Proceedings of the international Paris-Cologne seminar on

Variability in Speech Production and Perception

2022

Foreword

This volume presents virtual conference papers which document the outcome of the transnational seminar between Cologne (Universität zu Köln, IfL Phonetics) and Paris (Université de Paris). The seminar took place online and was supported by the EduVEnture Cologne (IVAC) program funded by the DAAD.

The seminar supported a strong knowledge transfer between the students of the two universities and was a very enjoyable experience overall. In mixed Cologne-Paris groups, several projects were created, dealing with the topic of variability in speech production and perception.

The experiments themselves were planned and possible outcomes were derived based on research literature and hypotheses; however, they were not carried out in the real world of experimental linguistics – not yet.

Doris Mücke

Ioana Chitoran

Simon Roessig

Cross-language speech perception
– Non-native perception of labial-velar consonants

Elisa Herbig[1], Nlabephee Kefas Othaniel[2]

[1]*Institute of Linguistics, University of Cologne, Germany*
[2]*M2: Phonetics and Phonology, University of Paris*
eherbig@smail.uni-koeln.de, othaniel_nlabephee@linguamail.org

Abstract

The phoneme inventory of one's native language has been said to greatly influence or even impede cross-linguistic speech perception. This leads to the question whether listeners can discriminate and how they assimilate non-native contrasts. Investigating non-native perception of labial-velar consonants that are common in Dza but do not occur in German, we predict differing results for the contrasts voiced labial-velar vs. velar plosive (/k͡p/ - /k/) and voiceless labial-velar vs. velar plosive (/g͡b/ -/g/) as compared to the contrasts voiced labial-velar vs. labial plosive (/k͡p/ - /p/) and voiceless labial-velar vs. labial plosive (/g͡b/ - /b/). Based on Best's (1995) PAM model, the former show a tendency towards being assimilated as TC (Two Category), while the latter, as hypothesised, show a tendency towards being assimilated as CG (Category Goodness).

Keywords: speech perception, cross-language, native and non-native speech inventory, assimilation

1. Introduction

According to Best et al. (2001: 775) "it has been assumed that mature listeners have difficulty discriminating phonetic distinctions that do not occur as a native phonological contrast." To take a step back, it seems valuable to have a look at the organization of the phonetic space. A differentiation is made here between the "universal phonetic domain", which includes all gestures that the human vocal tract allows for, and individual, language specific "native phonological spaces" (Best 1995: 186-187). The latter could be described as selections of gestures from the universal domain forming the languages' individual phoneme inventories (Best 1995: 189). While Best (1995: 193) does attest "a great amount of overlap among languages", she also asks one important question: "If phonetic implementations are language-specific rather than universal, then what exactly does the mature listener perceive in non-native phones and contrasts" (1995: 184).

This issue has led to the topic of the current paper where we want to consider non-native perception of labial-velar consonants as an example of a problem arising in cross-linguistic speech perception due to the phoneme inventory of the native language. Labial-velar consonants occur frequently in African languages but are rare to non-existent in other languages around the world, particularly in European languages (Connell 1994: 441). According to Westermann and Ward (1933, as cited in Connell 1994: 442) "the labial element is perceptually more salient for Europeans, [while] for Africans it is the velar element which predominates". These characteristics of labial-velar consonants make them worth investigating within the framework of cross-language speech perception. Specifically, in this paper we want to compare Dza, an Adamawa language of Taraba State Nigeria, which has both the voiced and the voiceless labial-velar stops /k͡p/ and /g͡b/ in its phoneme inventory and German, which does not employ labial-velar consonants. We will look at these two languages in more detail in the following section.

The contrasts in question are /k͡p/ - /k/ and /k͡p/ - /p/ as well as /g͡b/ -/g/ and /g͡b/ - /b/, whereby all six plosives (labial-velar, labial, and velar) from part of the Dza phoneme inventory but only labial and velar consonants occur in German. We now want to investigate how German native speakers assimilate the Dza doubly articulated plosives when they appear in places where labial or velar plosives would be expected. Our investigation will be based on Best's (1995) Perceptual Assimilation Model (PAM) as explained in the following section.

1.1. Dza vs. German

The two languages we want to compare in this study are Dza and German, whose consonant inventories differ amongst others regarding their plosives (cf. Table 1 & 2). Dza belongs to the Jen language cluster, "an area of high linguistic diversity", which consists of ten Adamawa language varieties in total (Norton, Othaniel 2020: 18-19). Its phoneme inventory includes the doubly articulated labial-velar plosives (/k͡p/ and /g͡b/) which are often described as being simultaneously articulated. However, Connell (1994: 446) adds that the two "component gestures are not simultaneous in the strict sense of the word", but rather the velar gesture precedes the labial one.

