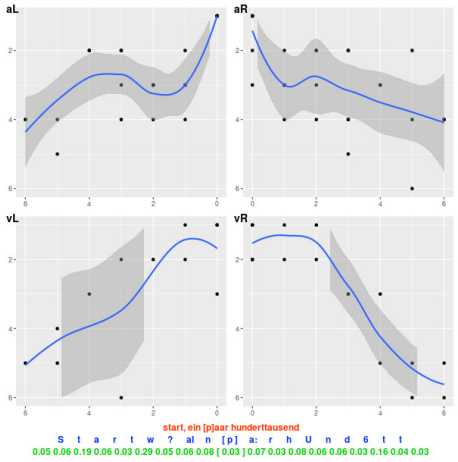
(a) Paar-a for grade



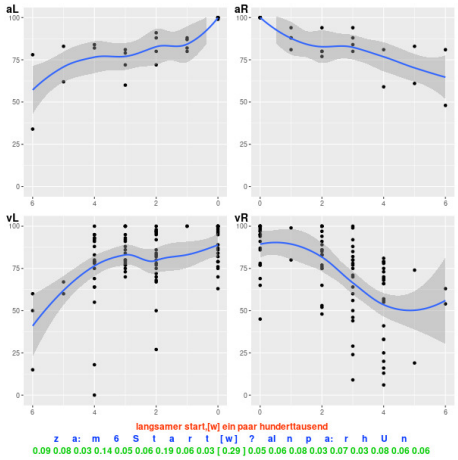
(b) Paar-a for rank



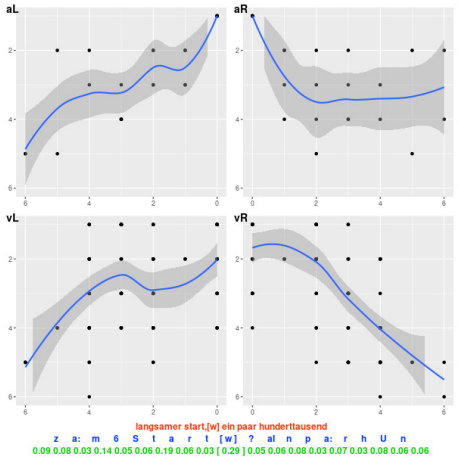
(c) Paar-p for grade



(d) Paar-p for rank

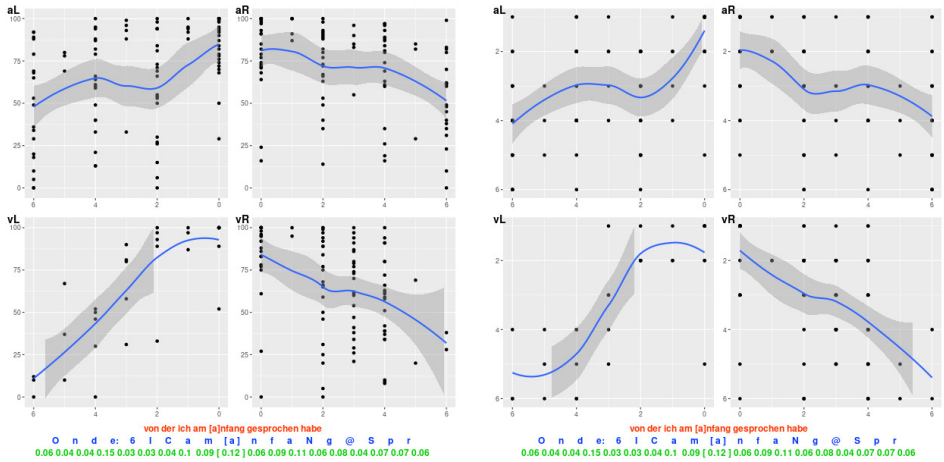


(e) Paar-w for grade



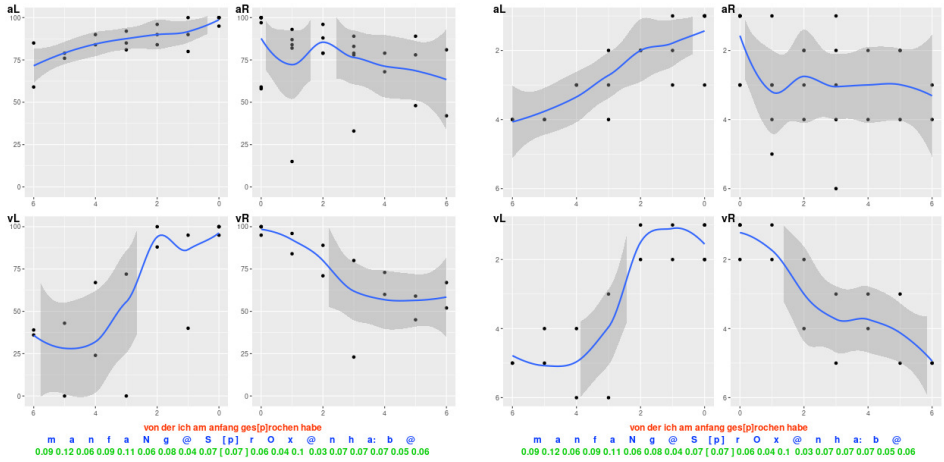
(f) Paar-w for rank

Table A.2.: Overview of exploratory analysis for excerpt Paar.
aL: audio shifted to the left; aR: audio shifted to the right;
vL: visual shifted to the left, vR: visual shifted to the right.



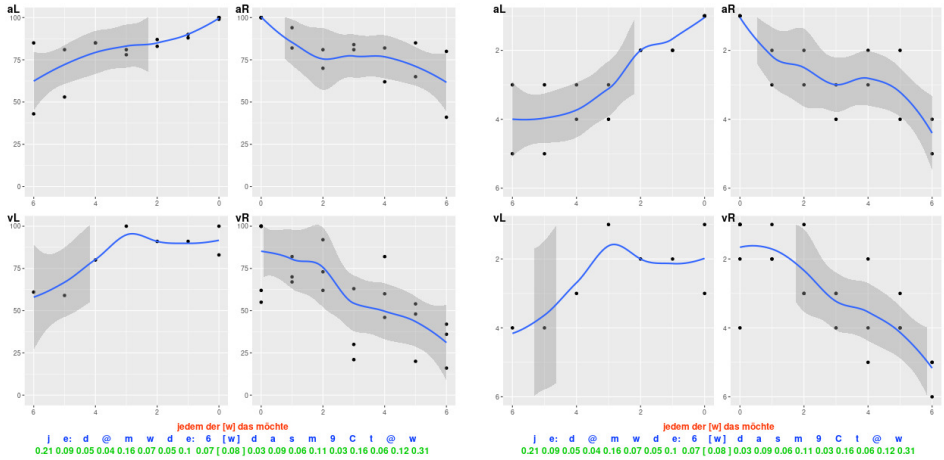
(a) Anfang-a for grade

(b) Anfang-a for rank



(c) Anfang-p for grade

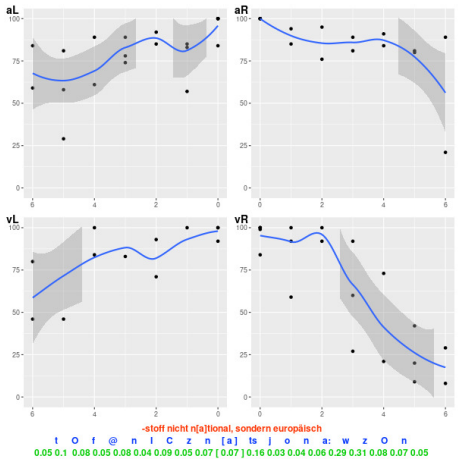
(d) Anfang-p for rank



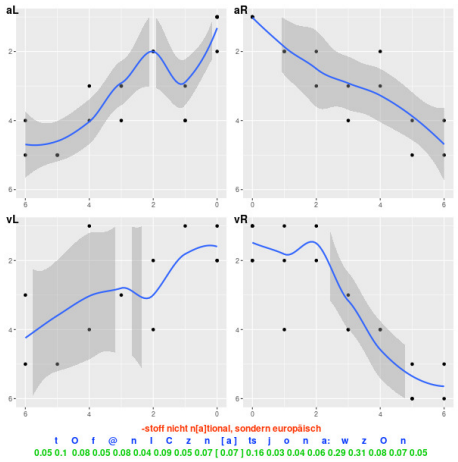
(e) Impfangebot-w for grade

(f) Impfangebot-w for rank

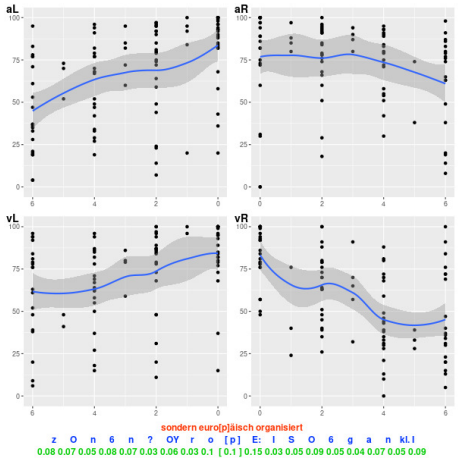
Table A.3.: Overview of exploratory analysis for excerpts Anfang and Impfangebot. aL: audio shifted to the left; aR: audio shifted to the right; vL: visual shifted to the left, vR: visual shifted to the right.