Table 1: *Consonant chart of the Jen languages (Norton, Othaniel 2020: 46).*

Proto-Jen consonants

	labial	alveolar	postalveolar	velar	labio-velar	glottal
plosives	*p *b	*t *d	*c	*k *g	(*k͡p) *g͡b	
affricates		*t͡s *d͡z	*t͡ʃ *d͡ʒ			
fricatives	*f *v	*s *z	*ʃ			(*h)
nasals	*m	*n	*ɲ			
implosives	*ɓ	*ɗ				
trill		*r				
approxi-mants		*l	*y *ɥ		*w	

Table 2: *Consonant chart of German (Dahmen, Weth 2017: 34).*

	Bilabial	Labio-dental	Alveolar	Postal-veolar	Palatal	Velar	Uvular	Glottal
Plosiv	p b		t d			k g		
Nasal	m		n			ŋ		
Frikativ		f v	s z	ʃ ʒ	ç		χ ʁ	h
(Affrikate)	(pf)		(ts)	(tʃ)				
Approxi-mant					j			
Lateral			l					

In comparison, the German consonant inventory contains the voiceless bilabial plosive /p/, the voiced bilabial plosive /b/, the voiceless velar plosive /k/ and the voiced velar plosive /g/. So, while German does not employ simultaneous gestural constellations of the lips and the velum as phonological contrasts, individually, these gestures can be found within the boundaries of the language's native phonological space (similarly, see Best 1995: 192). According to Best (1995: 194), the "[s]imilarity between non-native segments and native gestural constellations, as indexed by the spatial proximity of constriction locations and active articulators and by similarities in constriction degree and gestural phasing, are predicted to determine listeners' perceptual assimilation of the non-native phones to native categories". Considering this, we can assume that native German listeners will be able to perceive similarities and differences between the native and non-native phonemes in question to a certain degree.

1.2. PAM (Best 1995)

As mentioned at the beginning of this paper, it has been assumed that listeners have difficulty discriminating phonetic distinctions that do not occur as a native phonological contrast. It seems, however, that this assumption cannot be upheld entirely. Instead, Best (1995) proposes her Perception Assimilation Model (PAM) which categorizes assimilation into different levels (also cf. Best et al. 2001: 776). The three main ones are:

1. Assimilation to a native category, which can range between being a good or bad exemplar of this category.
2. Assimilation to an uncategorizable speech sound.
3. Perception as a non-speech sound.

(cf. Best 1995: 194-195)

When looking at non-native contrasts, Best (1995: 195) further distinguishes between Two-Category Assimilation (TC) (both non-native segments are assimilated to different native categories), Category-Goodness Difference (CG) (both non-native sounds are assimilated to the same native category, but the goodness of fit differs), Single-Category Assimilation (SC) (both non-native sounds are assimilated to the same native category which they fit equally well), both Uncategorizable (UU) (both non-native sounds are within the phonetic space but are not equivalent to a particular native category), Uncategorized versus Categorized (UC) (one non-native sound is assimilated to a native category, the other is within the phonetic space, but outside native categories) and Nonassimilable (NA) (both non-native categories are perceived as nonspeech sounds).

With the Perception Assimilation Model in mind, we hypothesise that the contrasts in question, /k͡p/ - /k/ and /k͡p/ - /p/ as well as /g͡b/ -/g/ and /g͡b/ - /b/, will be assimilated as a Category-Goodness Difference (CG). We thereby assume, that German native speakers will not be able to clearly discriminate between the labial-velar consonants and their labial or velar plosive counterparts, but that they will perceive them as better (labial and velar plosives) or worse (doubly articulated consonants) exemplars of the native category. As a sidenote, based on the assumption that labial elements are perceptually more salient for Europeans, we hypothesise that the listeners will rather assimilate the labial-velar plosives to /p/ and /b/ than to /k/ and /g/ respectively.

2. Method

To test our hypothesis, we will adapt Best et al.'s (2001: 782) experimental design. The experiment will consist of two parts: To begin with, the German native speaking participants will be presented with a discrimination task in the form of AXB stimuli where X is either equivalent to A or B. Then, participants will be asked "to write down what the [word] sounded like to them, using [German] orthography (i.e., ''spell as you would in [German])" (Best et al. 2001: 782). In contrast to Best et al. (2001) we will use entire German sounding pseudowords instead of syllable stimuli.

2.1. Participants

The (potential) listeners are 25 native speakers of German. 15 of them self-identify as female, 10 as male. The mean age is 23 years, ranging from 21 to 27. They were recruited through the Linguistics department at the University of Cologne and will receive course credits for their participation. None of the participants has previously studied Dza or any other language that has labial-velar consonants in its phoneme inventory. None of the participants reports any hearing or visual impairments.