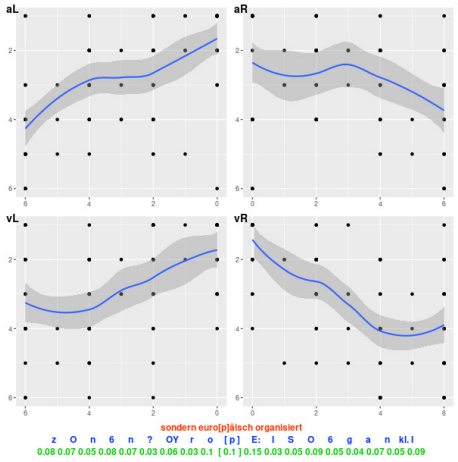
(a) Europaeisch-a for grade



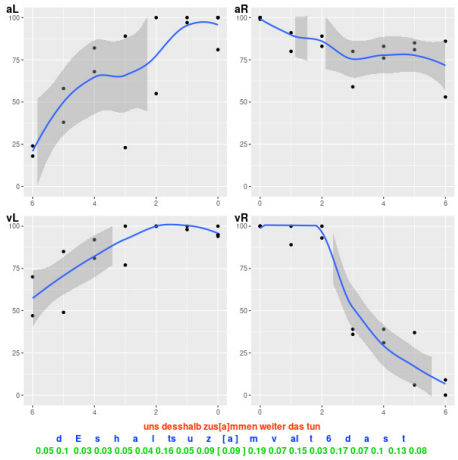
(b) Europaeisch-a for rank



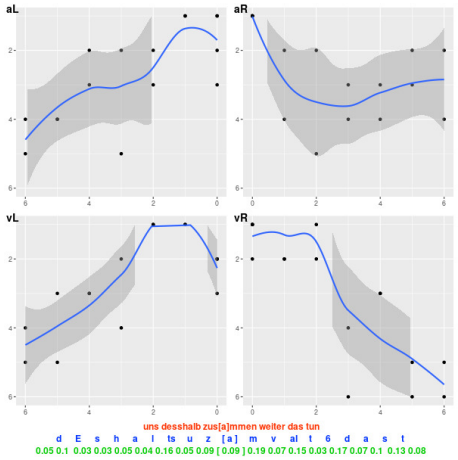
(c) Europaeisch-p for grade



(d) Europaeisch-p for rank

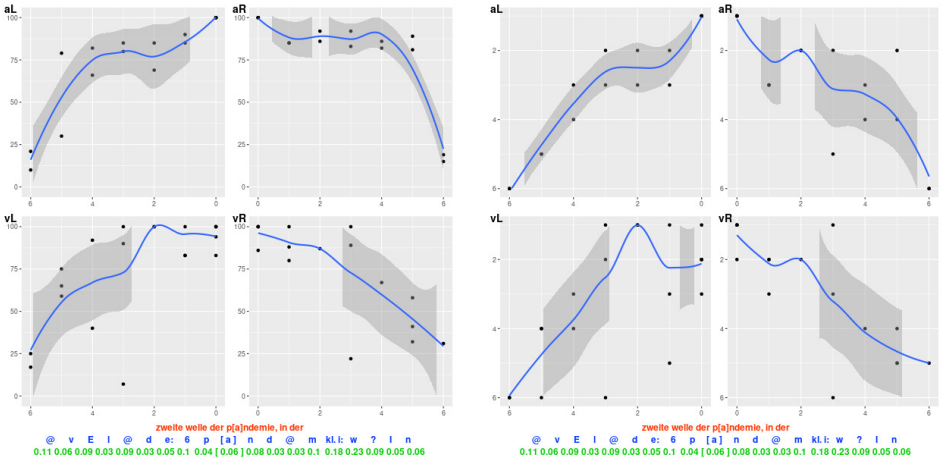


(e) Zusammen-a for grade



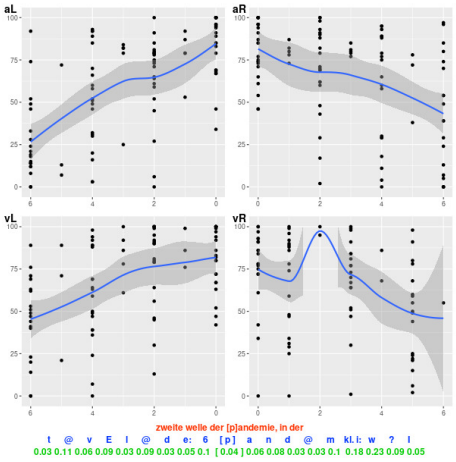
(f) Zusammen-a for rank

Table A.4.: Overview of exploratory analysis for excerpts Europaeisch and Zusammen. aL: audio shifted to the left; aR: audio shifted to the right; vL: visual shifted to the left, vR: visual shifted to the right.

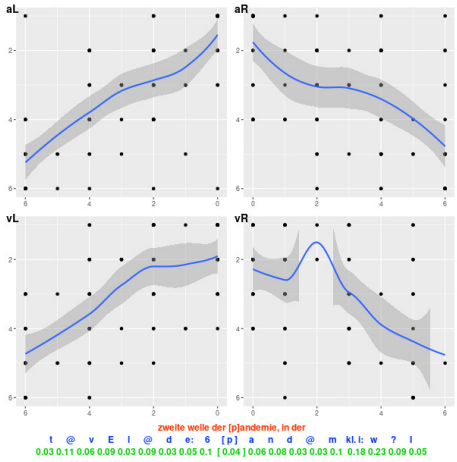


(a) Pandemie-a for grade

(b) Pandemie-a for rank



(c) Pandemie-p for grade



(d) Pandemie-p for rank

Table A.5.: Overview of exploratory analysis for excerpt Pandemie.
aL: audio shifted to the left; aR: audio shifted to the right;
vL: visual shifted to the left, vR: visual shifted to the right.