2.2. Stimulus materials

An adult male native Dza speaker from Taraba State, North Central Nigeria, will be recorded producing the language stimuli. These consist of German sounding pseudoword pairs that are constructed so that the onset of the first syllable contains one of the critical phonemes respectively. We opted for pseudowords instead of actual words to avoid the Ganong effect, which describes how lexical status affects the interpretation of acoustic input in favour of real words (cf. Ganong 1980). The reason behind placing the critical contrast in this position is that "labial-velars rarely occur finally" (Connell 1994: 468), that the onset position will draw the listeners' attention to it and the fact that this positioning seems most natural to our native Dza speaker. An example for each contrast is shown below. In total, there are 60 different pseudoword pair for each contrast.

Table 3: *Example stimuli.*

voiced labial-velar vs. velar plosive	**voiced labial-velar vs. labial plosive**	**voiceless labial-velar vs. velar plosive**	**voiceless labial-velar vs. labial plosive**
/k͡p/ - /k/	/k͡p/ - /p/	/g͡b/ -/g/	/g͡b/ - /b/
[k͡palsus] – [kalsus]	[k͡palsus] – [palsus]	[g͡balsus] – [galsus]	[g͡balsus] – [balsus]

2.3. Procedure

The first part of the experiment consists of a discrimination task targeting the assimilation patterns. Participants will be presented with the pre-recorded stimuli in an AXB sequence with X being equivalent to either A or B. An example for such a sequence could be A = [k͡palsus], X = [k͡palsus], B = [palsus] or alternatively: A = [k͡palsus], X = [palsus], B = [palsus] (voiced labial-velar vs. labial plosive contrast). The

experimental design will follow the same pattern for each contrast. Each session will consist of 60 trials. The design is within-subject and within-item as every participant will hear all four contrasts, while every pseudoword pair will appear only once in one of the contrasts. The pairing of pseudoword and contrast as well as the order of the stimuli will be randomized, the only restriction being that the same contrast cannot appear twice in a row. Additionally, the AXB (AAB, ABB, BBA or BAA) sequence will be randomized. All stimuli, contrasts and sequences will be equally represented. The participants' answers will be recorded by having them press A (if to them, X sounds like A) or B (if X sounds like B) on a keyboard.

We will then replay the X stimulus. To obtain further information about what is perceived by the native speakers of German, we will ask them to write down what the word sounded like to them using German orthography. If they feel like they need to make additional remarks to explain their perception, they are allowed to do so. However, we do not expect this to be necessary in most cases.

3. Results

3.1. Discrimination task

The predicted results for the discrimination task for the two voiced contrasts are shown in Figure 1. The probabilities for correct choices are predicted across participants. For the contrast voiced labial-velar vs. velar plosive, we expect both /k͡p/ and /k/ to be chosen correctly with similarly high probability, but for the accuracy on /k/ being slightly higher. For the contrast voiced labial-velar vs. labial plosive, we expect the probability of /k͡p/ being chosen correctly to be much lower than of /p/. We expect the probability of /p/ being chosen correctly to be very high as well as for /p/ being chosen incorrectly (when X is actually equivalent to /k͡p/ not /p/) to be relatively high.

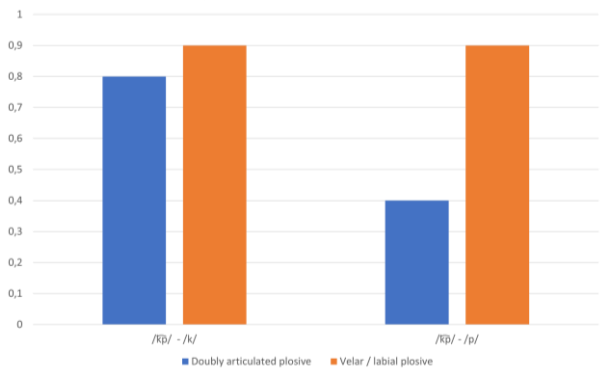


Figure 1: *Probability of correct choice for voiced contrasts.*

The predicted results for the discrimination task for the two voiceless contrasts are shown in Figure 2. The probabilities for correct choices are predicted across participants. We expect similar results for the voiceless contrasts as for the voiced contrasts. For the contrast voiceless labial-velar vs. velar plosive, we expect both /g͡b/ and /g/to be chosen correctly with similarly high probability, but for the accuracy on /g/ being slightly higher. For the contrast voiceless labial-velar vs. labial plosive, we expect the probability of /g͡b/ being chosen correctly to be much lower than of /b/. We expect the probability of /b/ being chosen correctly to be very high as well as for /b/ being chosen incorrectly (when X is actually equivalent to /g͡b/ not /b/) to be relatively high.

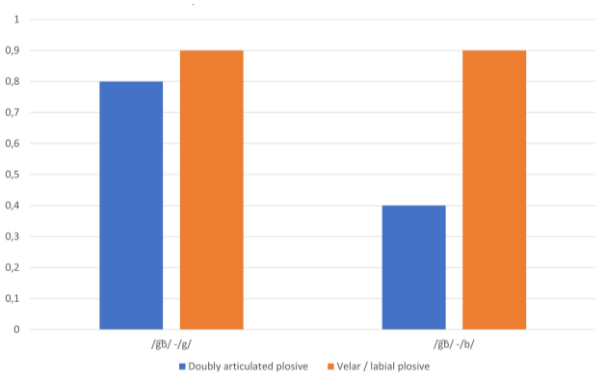


Figure 2: *Probability of correct choice for voiceless contrasts.*

3.2. German orthography task

At this point, it is difficult to predict how the participants might spell out the words they perceive. As Best et al. (2001: 783-784) describe, it might be possible for the non-native contrast to be spelled identically or phonologically equivalent in German or for one item of the pair to be "further modified by punctuation marks or by additional letters". We might also see the pseudowords "spelled with different letters or combinations of letters" (Best et al. 2001: 784, cf. also for further descriptions). All of these options seem to be plausible outcomes of our experiment. For the contrasts voiced labial-velar vs. labial plosive, /k͡p/ - /p/, and voiceless labial-velar vs. labial plosive, /g͡b/ - /b/, we might rather observe equivalent or slightly modified spelling, with the latter being predominant. For the contrasts voiced labial-velar vs. velar plosive, /k͡p/ - /k/, and voiceless labial-velar vs. velar plosive, /g͡b/ -/g/, we might rather observe slightly modified or even completely different spelling, again with the latter being predominant.

4. Discussion

If we were to obtain the predicted results, our hypothesis would be supported to some extent. Especially the results for the two labial-velar vs. labial plosive contrasts (/k͡p/ - /p/ and /g͡b/ - /b/) would show clear tendencies for /p/ and /b/ being chosen correctly almost 100% of the time and for the two consonants being chosen incorrectly, i.e., /k͡p/ and /g͡b/ being mistaken for /p/ and /b/, in many cases. These tendencies would let us assume that German native speakers in fact are not always able to clearly discriminate between the labial-velar consonants and their labial plosive counterparts. However, we do not expect to see the same, clear pattern when it comes to the results for the two labial-velar vs. velar plosive contrasts (/k͡p/ - /k/ and /g͡b/ - /g/). Rather, the doubly articulated and the velar plosives are expected to be discriminated well. There might be a slight difference between the two, with /k/ and /g/ being chosen correctly with higher probability than /k͡p/ and /g͡b/. This is assumed to be due to the fact, that the velar plosives are perceived clearly, while some participants might vary between perceiving the labial or the velar element of the non-native consonant as being predominant. Nonetheless, as labial elements (/p/ and /b/ in /k͡p/ and /g͡b/) are presumed to be perceptually more salient for Europeans, we expect the discrimination to be quite good.

Considering both the discrimination and the German orthography task of our experiment together, we would come to following conclusions. As already mentioned, the results for the labial-velar vs. labial contrasts let us assume that German native speakers assimilate both plosives to native phonemes.

Depending on the outcomes of the German orthography task, one possibility might be that we observe identical or phonologically equivalent spelling, which Best et al. (2001: 783) categorize as SC (Single Category Assimilation). Alternatively, we might observe the non-native element being predominantly transcribed with further modification in form of "punctuation marks or by additional letters to emphasize some phonetic feature", which Best et al. (2001: 783-784) then categorize as CG (Category Goodness Assimilation). As we expect for a majority of the participants to use this latter form of spelling the non-native phoneme, this finding would support our hypothesis: German native speakers perceive the critical items as better (labial plosives) or worse (doubly articulated consonants) exemplars of the native category. However, this conclusion will not hold for the labial-velar vs. velar contrasts. While we do expect some participants to use only slightly modified spelling for the doubly articulated consonants, the majority is expected to produce "different letters or combinations of letters that indicate phonologically different [German] pronunciation" which Best et al. (2001: 784) categorize as TC (Two Category Assimilation). As mentioned already, discrimination in this case is expected to be good.