Excerpt	Object	Modification	Attribute	meanRank	varRank	meanGrade	varGrade	number
Anfang	a	aL	0	1.368421	0.9122807	82.05263	350.4971	19
Anfang	a	aL	2	3.210526	1.3976608	56.05263	996.6082	19
Anfang	a	aL	4	2.947368	2.0526316	60.21053	776.7310	19
Anfang	a	aL	6	4.157895	2.2514620	47.15789	1085.6959	19
Anfang	a	aR	0	2.210526	1.5087719	76.57895	527.9240	19
Anfang	a	aR	2	3.000000	2.8888889	70.57895	560.1462	19
Anfang	a	aR	4	2.736842	1.7602339	69.21053	720.3977	19
Anfang	a	aR	6	3.842105	2.4736842	51.47368	665.0409	19
Anfang	a	iL	4	5.271186	0.9596727	38.98305	642.1204	59
Anfang	a	uR	4	3.932203	2.2367037	52.38983	690.2420	59
Anfang	a	vR	0	1.714286	2.1142857	81.23810	626.6905	21
Anfang	a	vR	2	3.190476	2.4619048	61.52381	969.8619	21
Anfang	a	vR	3	3.142857	1.0285714	62.80952	643.5619	21
Anfang	a	vR	4	3.761905	2.2904762	57.09524	785.3905	21
Europaeisch	p	aL	0	1.761905	1.2904762	82.61905	505.2476	21
Europaeisch	p	aL	2	2.619048	1.7476190	67.38095	862.1476	21
Europaeisch	p	aL	4	2.809524	1.0619048	60.85714	560.0286	21
Europaeisch	p	aL	6	4.190476	2.0619048	44.52381	736.6619	21
Europaeisch	p	aR	0	2.631579	3.9122807	71.42105	1056.7018	19
Europaeisch	p	aR	2	2.526316	1.1520468	74.73684	501.2047	19
Europaeisch	p	aR	4	2.789474	1.2865497	72.89474	370.6550	19
Europaeisch	p	aR	6	3.631579	2.5789474	61.57895	807.9240	19
Europaeisch	p	iL	4	5.391753	0.7407646	29.32990	430.3484	97
Europaeisch	p	iR	0	1.777778	1.4771242	86.05556	145.4673	18
Europaeisch	p	iR	1	2.777778	1.8300654	72.77778	606.3007	18
Europaeisch	p	iR	2	2.833333	0.9705882	70.38889	378.2516	18
Europaeisch	p	iR	3	3.388889	1.5457516	65.00000	533.6471	18
Europaeisch	p	uR	4	4.247423	2.2506443	46.04124	800.2691	97
Europaeisch	p	vL	0	1.789474	0.9532164	81.36842	491.4678	19
Europaeisch	p	vL	2	2.578947	1.9239766	73.47368	777.4854	19
Europaeisch	p	vL	4	3.368421	1.8011696	64.15789	593.4737	19
Europaeisch	p	vL	6	3.263158	2.3157895	62.42105	833.0351	19
Europaeisch	p	vR	0	1.400000	0.6736842	83.45000	227.5237	20
Europaeisch	p	vR	2	2.650000	0.9763158	64.70000	468.6421	20
Europaeisch	p	vR	4	4.000000	1.6842105	44.90000	639.3579	20
Europaeisch	p	vR	6	3.900000	2.2000000	45.75000	822.0921	20
Paar	w	iL	4	5.820513	0.3090418	23.00000	390.4211	39
Paar	w	uR	4	4.333333	1.2280702	53.25641	530.7746	39
Paar	w	vL	0	1.842105	1.1403509	87.57895	140.5906	19
Paar	w	vL	2	2.789474	1.5087719	79.68421	323.3392	19
Paar	w	vL	3	2.631579	1.4678363	83.47368	93.7076	19
Paar	w	vL	4	3.315789	1.7836257	73.26316	669.0936	19
Paar	w	vR	0	1.750000	1.0394737	87.25000	214.8289	20
Paar	w	vR	2	2.150000	1.2921053	80.35000	248.6605	20
Paar	w	vR	3	3.300000	1.6947368	65.55000	628.3658	20
Paar	w	vR	4	3.900000	1.3578947	51.65000	654.8711	20
Pandemie	p	aL	0	1.750000	1.5657895	80.65000	335.9237	20
Pandemie	p	aL	2	2.600000	1.7263158	63.20000	688.1684	20
Pandemie	p	aL	4	3.600000	1.2000000	52.55000	866.2605	20
Pandemie	p	aL	6	5.250000	0.8289474	27.85000	584.6605	20
Pandemie	p	aR	0	1.900000	1.5684211	77.90000	295.9895	20
Pandemie	p	aR	2	3.000000	1.7894737	67.15000	777.7132	20
Pandemie	p	aR	4	3.400000	2.7789474	58.45000	1116.2605	20
Pandemie	p	aR	6	4.650000	2.3447368	43.55000	1092.8921	20
Pandemie	p	iL	0	2.210526	2.7309942	69.31579	745.7836	19
Pandemie	p	iL	1	2.578947	1.4795322	61.36842	691.0234	19
Pandemie	p	iL	2	3.736842	3.0935673	49.31579	905.5614	19
Pandemie	p	iL	3	4.684211	1.6725146	41.73684	576.2047	19
Pandemie	p	iL	4	5.042105	1.0194849	37.29474	672.6356	95
Pandemie	p	uR	4	3.136842	2.2044793	63.74737	700.9568	95
Pandemie	p	vL	0	2.055556	0.9967320	76.88889	332.4575	18
Pandemie	p	vL	2	2.277778	1.5065359	75.00000	706.2353	18
Pandemie	p	vL	4	3.444444	2.3790850	59.27778	904.6830	18
Pandemie	p	vL	6	4.777778	1.3594771	43.33333	684.3529	18
Pandemie	p	vR	0	2.333333	1.2941176	72.77778	618.1830	18
Pandemie	p	vR	1	2.722222	2.3300654	64.72222	963.3889	18
Pandemie	p	vR	3	3.000000	2.0000000	69.66667	620.0000	18
Pandemie	p	vR	5	4.222222	3.2418301	48.77778	981.0065	18
Tempo	a	iL	4	5.621622	0.7417417	20.35135	305.7898	37
Tempo	a	uR	4	3.783784	2.0075075	47.27027	598.0916	37
Tempo	a	vL	0	1.944444	1.8202614	73.38889	724.3693	18
Tempo	a	vL	2	2.555556	1.2026144	68.27778	703.6242	18
Tempo	a	vL	3	3.055556	1.8202614	64.66667	646.8235	18
Tempo	a	vL	4	3.048780	2.1975610	61.29268	758.0122	41
Tempo	a	vR	0	1.736842	1.0935673	79.26316	516.0936	19
Tempo	a	vR	2	2.894737	1.4327485	52.89474	744.6550	19
Tempo	a	vR	3	3.000000	2.3333333	51.31579	855.6725	19
Tempo	a	vR	4	4.473684	1.7076023	31.89474	471.0994	19
Tempo	p	aL	0	1.526316	0.5964912	80.52632	328.2632	19
Tempo	p	aL	2	2.789474	2.7309942	62.47368	1092.3743	19
Tempo	p	aL	4	3.263158	1.4269006	57.42105	948.3684	19
Tempo	p	aL	6	4.263158	2.4269006	42.47368	1137.7076	19
Tempo	p	aR	0	2.000000	2.1000000	73.95238	414.9476	21
Tempo	p	aR	2	2.809524	1.9619048	66.38095	710.3476	21
Tempo	p	aR	4	2.952381	1.9476190	67.85714	623.7286	21
Tempo	p	aR	6	3.666667	2.1333333	55.71429	948.7143	21
Tempo	p	iL	4	5.095238	1.1428571	30.22222	407.5982	126
Tempo	p	iR	0	2.050000	1.6289474	74.70000	483.3789	20
Tempo	p	iR	1	2.500000	1.0000000	63.55000	381.2079	20
Tempo	p	iR	2	2.850000	1.6078947	61.25000	737.9868	20
Tempo	p	iR	3	3.450000	3.1026316	56.40000	928.0421	20
Tempo	p	uR	4	4.357143	2.1354286	42.00794	699.6079	126
Tempo	p	vL	0	1.955556	1.6797980	76.68889	517.9010	45
Tempo	p	vL	2	2.363636	2.0519481	66.27273	751.9221	22
Tempo	p	vL	4	3.244444	1.7797980	57.24444	649.4616	45
Tempo	p	vL	6	4.454545	1.4025974	41.36364	560.9091	22
Tempo	p	vR	0	1.571429	0.7571429	80.71429	252.5143	21
Tempo	p	vR	2	2.809524	2.1619048	62.90476	738.9905	21
Tempo	p	vR	4	3.714286	1.7142857	49.19048	659.6619	21
Tempo	p	vR	6	4.380952	2.0476190	37.19048	684.3619	21
Tempo	w	vL	4	2.869565	1.3003953	64.60870	602.5217	23

Table A.6.: Analyse of rating tendencies of listening panel from study part Blue.

SubjectName	meanRankDistanceMean	meanGradeDistanceMean	meanRuntime	numberTrials
UristMc3e5d53	0.591666666666667	25.537500	94629.50	12
UristMccc2211	0.673971163372715	16.949150	60427.91	138
UristMc299a6b	0.675438596491228	3.149123	66043.00	6
UristMc562e14	0.743635762385762	15.583751	47040.25	24
UristMc74404a	0.843743981865471	11.106226	165938.91	66
UristMcaae4cb	0.846568018916417	21.269719	55255.65	120
UristMc2ed68c	0.856481481481481	22.641049	52467.67	18
UristMc76a32a	0.875963055787617	29.170877	74297.08	72
UristMc33f2f0	0.912459988211157	14.556499	86755.45	138
UristMcc57a78	0.937817513975981	11.349591	55311.38	48
UristMc95275d	0.944305207463102	17.593428	48970.67	18
UristMcc3cee5	0.954007482474302	34.930324	75280.90	120
UristMc94defa	0.980877749746934	25.714076	134481.50	72
UristMc916822	0.984796985829221	22.126456	36104.39	138
UristMca712b6	0.985028473537419	11.099284	100734.32	132
UristMca0ceed	1.00907704042715	20.174194	46871.75	24
UristMc941662	1.01261569327359	20.650002	106146.60	60
UristMc0d6bca	1.02268610581465	24.959321	90752.48	138
UristMc69243d	1.02469992623957	16.656666	54674.30	138
UristMce9593f	1.02963527982597	27.329887	55979.67	18
UristMc2da0b6	1.03032415273735	19.335856	63420.52	138
UristMc865aa3	1.05340657428261	17.385099	42626.43	138
UristMce57309	1.0577233064683	25.069018	59669.75	96
UristMce92c85	1.05958764820818	18.841437	46664.22	138
UristMce05022	1.06349206349206	21.809524	96096.00	6
UristMcae6da2	1.07619304964594	19.879686	30079.09	138
UristMc234d91	1.09373863834881	21.791359	43227.71	84
UristMc534634	1.15026547029582	18.752570	26785.00	138
UristMcb42403	1.1874537153623	19.292413	81174.16	114
UristMc390561	1.21927214137193	19.875659	47724.61	138
UristMc8dc0db	1.22152777777778	19.709325	68439.75	24
UristMc56762b	1.38824632319238	20.204932	69713.43	42
UristMcb958bc	1.46031746031746	21.444444	54147.00	6
UristMc463fa2	1.55118472509777	20.546354	36840.67	18

Table A.7.: Subjects tendency to rate according to the rating tendencies of listening panel sorted by distance from mean of rank from study part Blue.