5. Conclusion

To conclude, our predicted results support our hypothesis, that German native speakers are not able to clearly discriminate between the labial-velar consonants and their labial or velar plosive counterparts, but that they perceive them as better or worse exemplars of the native category, only for the labial plosive counterparts but not for velar counterparts. Nonetheless, the predicted findings for the labial-velar vs. velar plosive contrasts show, as expected, that the listeners rather assimilate the labial-velar plosives to /p/ and /b/ than to /k/ and /g/ respectively.

The presented work will provide a significant contribution to the discussion and understanding of non-native speech perception. It might help answer the question posed by Best (1995: 184) regarding what mature listeners perceive in non-native phones and contrasts. The investigation of the two languages in question, Dza and German, might allow us to make more general inferences about non-native speech perception between other European and African languages that have similar phoneme inventories. In this context, it might also be interesting to dig deeper into why "the labial element is perceptually more salient for Europeans, [while] for Africans it is the velar element which predominates" (Westermann and Ward (1933, as cited in Connell 1994: 442) and any other consequences this might have on the perception of speech. Focusing only on the CG assimilation pattern, and thereby on the /k͡p/ - /p/ and /g͡b/ - /b/ contrasts, it might be possible to obtain further clarification and more nuanced results by adding a rating scale task alongside or as an alternative to the German orthography task. This might allow us to zoom into how close the critical item is to the native ideal. Finally, it also seems valuable to compare our pseudoword analysis to an experiment that uses real words and thereby investigate how non-native phonemes, e.g., labial-velar plosives, are perceived in positions where their counterparts, e.g., labial plosives, would be expected. Such a design would test the magnitude of the Ganong effect on non-native speech perception.

6. Acknowledgements

I would like to thank my colleague, who is a native speaker of Dza, for the common work effort put into the design of this experiment.

7. References

Best, C.T. (1995). A direct realist view of cross-language speech perception. *Speech Perception and Linguistic Experience*, 171-206.

Best, C. T., McRoberts, G. W., & Goodell, E. (2001). Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. *The Journal of the Acoustical Society of America*, 109(2), 775–794. https://doi.org/10.1121/1.1332378.

Connell, B. (1994). The structure of labial-velar stops. *Journal of Phonetics*, 22(4), 441-476. https://doi.org/10.1016/S0095-4470(19)30295-5.

Dahmen, S., Weth, C. (2017). *Phonologie und Schrift*. Paderborn: Ferdinand Schöningh.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, 6(1), 110–125. https://doi.org/10.1037/0096-1523.6.1.110.

Norton, R., Othaniel, N. (2020). The Jen language cluster: A comparative analysis of wordlists. *Language in Africa*, 1(3), 17–99. doi: 10.37892/2686-8946-2020-1-3-17-99.

Westermann, D., Ward, I. C. (1933) *Practical Phonetics for Students of African Languages*. London: Oxford University Press.

The influence of the phonetic specifics of regional varieties of Spanish on L2 acquisition in an immersion setting

Franziska Sprenger[1], Maxime Klingelschmitt[2], Miranda Lalinde[3]

[1]*University of Cologne, Germany*
[2]*University of Paris, France*
[3]*University of Cologne, Germany*

fspreng1@smail.uni-koeln.de

Abstract

This study aims to examine the influence the accent-specific mannerism debuccalization of an L1 speaker on their L2 students. In certain Spanish accents, like Venezuelan, the /s/ in coda position gets debuccalized. That means that the /s/ turns into an [h]-like sound at the end of a syllable. To avoid debuccalization, a Venezuelan teacher might tend to over-articulate the coda /s/ in front of their students. Since it is common for language learning students to adopt speech patterns of their teachers, we assume that students from a Spanish immersion elementary school are going to over-articulate their coda /s/. We expect to find that students of a debuccalizing teacher will pronounce a coda /s/ with a longer temporal duration and higher center of gravity than their peers from a class with a non-debuccalizing teacher.

Keywords: debuccalization, language acquisition, language immersion, over-articulation

1. Introduction

Learning at least one language that is not the official language of the country one resides in, is a mandatory part of many countries' school curriculums. Some schools go even further and offer their students a multilingual education in an immersive setting, where the students have regular classes in their L2 and only use their L1 in certain settings in school. The teachers are usually L1 speakers of the language the students get immersed in. This makes it interesting to look at the students' acquisition of their L2, especially in respect to their articulation. An L1 speaker as a teacher might influence their students with their accent-specific mannerisms of speech.

One mannerism that is present in only some varieties of the language is debuccalization in Spanish. The term debuccalization contains the Latin word *bucca*, which means mouth, thus expressing that something is taken away from the mouth. Phonetically speaking, debuccalization means that an oral consonant is turned into a laryngeal consonant. (O'Brien, 2012, 3) In the case of Spanish, debuccalization occurs in the form of a word-final [s] being turned into an [h]. *Las palabras* is pronounced [lah paˈlaβrah] instead of [las paˈlaβras].