SubjectName	meanRankDistanceMean	meanGradeDistanceMean	meanRuntime	numberTrials
UristMccc2211	0.492307685050411	10.480797	84430.66	1260
UristMcaae4cb	0.507633364382212	11.391278	52254.68	1194
UristMc3e5d53	0.681150793650794	24.729365	111682.33	18
UristMc48ab73	0.711539550443059	23.402093	62304.33	90
UristMcc3e95a	0.759120022277917	9.009156	76569.60	90
UristMc1c344c	0.829211048454469	7.894295	137270.08	72
UristMc33f2f0	0.832541218728932	13.728314	68774.96	294
UristMc98e46a	0.833333333333333	12.608466	91582.00	6
UristMcb5d87f	0.838624338624339	21.489418	125252.00	6
UristMcacf358	0.849206349206349	11.058201	103785.00	6
UristMc74404a	0.85963127790931	10.787850	159424.42	72
UristMc79d4ce	0.867155156409542	17.901741	46882.27	90
UristMc9711aa	0.879519400352734	17.219444	43546.60	90
UristMcc2077e	0.882355729066255	13.147442	101256.80	90
UristMc19ff22	0.891534391534392	19.891534	239569.00	6
UristMc76a32a	0.894887267772606	29.019246	73096.86	84
UristMc299a6b	0.917937204121415	8.769505	135204.33	18
UristMcb14fc3	0.919642857142857	26.660714	177700.00	6
UristMc562e14	0.930556557639891	16.098713	59558.67	36
UristMc51e3cd	0.94973544973545	18.756614	69797.00	6
UristMc941662	0.955773494937582	21.853724	118186.12	102
UristMcf6c881	0.967846967846968	17.849006	34147.62	48
UristMc69243d	0.967875571289791	16.249920	59954.61	216
UristMcc57a78	0.975551630228404	11.331988	57766.50	60
UristMcb6d717	0.976741622574956	15.748787	300000.00	18
UristMcc3cee5	0.979578110570673	34.421296	81553.91	132
UristMcde08e0	0.981484462580954	17.143609	107451.87	90
UristMca0ceed	0.987618775198867	18.140546	51560.60	30
UristMc94defa	0.990361512586913	26.735613	134705.08	78
UristMc916822	1.00713756558722	22.250149	52447.72	150
UristMc95275d	1.00884502923977	16.799046	62364.40	30
UristMca712b6	1.01302047904951	11.348175	101491.77	144
UristMc2da0b6	1.02661515173529	19.632797	61041.88	150
UristMc0d6bca	1.03675746073572	24.793978	92401.68	150
UristMc0ec85c	1.04189919242551	17.712830	53085.00	30
UristMce57309	1.07060933045756	24.937157	62029.82	102
UristMce05022	1.0706569664903	22.287257	102337.33	18
UristMc30f1e6	1.08500116030818	19.544442	64649.27	90
UristMc76f087	1.08862664550924	21.826058	80291.77	78
UristMce92c85	1.09414603317693	19.656291	45502.32	150
UristMc865aa3	1.09782452453047	17.660031	45235.36	150
UristMc234d91	1.10643020405785	22.090959	50328.62	96
UristMc47bb27	1.11026713864433	15.135935	65413.60	90
UristMcae6da2	1.11302088609755	21.060316	30055.88	150
UristMc2ed68c	1.11712962962963	25.671402	54312.20	30
UristMc534634	1.1531251850531	19.504296	27837.44	150
UristMcbdd99e	1.16666666666667	18.275132	112231.00	6
UristMccbcbc6	1.16749338624339	13.221230	141114.50	12
UristMc332197	1.18253968253968	23.206349	111737.00	6
UristMcb64697	1.18766969507101	32.021888	64387.00	24
UristMcb42403	1.20832295083548	19.550310	95057.71	126
UristMce9593f	1.21321767583209	25.542906	68184.60	30
UristMcf54dcf	1.21428571428571	22.058201	217540.00	6
UristMcc0c2a2	1.21478174603175	25.865410	72306.50	12
UristMc390561	1.21907163990345	20.099231	46972.88	150
UristMce54b2c	1.22794261294261	17.508851	67754.73	90
UristMc3162e2	1.22804324608836	16.997273	67816.86	84
UristMc10cd42	1.24074074074074	19.362434	121790.00	6
UristMc3126dd	1.24074074074074	14.891534	94639.00	6
UristMc2cc0ac	1.24404761904762	27.503604	74789.20	30
UristMc5bad5b	1.2729828042328	30.973876	174070.00	12
UristMcb958bc	1.27645502645503	19.100529	84504.00	12
UristMc8dc0db	1.3145171957672	20.270392	69274.00	36
UristMc29aeec	1.3275462962963	22.893519	72420.00	12
UristMc463fa2	1.33963940648723	18.741966	46952.60	30
UristMc56762b	1.34157565184096	19.242559	75617.00	48
UristMc308a9a	1.40443121693122	14.353450	60812.67	18
UristMc753ece	1.43086964205385	12.573716	102280.60	90
UristMc115cdd	1.56349206349206	38.108466	77449.00	6
UristMcd36768	1.67080026455026	21.895503	92939.50	12

Table A.8.: Subjects tendency to rate according to the rating tendencies of listening panel sorted by distance from mean of rank.

References

Alves Veiga, M. J. (2006). *Subtitling reading practices.* (In A. Assis Rosa & T. Seruya (Eds.), Translation studies at the interface of disciplines (pp. 161-168). Amsterdam: John Benjamins.)

Ameri, S., Khoshsaligheh, M., & Farid, A. K. (2015). Investigating Expectancy Norms in Dubbing in Iran: An Exploratory Study. *undefined*. Retrieved 2022-01-19, from https://www.semanticscholar.org/paper/Investigating-Expectancy-Norms-in-Dubbing-in-Iran%3A-Ameri-Khoshsaligheh/26eb6230c9a07e6fb46d0c95bc50c20b841e2493

Ameri, S., Khoshsaligheh, M., & Farid, A. K. (2018, May). The reception of Persian dubbing: a survey on preferences and perception of quality standards in Iran. *Perspectives*, *26*(3), 435–451. Retrieved 2021-01-28, from https://www.tandfonline.com/doi/full/10.1080/0907676X.2017.1359323 (Number: 3) doi: 10.1080/0907676X.2017.1359323

Antonini, R. (2005). *The perception of subtitled humour in italy: An empirical study.* (Humor, Journal of Humor Research, 18(2), 209-225.)

Antonini, R. (2007). *SAT, BLT, Spirit Biscuits, and the Third Amendment.* Retrieved 2022-01-19, from https://benjamins.com/catalog/btl.72.17ant (Publisher: John Benjamins Publishing Company)

Antonini, R. (2008). *The perception of dubbese: An italian study.* (In D. Chiaro, C. Heiss, & C. Bucaria (Eds.), Between text and image: Updating research in screen translation (pp. 135-147). Amsterdam: John Benjamins.)

Antonini, R., & Chiaro, D. (2005). *The quality of dubbed television programmes in italy: The experimental design of an empirical study.* (In M. Bondi & N. Maxwell (Eds.), Cross-cultural encounters: Linguistic perspectives (pp. 33-44). Rome: Officina Edizioni.)

Antonini, R., & Chiaro, D. (2009). *Perception of dubbing by italian audiences.* (In J. Díaz-Cintas & G. Anderman (Eds.), Audiovisual translation: Language transfer on screen (pp. 97-104). Basingstoke: Palgrave MacMillan.)

Arduini, S., & Hodgson, R. (2007). *Similarity and Difference in Translation.* Ed. di Storia e Letteratura.

Aschenberner, B., & Weiss, C. (2005). Phoneme-Viseme Mapping for German Video-Realistic Audio-Visual-Speech-Synthesis IKP - Working Paper NF 1. , 11.

Barbe, K. (1996). *Dubbing in the translation classroom.* (In: Perspectives: Studies in Translatology, 4 (2), pp. 255-274. https://doi.org/10.1080/0907676X.1996.9961291)

Barman, N., Martini, M. G., Zadtootaghaj, S., Moller, S., & Lee, S. (2018, May). A Comparative Quality Assessment Study for Gaming and Non-Gaming Videos. In *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)* (pp. 1–6). Cagliari: IEEE. Retrieved 2021-01-28, from https://ieeexplore.ieee.org/document/8463403/ doi: 10.1109/QoMEX.2018.8463403

Barsam, R., & Mohanan, D. (2010). *Looking at movies: An introduction to film.* (3 rd ed. New York:W. W. Norton & Company.)

Baumann, T. (2017, August). Large-Scale Speaker Ranking from Crowdsourced Pairwise Listener Ratings. In *Interspeech 2017* (pp. 2262–2266). ISCA. doi: 10.21437/Interspeech.2017-1697

Bear, H., & Harvey, R. (2019, September). Alternative Visual Units for an Optimized Phoneme-Based Lipreading System. *Applied Sciences*, *9*(18), 3870. Retrieved 2021-12-14, from http://arxiv.org/abs/1909.07147 (arXiv: 1909.07147) doi: 10.3390/app9183870

Bear, H. L., & Harvey, R. (2017, December). Phoneme-to-viseme mappings: the good, the bad, and the ugly. *Speech Communication*, *95*, 40–67. Retrieved 2021-10-12, from http://arxiv.org/abs/1805.02934 (arXiv: 1805.02934) doi: 10.1016/j.specom.2017.07.001

Borell, J. (2000). *Subtitling or dubbing? An investigation of the effects from reading subtitles on understanding audiovisual material.*

Bucaria, C. (2005). *The perception of humour in dubbing vs. subtitling: The case of six feet under.* (ESP Across Cultures, 2, pp.34-46.)