In order to assess if and how students who learn Spanish as their L2 in an immersion setting would incorporate the input of a debuccalizing teacher into their own speech, we must first look at how L2 learners of Spanish incorporate Spanish phonetics into their own speech in general.

For rhotics, Olsen (2012) found evidence that a production of rhotics in English that resembles the one in Spanish caused a higher pronunciation accuracy of L1 English speakers in their L2 Spanish. This leads to the assumption that L2 learners of Spanish should be able to pronounce sounds more accurately in general if their L1 has similar sounds.

In the case of the debuccalization of /s/, predictions on the aforementioned basis are not as easy to make. On the one hand, [h] is not a sound that is used in coda position in English. On the other hand, [h] is a sound that is easy to articulate for English speakers, which should make them capable of also using it in coda position. It would be plausible for the students to either debuccalize a coda /s/ or not to debuccalize a coda /s/. A study by Agostinelli (2013), however, delivers a clearer prediction. She found that when being explicitly made aware of debuccalization, L2 learners of Spanish with English as their L1 tend to debuccalize a coda /s/, whereas learners who are not made aware of the debuccalization realize the coda /s/ as an [s].

It is important to note that the previously mentioned studies were conducted with adult participants. Nevertheless, we can assume that the results would not be significantly different for elementary school students in an immersion setting due to evidence found by Menke (2015). Her study's results indicate that even though students in an immersion school for Spanish are permanently exposed to their L2 throughout the school day, their pronunciation is affected by L1 transfer.

Summing up what we can pull from previously done studies on L2 learners with English as their L1, including research done specifically on debuccalization and students visiting an immersion school, we can make several hypotheses about how students with L1 English who go to a Spanish immersion school would be influenced by a teacher who debuccalizes their coda /s/.

First of all, we can assume that the students will try to imitate their teacher's debuccalization if made aware of it. If the teacher points out this aspect of their accent or might find another way to make the students aware of it (like over-articulating in formal settings while not doing so when casually talking), the students will probably follow their example and debuccalize. However, it is also possible that the

students will not debuccalize if the teacher over-articulates their coda /s/ when talking in front of the class and teaching them vocabulary. It is even likely that the students will also start over-articulating their coda /s/ when realizing their teacher does so in formal settings.
Secondly, we can assume that the participants will be capable of debuccalizing a coda /s/ since the [h]-sound is also used in English albeit never in coda position.

2. Method

2.1 Design

To test the hypotheses, we are going to conduct an experiment where we are going to look at the articulation of the coda /s/ of young L2 learners of Spanish.
The experiment has a 2x2 design. The first factor is the accent of the teacher. One group has a teacher with an accent in which the word-final coda /s/ is debuccalized but who over-articulates in a formal classroom setting while the other group has a teacher who does not debuccalize their coda /s/. With these two groups, we are going to be able to discern the teacher's manner of speaking on the L2 language acquisition of their students.
The second factor is the position of the /s/ in the word. In addition to words containing a coda /s/ there are also going to be words with an onset /s/. This allows us to analyze the articulation of the coda /s/ within each group by comparing it with the onset /s/.

2.2 Participants

The 20 participants are first-grade students, aged 5 to 7, from a Spanish immersion school in Texas, USA. The students have English as their L1 but the main language of instruction in their school is Spanish. Only some classes are held in English. For the participating students this amounts to about 80% Spanish and 20% English during class.
Ten of the participants have a Venezuelan teacher who debuccalizes their coda /s/ in casual talk but over-articulates it when standing in front of the class. The other ten students have a Mexican teacher who does not debuccalize their coda /s/, thus not displaying any difference between their pronunciation of the sound in casual talk and talking in front of their class.

2.3 Material

The material consists of 4 words with a coda /s/, 4 words with an onset /s/ and 4 words without an /s/. The words are taken from the students current "unit of inquiry". These themed units come with specific vocabulary and take about six to eight weeks of time in class to be completed. During this period of time, the vocabulary is displayed in the classroom – the Spanish word in written form and its meaning as a picture (see Fig. 1). Additionally, the students will be asked to produce numerals since these are also high-frequency words and easy to depict.

Figure 1: *Vocabulary in classroom*

2.4 Procedure

The students are tested separately in a setting that resembles a vocabulary test. The teacher then presents the pictures associated with the vocabulary the children are currently learning and the students have to say the corresponding Spanish word. The teacher shows the pictures one at a time. This experimental setting ensures that the students will only say one word at a time, which is important to avoid the potential influence of following sounds on the coda /s/. It is also a situation that is, on the one hand, not unknown to the students, which should lead to them performing not differently than any other day, and, on the other hand, is controlled enough to provide comparable results from all participants.
During the experiment, an audio recording is taken.

2.5 Expected results

We expect the students from the class with the debuccalizing teacher to pronounce their coda /s/ with a longer temporal duration and a higher center of gravity than the children from the class with the non-debuccalizing teacher. We also expect their coda /s/ to have a longer temporal duration and higher center of gravity than their onset /s/.
For the students from the class with the non-debuccalizing teacher we expect no difference in the temporal duration and the center of gravity of the coda /s/ and the onset /s/.

3. Results

The audio recordings of the students are going to be analyzed in Praat (Boersma & Weenink, 2022). The statistical analyses are done in RStudio (RStudio Team, 2020).
First, we are going to look at the temporal duration of the coda sounds produced by the students of the debuccalizing teacher and the students of the non-debuccalizing teacher. Due to the small number of participants all recorded words are going to be analyzed separately instead of first calculating a mean value for each student. The data for each group (as presented

in Fig. 2) are then going to be compared with a *t*-test. We expect the temporal duration of the coda /s/ of the students with a debuccalizing teacher to be significantly longer ($p < .05$) than the other group's temporal duration.

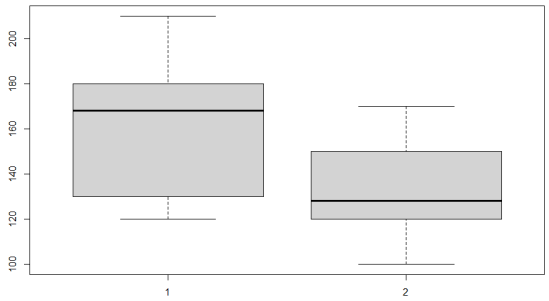


Figure 2: *Temporal duration of coda /s/ in ms for group 1 (debuccalizing teacher) and group 2 (non-debuccalizing teacher)*

The center of gravity for the coda sounds of both groups is analyzed in the same manner (see Fig. 3 for a visual representation). The mean value and standard deviation for each group are calculated and then compared by using a *t*-test. As for the temporal duration, we expect the center of gravity to be significantly higher for the group with the debuccalizing teacher.

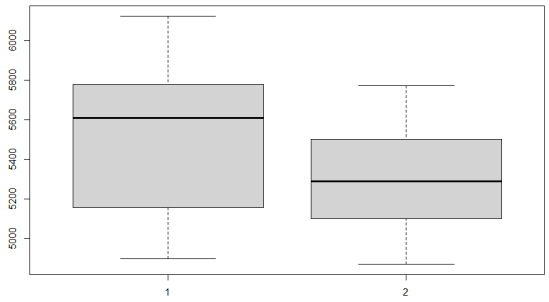


Figure 3: *Center of gravity of coda /s/ in Hz for group 1 (debuccalizing teacher and group 2 (non-debuccalizing teacher)*

Then the temporal duration of the coda /s/ of each group needs to be compared to their onset /s/ (see Fig. 4). Like the comparison between the groups, the comparison within each group will be done by carrying out a *t*-test. We expect the temporal duration of the coda /s/ of the group with the debuccalizing teacher to be significantly different from their onset /s/, but the onset /s/ of the group with the non-debuccalizing teacher not to differ from their coda /s/.
Additionally, we are going to compare the temporal duration of the onset /s/ of each group. We expect them to be not significantly different from each other.

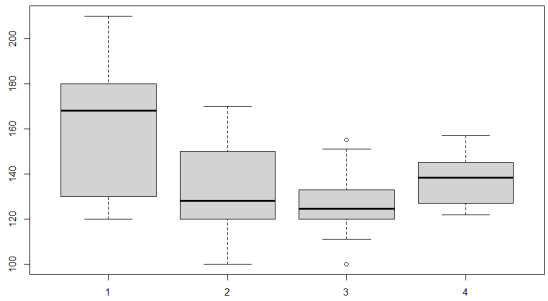


Figure 4: *Temporal duration for coda sound from class with debuccalizing teacher (1) and non-debuccalizing teacher (2) and onset sound from class with debuccalizing teacher (3) and non-debuccalizing teacher (4)*

Similarly, the centers of gravity for the coda /s/ and the onset /s/ of each group will be compared (see Fig. 5). We expect that the center of gravity of the coda /s/ and the onset /s/ of the participants with the debuccalizing teacher will differ, whereas we expect no difference for the students of the non-debuccalizing teacher.
As done with the temporal duration of the onset /s/, we are also going to compare the center of gravity of the debuccalizing teacher's group with the non-debuccalizing teacher's group. We expect no significant difference.