Bucaria, C. (2008). *Acceptance of the norm or suspension of disbelief?: The case of formulaic language in dubbese.* (In D. Chiaro, C. Heiss, & C. Bucaria (Eds.), Between text and image: Updating research in screen translation (pp. 149-163). Amsterdam: John Benjamins.)

Buchan, J. N., Paré, M., & Munhall, K. G. (2008, November). The effect of varying talker identity and listening conditions on gaze behavior during audiovisual speech perception. *Brain Research*, *1242*, 162–171. Retrieved 2021-01-28, from https://linkinghub.elsevier.com/retrieve/pii/S0006899308014832 doi: 10.1016/j.brainres.2008.06.083

Caffrey, C. (2008). *Viewer perception of visual nonverbal cues in subtitled tv anime.* (European Journal of English Studies, 12(2), 163-178. doi:10.1080/13825570802151439)

Cappelletta, L., & Harte, N. (2012). PHONEME-TO-VISEME MAPPING FOR VISUAL SPEECH RECOGNITION:. In *Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods* (pp. 322–329). Vilamoura, Algarve, Portugal: SciTePress - Science and and Technology Publications. Retrieved 2021-10-10, from http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0003731903220329 doi: 10.5220/0003731903220329

Cavaliere, F. (2008). *Measuring the perception of the screen translation of un posto al sole: A cross-cultural study.* (In D. Chiaro, C. Heiss, & C. Bucaria (Eds.), Between text and image: Updating research in screen translation (pp. 165-180). Amsterdam: John Benjamins.)

Chaume, F. (2007). Dubbing practices in Europe: localisation beats globalisation. (Linguistica Antverpiensia, New Series - Themes in Translation Studies) doi:

10.52034/lanstts.v6i.188

Chaume, F. (2012). *Audiovisual Translation: Dubbing.* doi: 10.4324/9781003161660

Chaume, F. (2013, January). Research Paths in Audiovisual Translation: The Case of Dubbing. In *Routledge Handbook of Translation Studies* (pp. 288–302). Routledge.

Chaume, F. (2016, April). Audiovisual Translation Trends: Growing Diversity, Choice, and Enhanced Localization. In *Media Across Borders. Localizing TV, Film and Video Games* (pp. 68–84). Routledge.

Chaume, F. (2018a, July). Is audiovisual translation putting the concept of translation up against the ropes? *The Journal of Specialised Translation*, 84–104.

Chaume, F. (2018b, November). An overview of audiovisual translation: Four methodological turns in a mature discipline. *Journal of Audiovisual Translation*, *1*, 40–63. doi: 10.47476/jat.v1i1.43

Chaume, F. (2020a, November). The language of dubbing: a matter of compromise. In *Audiovisual translation : dubbing* (pp. 81–99). Routledge. doi: 10.4324/9781003161660-5

Chaume, F. (2020b, November). Research in dubbing. In *Audiovisual translation : dubbing* (pp. 158–179). Routledge. doi: 10.4324/9781003161660-8

Chaume, F. (2021, January). *Textual constraints and the translator's creativity in dubbing.* Retrieved 2021-01-29, from `https://benjamins.com/catalog/btl.27.04cha` (Publisher: John Benjamins Publishing Company)

Chaume Varela, F. (2004). Synchronization in dubbing: A translational approach. In P. Orero (Ed.), *Benjamins Translation Library* (Vol. 56, pp. 35–52). Amsterdam: John Benjamins Publishing Company. Retrieved 2021-01-28, from `https://benjamins.com/catalog/btl.56.07cha` doi: 10.1075/btl.56.07cha

Cheng, S., Zeng, H., Chen, J., Hou, J., Zhu, J., & Ma, K.-K. (2020). Screen Content Video Quality Assessment: Subjective and Objective Study. *IEEE Transactions on Image Processing*, *29*, 8636–8651. Retrieved 2021-01-28, from `https://ieeexplore.ieee.org/document/9178481/` doi: 10.1109/TIP.2020.3018256

Chiaro, D. (2004). *Investigating the perception of translated verbally expressed humour on italian tv.* (ESP Across Cultures, 1, 35-52.)

Chiaro, D. (2007). *The effect of translation on humour response: The case of dubbed comedy in italy.* (In Y. Gambier, M. Shlesinger, & R. Stolze (Eds.), Doubts and directions in translation studies: Selected contributions from the EST congress, Lisbon 2004 (pp. 137-152). Amsterdam: John Benjamins.)

Chiaro, D. (2014). *The eyes and ears of the beholder? translation, humor, and perception.* (In D. Abend-David (Ed.), Media and translation: An interdisciplinary approach (pp. 197-219). Fakenham: Bloomsbury.)

Chmiel, A., & Mazur, I. (2012). *Ad reception research: Some methodological considerations.* (In E. Perego (Ed.), Emerging topics in translation: Audio description (pp. 57-80). Trieste: Università di Trieste.)

Cisco. (2020, March). *Cisco Annual Internet Report - Cisco Annual Internet Report (2018-2023) White Paper.* Retrieved 2021-12-17, from https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html

Delabastita, D. (1989). *Translation and mass-communication: Film and t.v. translation as evidence of cultural dynamics.* (In: Babel, 35 (4), pp. 193-218. https://doi.org/10.1075/babel.35.4.02del)

Denton, J., & Ciampi, D. (2012). *A new development in audiovisual translation studies:focus on target audience perception.* (LEA - Lingue e letterature d'Oriente e d'Occidente, 1, pp. 399-422.)

Duanmu, Z., Zeng, K., Ma, K., Rehman, A., & Wang, Z. (2016, September). A Quality-of-Experience Index for Streaming Video. *IEEE Journal of Selected Topics in Signal Processing*, *PP*, 154-166. doi: 10.1109/JSTSP.2016.2608329

d'Ydewalle, G., Muylle, P., & Van Rensbergen, J. (1985). *Attention shifts in partially redundant information situations.* (In R. Groner, G. McConkie, & C. Menz (Eds.), Eye movements and human information processing (pp. 375-384). Amsterdam: Elsevier.)

d'Ydewalle, G., Van Rensbergen, J., & Pollet, J. (1987, January). READING A MESSAGE WHEN THE SAME MESSAGE IS AVAILABLE AUDITORILY IN ANOTHER LANGUAGE: THE CASE OF SUBTITLING. In J. K. O'regan & A. Levy-schoen (Eds.), *Eye Movements from Physiology to Cognition* (pp. 313–321). Amsterdam: Elsevier. Retrieved 2022-01-19, from https://www.sciencedirect.com/science/article/pii/B9780444701138500473 doi: 10.1016/B978-0-444-70113-8.50047-3

Fernández Costales, A. (2016). *Analyzing players' perceptions on the translation of video games: Assessing the tension between the local and the global concerning language use.* (In A. Esser, M. ÃĄ. Bernal-Merino, & I. R. Smith (Eds.), Media across borders localizing TV, film, and video games (pp. 183-201). Abingdon: Routledge.)

Figuerola Salas, Ó., Adzic, V., Shah, A., & Kalva, H. (2013, October). Assessing internet video quality using crowdsourcing. In *Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia* (pp. 23–28). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/2506364.2506366

Flis, G., & Sikorski, A. (2019). *Does the dubbing effect apply to voice-over ? A conceptual replication study on visual attention and immersion.* Retrieved 2021-01-29, from /paper/Does-the-dubbing-effect-apply-to-voice-over-A-study-Flis-Sikorski/a8eeefbc7a7d4141d81fe53ff1a6c755682cbad0

Fodor, I. (1976). *Film dubbing phonetic, semiotic, esthetic and psychological aspects.* (Buske, Hanburg.)

Fryer, L., & Freeman, J. (2013). *Cinematic language and the description of film: Keeping ad users in the frame.* (Perspectives, 21(3), 412-426. doi:10.1080/0907676X.2012.693108)

Fryer, L., Pring, L., & Freeman, J. (2013). *Audio drama and the imagination: The*

influence of sound effects on presence in people with and without sight. (Journal of Media Psychology, 25(2), 65-71. doi:10.1027/1864-1105/a000084)

Fuentes Luque, A. (2003, November). An Empirical Approach to the Reception of AV Translated Humour. *The Translator*, *9*(2), 293–306. Retrieved 2022-01-19, from https://doi.org/10.1080/13556509.2003.10799158 (Publisher: Routledge _eprint: https://doi.org/10.1080/13556509.2003.10799158) doi: 10.1080/13556509.2003.10799158

Gambier, Y. (2019, February). Audiovisual translation and reception. *Slovo.ru: Baltic accent*, *10*, 62–68. doi: 10.5922/2225-5346-2019-1-4

Ghadiyaram, D., & Bovik, A. (2015, November). Massive Online Crowdsourced Study of Subjective and Objective Picture Quality. *IEEE Transactions on Image Processing*, *25*, 1–1. doi: 10.1109/TIP.2015.2500021

Giovanni, E. D. (2012). Italians and television: a comparative study on the reception of subtitles and voice over. *undefined*. Retrieved 2022-01-19, from https://www.semanticscholar.org/paper/Italians-and-television%3A-a-comparative-study-on-the-Giovanni/4ca109a59a88475dfdf1342c401f692ed39b22ca

Giovanni, E. D. (2016, February). Dubbing and Redubbing Animation: Disney in the Arab World. *Altre Modernità*, 92–106. Retrieved 2021-02-05, from https://riviste.unimi.it/index.php/AMonline/article/view/6850 doi: 10.13130/2035-7680/6850

Giovanni, E. D., & Fresco, P. R. (2019). Are we all together across languages?: An eye tracking study of original and dubbed films. In *Reassessing Dubbing: Historical approaches and current trends, 2019, ISBN 9789027262271, pp. 126-144* (pp. 126–144). John Benjamins. Retrieved 2022-01-19, from https://dialnet.unirioja.es/servlet/articulo?codigo=7794010 (Section: Reassessing Dubbing: Historical approaches and current trends)

Gottlieb, H. (1997). *Subtitling, translation and idioms.* ((Unpublished doctoral dissertation). University of Copenhagen, Copenhagen, Denmark.)

Grillo, V., & Kawin, B. (1988). *Reading at the movies: Subtitles, silence and the structure of the brain.* (Post Script: Essays in Film and Humanities, 1, pp. 25-32.)

Hands, D. S., & Avons, S. E. (2001). Recency and duration neglect in subjective assessment of television picture quality. *Applied Cognitive Psychology*, *15*(6), 639–657. Retrieved 2021-01-29, from https://onlinelibrary.wiley.com/doi/abs/10.1002/acp.731 (Number: 6 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/acp.731) doi: https://doi.org/10.1002/acp.731

Häsk. (2009, April). *English: Dubbing films in Europe:.* Retrieved 2022-01-12, from https://commons.wikimedia.org/wiki/File:Dubbing_films_in_Europe.png

Herbst, T. (1997). Dubbing and the Dubbed text - style and cohesion. In *Text typology and translation* (p. 291-). doi: 10.1075/btl.26.21her

Hoßfeld, T., Hirth, M., Redi, J., Mazza, F., Korshunov, P., Naderi, B., ... Keimel, C. (2014). *Best Practices and Recommendations for Crowdsourced QoE - Lessons learned from the Qualinet Task Force "Crowdsourcing"*. Retrieved 2021-01-29, from https://hal.archives-ouvertes.fr/hal-01078761

Huber, L., & Kairys, A. (2021, July). Culture Specific Items in Audiovisual Translation: Issues of Synchrony and Cultural Equivalence in the Lithuanian Dub of "Shrek the Third". *Studies about Languages*(38), 5–16. Retrieved 2021-11-05, from https://kalbos.ktu.lt/index.php/KStud/article/view/24743 (Number: 38) doi: 10.5755/j01.sal.1.38.24743

Janowski, L. (2019, June). Evaluating experiment design with unrepeated scenes for video quality subjective assessment. (NA). Retrieved 2021-11-09, from https://www.its.bldrdoc.gov/publications/details.aspx?pub=3240 (Publisher: ITS) doi: NA

Janowski, L., & Pinson, M. (2015, December). The Accuracy of Subjects in a Quality Experiment: A Theoretical Subject Model. *IEEE Transactions on Multimedia*, *17*(12), 2210–2224. (Number: 12 Conference Name: IEEE Transactions on Multimedia) doi: 10.1109/TMM.2015.2484963

Joskowicz, J., Sotelo, R., Juayek, M., Durán, D., & Garella, J. P. (2014). Automation of Subjective Video Quality Measurements. In *Proceedings of the Latin America Networking Conference on LANC 2014 - LANC '14* (pp. 1–5). Montevideo, Uruguay: ACM Press. Retrieved 2021-01-28, from http://dl.acm.org/citation.cfm?doid=2684083.2684090 doi: 10.1145/2684083.2684090

Karakanta, A., Bhattacharya, S., Nayak, S., Baumann, T., Negri, M., & Turchi, M. (2020). The Two Shades of Dubbing in Neural Machine Translation. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 4327–4333). Barcelona, Spain (Online): International Committee on Computational Linguistics. Retrieved 2021-01-28, from https://www.aclweb.org/anthology/2020.coling-main.382 doi: 10.18653/v1/2020.coling-main.382

Kizeweter, M. (2015). *Voices about polish voices in foreign films: Using an internet forum as a source of information about the opinions of polish viewers on dubbing as a mode of avt.* (In Ł. Bogucki & M. Deckert (Eds.), Accessing audiovisual translation (pp. 163-178). Frankfurt: Peter Lang.)

Knapp, K. (2019). *CROSSLINGUISTIC INFLUENCE IN SECOND LANGUAGE ACQUISITION - Phonological Transfer in German Learners of English* (Unpublished doctoral dissertation). Karl-Franzens-Universität Graz.

Koolstra, C. M., Peeters, A. L., & Spinhof, H. (2002, September). The Pros and Cons of Dubbing and Subtitling. *European Journal of Communication*, *17*(3), 325–354. Retrieved 2022-01-19, from http://journals.sagepub.com/doi/10.1177/

0267323102017003694 doi: 10.1177/0267323102017003694

Koveriene, I. (2015). Dubbing as an Audiovisual Translation Mode: English and Lithuanian Phonemic Inventories in the Context of Visual Phonetics. *undefined*. Retrieved 2022-01-13, from https://www.semanticscholar.org/paper/Peculiarities-of-Multilingual-Films-in-the-Context-Alosevi%C4%8Dien%C4%97/4dd378faff10fa47539110636bbc261ef0f24f65

Koverienė, I., & Čeidaitė, K. (2020, October). Lip Synchrony of Rounded and Protruded Vowels and Diphthongs in the Lithuanian-Dubbed Animated Film "Cloudy with a Chance of Meatballs 2". *Respectus Philologicus*(38(43)), 214–229. Retrieved 2021-01-28, from https://www.journals.vu.lt/respectus-philologicus/article/view/17155 (Number: 38(43)) doi: 10.15388/RESPECTUS.2020.38.43.69

Kraft, S., & Zölzer, U. (2014). *BeaqleJS: HTML5 and JavaScript based Framework for the Subjective Evaluation of Audio Quality.*

Kruger, J.-L. (2012). *Making meaning in avt: Eye tracking and viewer construction of narrative.* (Perspectives, 20(1), 67-86. doi:10.1080/0907676X.2011.632688)

Kumcu, A., Bombeke, K., Platisa, L., Jovanov, L., Van Looy, J., & Philips, W. (2017, February). Performance of Four Subjective Video Quality Assessment Protocols and Impact of Different Rating Preprocessing and Analysis Methods. *IEEE Journal of Selected Topics in Signal Processing*, *11*(1), 48–63. Retrieved 2021-01-28, from http://ieeexplore.ieee.org/document/7781646/ (Number: 1) doi: 10.1109/JSTSP.2016.2638681

Li, Z., & Bampis, C. G. (2017, April). Recover Subjective Quality Scores from Noisy Measurements. In *2017 Data Compression Conference (DCC)* (pp. 52–61). (ISSN: 2375-0359) doi: 10.1109/DCC.2017.26

Lin, J. Y., Song, R., Wu, C.-H., Liu, T., Wang, H., & Kuo, C. C. J. (2015, July). MCL-V: A streaming video quality assessment database. *Journal of Visual Communication and Image Representation*, *30*, 1–9. Retrieved 2021-01-29, from http://www.sciencedirect.com/science/article/pii/S1047320315000425 doi: 10.1016/j.jvcir.2015.02.012

Mangiron, C. (2016). *Reception of game subtitles: An empirical study.* (The Translator, 22(1), 72-93. doi:10.1080/13556509.2015.1110000)

Mayoral, R., Kelly, D., & Gallardo, N. (1988). Concept of Constrained Translation. Non-Linguistic Perspectives of Translation.
doi: 10.7202/003608AR

Mazur, I., & Chmiel, A. (2012, March). Towards common European audio description guidelines: Results of the Pear Tree Project. *Perspectives-studies in Translatology*, *20*, 5–23. doi: 10.1080/0907676X.2011.632687

McGurk, H., & MacDonald, J. (1976, December). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748. Retrieved 2021-01-29, from https://www.nature.com/

articles/264746a0 (Number: 5588 Publisher: Nature Publishing Group) doi: 10.1038/264746a0

Mendonça, C., & Delikaris-Manias, S. (2018). *Statistical tests with MUSHRA data.*

Mereu Keating, C. (2016). *The Politics of Dubbing.* Retrieved 2022-01-20, from https://www.peterlang.com/document/1053381 (ISSN: 1664-249X)

Min, X., Zhai, G., Zhou, J., Farias, M. C. Q., & Bovik, A. C. (2020). Study of Subjective and Objective Quality Assessment of Audio-Visual Signals. *IEEE Transactions on Image Processing*, *29*, 6054–6068. Retrieved 2021-01-28, from https://ieeexplore.ieee.org/document/9075375/ doi: 10.1109/TIP.2020.2988148

Miquel Iriarte, M. (2014). *The reception of subtitling by the deaf and hard of hearing. preliminary findings.* (In E. Torres-Simón & D. Orrego-Carmona (Eds.), Translation research projects 5 (pp. 63-76). Tarragona: Universitat Rovira i Virgili.)

Mohammed, H., & Färber, N. (2010). *Subjective evaluation of Hierarchical B-Frames using video-MUSHRA.* (Journal Abbreviation: 28th Picture Coding Symposium, PCS 2010 Pages: 449 Publication Title: 28th Picture Coding Symposium, PCS 2010) doi: 10.1109/PCS.2010.5702532

Moorthy, A. K., Choi, L. K., Bovik, A. C., & Veciana, G. d. (2012, October). Video Quality Assessment on Mobile Devices: Subjective, Behavioral and Objective Studies. *IEEE Journal of Selected Topics in Signal Processing*, *6*(6), 652–671. (Number: 6 Conference Name: IEEE Journal of Selected Topics in Signal Processing) doi: 10.1109/JSTSP.2012.2212417

Nayak, S., Baumann, T., Bhattacharya, S., Karakanta, A., Negri, M., & Turchi, M. (2020, October). See me Speaking? Differentiating on Whether Words are Spoken On Screen or Off to Optimize Machine Dubbing. In *Companion Publication of the 2020 International Conference on Multimodal Interaction* (pp. 130–134). Virtual Event Netherlands: ACM. Retrieved 2021-01-28, from https://dl.acm.org/doi/10.1145/3395035.3425640 doi: 10.1145/3395035.3425640

O'Hagan, M. (2009). *Towards a cross-cultural game design: An explorative study in understanding the player experience of a localised japanese video game.* (The Journal of Specialised Translation, 11, 211-233.)

Orrego-Carmona, D. (2014). *Subtitling, video consumption and viewers: The impact of the young audience.* (Translation Spaces, 3, 51-70. doi:10.1075/ts.3.03orr)

Orrego-Carmona, D. (2016). *A reception study on non-professional subtitling: Do audiences notice any difference?* (Across Languages and Cultures, 17(2), 163-181. doi:10.1556/084.2016.17.2.2)

Peeters, A. L., Scherpenzeel, A. C., & Zantinge, J. H. (1988). *Ondertiteling of nasynchronisatie van kinderprogramma's [subtitling or dubbing children's programs].*

Perego, E. (2016). History, development, challenges and opportunities of empirical research in audiovisual translation. *Across Languages and Cultures*, *17*, 155–162. doi: 10.1556/084.2016.17.2.1

Perego, E. (2018, November). Cross-national research in audiovisual translation: Some methodological considerations. *Journal of Audiovisual Translation*, *1*(1), 64–80. Retrieved 2021-02-05, from https://www.jatjournal.org/index.php/jat/article/view/44 (Number: 1) doi: 10.47476/jat.v1i1.44

Perego, E., Del Missier, F., & Bottiroli, S. (2015, January). Dubbing versus subtitling in young and older adults: cognitive and evaluative aspects. *Perspectives*, *23*(1), 1–21. Retrieved 2022-01-19, from https://doi.org/10.1080/0907676X.2014.912343 (Publisher: Routledge _eprint: https://doi.org/10.1080/0907676X.2014.912343) doi: 10.1080/0907676X.2014.912343

Perego, E., Laskowska, M., Matamala, A., Remael, A., Robert, I. S., Szarkowska, A., ... Bottiroli, S. (2016, December). Is subtitling equally effective everywhere? A first cross-national study on the reception of interlingually subtitled messages. *Across Languages and Cultures*, *17*(2), 205–229. Retrieved 2021-02-05, from https://akjournals.com/view/journals/084/17/2/article-p205.xml (Number: 2 Publisher: AkadÃľmiai KiadÃş Section: Across Languages and Cultures) doi: 10.1556/084.2016.17.2.4

Pinson, M. H. (2011, November). Audiovisual Quality Components: An Analysis. (NA). Retrieved 2021-11-09, from https://www.its.bldrdoc.gov/publications/details.aspx?pub=2565 (Publisher: ITS)

Pinson, M. H., Janowski, L., Pepion, R., Huynh-Thu, Q., Schmidmer, C., Corriveau, P., ... Ingram, W. (2012, October). The Influence of Subjects and Environment on Audiovisual Subjective Tests: An International Study. *IEEE Journal of Selected Topics in Signal Processing*, *6*(6), 640–651. (Number: 6 Conference Name: IEEE Journal of Selected Topics in Signal Processing) doi: 10.1109/JSTSP.2012.2215306

Pinson, M. H., & Wolf, S. (2003, June). Comparing subjective video quality testing methodologies. In *Visual Communications and Image Processing 2003* (Vol. 5150, pp. 573–582). SPIE. Retrieved 2021-12-15, from https://www.spiedigitallibrary.org/conference-proceedings-of-spie/5150/0000/Comparing-subjective-video-quality-testing-methodologies/10.1117/12.509908.full doi: 10.1117/12.509908

Prajwal, K. R., Mukhopadhyay, R., Namboodiri, V., & Jawahar, C. V. (2020, October). A Lip Sync Expert Is All You Need for Speech to Lip Generation In the Wild. *28th ACM International Conference on Multimedia (ACM MM)*. Retrieved 2021-03-07, from https://researchportal.bath.ac.uk/en/publications/a-lip-sync-expert-is-all-you-need-for-speech-to-lip-generation-in (Publisher: Association for Computing Machinery) doi: 10.1145/3394171.3413532

Raffi, F. (2020, June). The Impact of Italian Dubbing on Viewers' Immersive Experience: An Audience Reception Study. *Online Journal of Communication and Media Technologies*, *10*(3), e202019. Retrieved 2021-10-30, from https://www.ojcmt.net/

article/the-impact-of-italian-dubbing-on-viewers-immersive-experience-an-audience-reception-study-8371 (Publisher: Bastas) doi: 10.30935/ojcmt/8371

Rainer, B., Petscharnig, S., Timmerer, C., & Hellwagner, H. (2015, May). Is one second enough? Evaluating QoE for inter-destination multimedia synchronization using human computation and crowdsourcing. In *2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX)* (pp. 1–6). Pylos-Nestoras: IEEE. Retrieved 2021-01-28, from http://ieeexplore.ieee.org/document/7148107/ doi: 10.1109/QoMEX.2015.7148107

Ramos, M. (2015). *The emotional experience of films: Does audio description make a difference?* (The Translator, 21(1), 68-94. doi:10.1080/13556509.2014.994853)

Ramos, M. (2016). *Testing audio narration: The emotional impact of language in audio description.* (Perspectives, 24(4), 606-634. doi:10.1080/0907676X.2015.1120760)

Ramos, M., & Rojo, A. (2014). *"feeling" audio description: Exploring the impact of ad on emotional response.* (Translation Spaces, 3, 133-150. doi:10.1075/ts.3.06ram)

Ranzato, I., & Zanotti, S. (2019). *Reassessing Dubbing: Historical approaches and current trends*. John Benjamins Publishing Company. (Google-Books-ID: _LGoDwAAQBAJ)

Reiter, U. (2011, October). Quality of experience: a buzzword or the key to successful multimedia delivery across networks? In *Proceedings of the 6th Latin America Networking Conference* (pp. 20–24). New York, NY, USA: Association for Computing Machinery. Retrieved 2021-01-29, from https://doi.org/10.1145/2078216.2078220 doi: 10.1145/2078216.2078220

Reyes Lozano, J. d. l. (2015). *La traduction du cinéma pour les enfants: Une étude sur la réception [the translation of cinema for children: A reception-oriented study].* ((Unpublished doctoral dissertation). Jaume I University, Spain.)

Riniolo, T. C., & Capuana, L. J. (2020, July). Directly comparing subtitling and dubbing using Netflix: Examining enjoyment issues in the natural setting. *Current Psychology*. Retrieved 2021-02-05, from https://doi.org/10.1007/s12144-020-00948-1 doi: 10.1007/s12144-020-00948-1

Romero-Fresco, P. (2013, January). Accessible filmmaking: Joining the dots between audiovisual translation, accessibility and filmmaking. *The Journal of Specialised Translation*, 201–223.

Romero-Fresco, P. (2016). *The Reception of Subtitles for the Deaf and Hard of Hearing in Europe: UK, Spain, Italy, Poland, Denmark, France and Germany.*

Romero-Fresco, P. (2018, November). In support of a wide notion of media accessibility: Access to content and access to creation. *Journal of Audiovisual Translation*, *1*(1), 187–204. Retrieved 2021-01-28, from http://jatjournal.org/index.php/jat/article/view/53 (Number: 1) doi: 10.47476/jat.v1i1.53

Romero-Fresco, P. (2019, May). The dubbing effect: An eye-tracking study on how viewers make dubbing work. *The Journal of Specialised Translation*.

Saboo, A., & Baumann, T. (2019). Integration of Dubbing Constraints into Machine Translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)* (pp. 94–101). Florence, Italy: Association for Computational Linguistics. Retrieved 2021-01-28, from https://www.aclweb.org/anthology/W19-5210 doi: 10.18653/v1/W19-5210

Saha, D., Nayak, S., & Baumann, T. (2022). *Merkel podcast corpus: A multimodal dataset compiled from 16 years of angela merkel's weekly video podcasts.* (LREC paper.)

Sánchez-Mompeán, S. (2020, January). Dubbing and Prosody at the Interface. In *The prosody of dubbed speech* (pp. 19–88). doi: 10.1007/978-3-030-35521-0_2

Sánchez-Mompeán, S. (2021, September). Netflix likes it dubbed: Taking on the challenge of dubbing into English. *Language & Communication*, *80*, 180–190. Retrieved 2021-10-30, from https://www.sciencedirect.com/science/article/pii/S0271530921000562 doi: 10.1016/j.langcom.2021.07.001

Schmidt, C., Koller, O., Ney, H., Hoyoux, T., & Piater, J. (2013). *Using Viseme Recognition to Improve a Sign Language Translation System.*

Seshadrinathan, K., Soundararajan, R., Bovik, A., & Cormack, L. (2010). *A Subjective Study to Evaluate Video Quality Assessment Algorithms.* doi: 10.1117/12.845382

Shahid, M. (2014). *Methods for objective and subjective video quality assessment and for speech recognition.* Karlskrona: Department of Applied Signal Processing, Blekinge Institute of Technology. Retrieved 2021-01-28, from http://urn.kb.se/resolve?urn=urn:nbn:se:bth-00603 (OCLC: 941390291)

Shdaifat, I., Grigat, R.-R., & Lütgert, S. (2001). *Viseme recognition using multiple feature matching.* (Pages: 2434)

Sinno, Z., & Bovik, A. C. (2018, October). Large Scale Subjective Video Quality Study. In *2018 25th IEEE International Conference on Image Processing (ICIP)* (pp. 276–280). Athens: IEEE. Retrieved 2021-01-28, from https://ieeexplore.ieee.org/document/8451467/ doi: 10.1109/ICIP.2018.8451467

Sinno, Z., & Bovik, A. C. (2019, February). Large-Scale Study of Perceptual Video Quality. *IEEE Transactions on Image Processing*, *28*(2), 612–627. Retrieved 2021-01-28, from https://ieeexplore.ieee.org/document/8463581/ (Number: 2) doi: 10.1109/TIP.2018.2869673

Smith, T., & Henderson, J. (2008, January). Attentional Synchrony in Static and Dynamic Scenes. *Journal of Vision*, *8*, 773. doi: 10.1167/8.6.773

Sokolovsky, Y. V. (2010). On the Linguistic Definition of Translation. *undefined*. Retrieved 2022-01-13, from https://www.semanticscholar.org/paper/On-the-Linguistic-Definition-of-Translation-Sokolovsky/b08bccc1d956ed35b5d1c5f89d7e9972cd3532ae

Sotelo, R. (2015, October). Subjective and objective video quality assessment. A parametric model for digital television coded with H.264. In *2015 CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies*

(CHILECON) (pp. 437–440). Santiago, Chile: IEEE. Retrieved 2021-01-28, from http://ieeexplore.ieee.org/document/7400414/ doi: 10.1109/Chilecon.2015.7400414

Sotelo, R., Joskowicz, J., Anedda, M., Murroni, M., & Giusto, D. D. (2017, June). Subjective video quality assessments for 4K UHDTV. In *2017 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)* (pp. 1–6). Cagliari, Italy: IEEE. doi: 10.1109/BMSB.2017.7986225

Szarkowska, A., & Jankowska, A. (2012). *Text-to-speech audio description of voiced-over films: A case study of audio described volver in polish.* (In E. Perego (Ed.), Emerging topics in translation: Audio description (pp. 81-98). Trieste: Università di Trieste.)

Szarkowska, A., Krejtz, I., Pilipczuk, O., Dutka, Å., & Kruger, J.-L. (2016). *The effects of text editing and subtitle presentation rate on the comprehension and reading patterns of interlingual and intralingual subtitles among deaf, hard of hearing and hearing viewers.* (Across Languages and Cultures, 17(2), 183-204. doi:10.1556/084.2016.17.2.3)

Szarkowska, A., & Laskowska, M. (2015). *Poland - a voice-over country no more? a report on an online survey on subtitling preferences among polish hearing and hearing-impaired viewers.* (In Ł. Bogucki & M. Deckert (Eds.), Accessing audiovisual translation (pp. 179-197). Frankfurt: Peter Lang.)

Topiwala, P., Dai, W., & Pian, J. (2020, September). Deep learning and video quality analysis: towards a unified VQA. In *Applications of Digital Image Processing XLIII* (Vol. 11510, p. 115100V). International Society for Optics and Photonics. Retrieved 2021-01-29, from https://www.spiedigitallibrary.org/conference-proceedings-of-spie/11510/115100V/Deep-learning-and-video-quality-analysis--towards-a-unified/10.1117/12.2571309.short doi: 10.1117/12.2571309

Tuominen, T. (2012). *The art of accidental reading and incidental listening: An empirical study on the viewing of subtitled films.* ((Unpublished doctoral dissertation). University of Tampere, Tampere, Finland.)

Varela, F. C. (2007, December). Quality standards in dubbing: a proposal. *Tradterm*, *13*, 71. Retrieved 2021-01-28, from http://www.revistas.usp.br/tradterm/article/view/47466 doi: 10.11606/issn.2317-9511.tradterm.2007.47466

Vöge, H. (1977). *The translation of films: Sub-titling versus dubbing.* (In: Babel, 23 (3), pp. 120-125. https://doi.org/10.1075/babel.23.3.05vog)

Völker, C., Bisitz, T., Huber, R., & Ernst, S. (2015). *Adaptions for the MUlti Stimulus test with Hidden Reference and Anchor (MUSHRA) for elder and technical unexperienced participants.*

Völker, C., Bisitz, T., Huber, R., Kollmeier, B., & Ernst, S. M. A. (2018, May). Modifications of the MUlti stimulus test with Hidden Reference and Anchor (MUSHRA) for use in audiology. *International Journal of Audiology*, *57*(sup3), S92–S104. Retrieved 2021-01-29, from https://doi.org/10.1080/

14992027.2016.1220680 (Number: sup3 Publisher: Taylor & Francis _-eprint: https://doi.org/10.1080/14992027.2016.1220680) doi: 10.1080/14992027.2016.1220680

Whitman-Linsen, C. (1992). *Through the dubbing glass.* (Frankfurt: Peter Lang.)

Wickham, H. (2014, September). Tidy data. *The American Statistician*, *14*. doi: 10.18637/jss.v059.i10

Widler, B. (2004). *A survey among audiences of subtitled films in viennese cinemas.* (Meta: Journal des Traducteurs, 49(1), 98-101. doi:10.7202/009025ar)

Wilken, N., & Kruger, J.-L. (2016). *Putting the audience in the picture: Mise-en-shot and psychological immersion in audio described film.* (Across Languages and Cultures, 17(2), 251-270. doi:10.1556/084.2016.17.2.6)

Wissmath, B., Weibel, D., & Groner, R. (2009, January). Dubbing or Subtitling?: Effects on Spatial Presence, Transportation, Flow, and Enjoyment. *Journal of Media Psychology: Theories, Methods, and Applications*, *21*, 114–125. doi: 10.1027/1864-1105.21.3.114

Yang, Y., Shillingford, B., Assael, Y., Wang, M., Liu, W., Chen, Y., ... de Freitas, N. (2020, November). Large-scale multilingual audio visual dubbing. *arXiv:2011.03530 [cs, eess]*. Retrieved 2021-01-28, from http://arxiv.org/abs/2011.03530 (arXiv: 2011.03530)

Zabalbeascoa, P. (1997). *Dubbing and the nonverbal dimension of translation.* (In: Nonverbal Communication in Translation: New Perspectives and Challenges in Literature, Interpretation and the Media. Ed. F. Poyatos. Amsterdam / Philadelphia: John Benjamins, pp. 327-342. https://doi.org/10.1075/ btl.17.26zab)

Zanotti, S. (2013, February). Censorship or Profit? The Manipulation of Dialogue in Dubbed Youth Films. *Meta*, *57*(2), 351–368. Retrieved 2022-01-20, from http://id.erudit.org/iderudit/1013950ar doi: 10.7202/1013950ar

Zieliński, S., Hardisty, P., Hummersone, C., & Rumsey, F. (2007, January). Potential biases in MUSHRA listening tests. *Audio Engineering Society - 123rd Audio Engineering Society Convention 2007*, *2*.

Declaration on oath

I hereby declare on oath that I have completed this work independently and without outside help and that I have not used any resources other than those specified in the attached bibliography. All passages that were taken verbatim or analogously from publications are marked as such. I further assure that I have not previously submitted the work in another examination procedure and that the written version submitted corresponds to that on the electronic storage medium.
I consent to this work being placed in the library of the department.

Eidesstattliche Versicherung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit im Bachelorstudiengang Informatik selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel – insbesondere keine im Quellenverzeichnis nicht benannten Internet-Quellen – benutzt habe. Alle Stellen, die wörtlich oder sinngemäß aus Veröffentlichungen entnommen wurden, sind als solche kenntlich gemacht. Ich versichere weiterhin, dass ich die Arbeit vorher nicht in einem anderen Prüfungsverfahren eingereicht habe und die eingereichte schriftliche Fassung der elektronischen Abgabe entspricht.
Ich bin mit einer Einstellung in den Bestand der Bibliothek des Fachbereiches einverstanden.

Hamburg, den _____ Unterschrift: _____