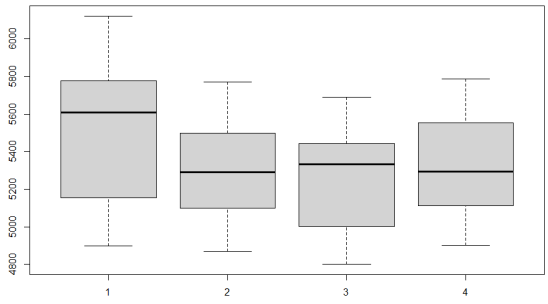


Figure 5: *Center of gravity in Hz for coda sound from class with debuccalizing teacher (1) and non-debuccalizing teacher (2) and onset sound from class with debuccalizing teacher (3) and non-debuccalizing teacher (4)*

4. Discussion and conclusion

The longer temporal duration as well as the higher center of gravity of the realization of the coda /s/ by the students with the debuccalizing teacher support the influence of the teacher's phonetic input on the students' phonetic output. However, we cannot determine whether these findings are due to conscious or subconscious input since we do not know if the teacher made the students explicitly aware of the pronunciation of the word-final coda /s/. If the input was consciously given by the teacher, the findings would be in line with the research results of Agostinelli (2013). A subconscious acquisition of an over-articulated coda /s/, on

the other hand, would require further research on how L2 learners might acquire specific phonetic behaviors.
A possible explanation for the subconscious incorporation of the teacher's input would be the contrast of the teacher's over-articulation when talking in front of the class and their debuccalization in casual conversation. If this was the case, it would give the insight that L2 learners can acquire phonetic specifics of their L2 that are different from their L1 without designated pronunciation training – as opposed to Agostinelli's (2013) findings – but by using the contrast between different ways of articulating a sound.
We do not expect the temporal duration and center of gravity of the onset /s/ to differ significantly between the two groups since there should be no difference in the two teachers' input. However, due to the limited number of participants and the small sample size it is possible that the groups' results might accidentally differ (although probably not to a significant degree) even if this is not representative of the larger population. Therefore, the limitations of the data must be kept in mind when analyzing and interpreting them.
Overall, the previously explained expected results would provide evidence that young language learning students imitate their teacher's way of speaking – at least in a formal setting. By not debuccalizing their coda /s/ but instead over-articulating it, the students of the Venezuelan teacher show that they are susceptible to their teacher's phonetic input and incorporate it into their own speech.
Although we expect the participants to realize the word-final coda /s/ as an (over-articulated) [s] and not an [h], we cannot rule out that the students would debuccalize their coda /s/ in another situational context. Thus, it would be interesting to test whether the students would try to debuccalize their coda /s/ in an informal setting like their teacher does. In this instance, it would come into play that the students' L1 English does not allow /h/ in coda position. Would the students still be able to debuccalize their /s/ although it contradicts the phonological rules of English, thus making it counterintuitive for the children to meet the phonetic requirements of debuccalization? This would be a topic for further investigation.

In conclusion, our expected results show that L2 learners in an immersion school setting are affected by their teacher's phonetic input. They are aware in which situations their teacher debuccalizes their coda /s/ and in which situations their teacher over-articulates the [s] sound, and then follow their teacher's example by imitating the teacher's pronunciation. Due to the experiment's setting, however, we cannot make any representative assumptions about the student's capability of debuccalizing a word-final coda /s/.

5. Acknowledgements

I must thank our lecturers for teaching us enough about phonetics that I was able to write this paper without prior knowledge whatsoever. Furthermore, I am thankful to my caffeine addiction for not becoming concerningly severe during the writing process.

6. References

Agostinelli, C. 2013. *The Effect of Pronunciation Instruction on the Perception of a Novel L2 Contrast*. Buffalo.

Boersma, P. & Weenink, D. 2022. *Praat: doing phonetics by computer*, Version 6.2.10. (URL: http://www.praat.org/, last time accessed: 29 March 2022)

Menke, M.R. 2015. How Native Do They Sound? An Acoustic Analysis of the Spanish Vowels of Elementary Spanish Immersion Students. *Hispania*, 98:4. 804–824.

O'Brien, J. 2012. *An experimental approach to debuccalization and supplementary gestures*. Santa Cruz.

Olsen, M.K. 2012. The L2 Acquisition of Spanish Rhotics by L1 English Speakers: The Effect of L1 Articulatory Routines and Phonetic Context for Allophonic Variation. *Hispania*, 95:1. 65–82.

RStudio Team. 2020. *RStudio: Integrated Development for R. RStudio*. Boston. (URL http://www.rstudio.com/, last time accessed: 29 March 2022).

Enhancing salience: Does it change speech perception?

A follow-up experiment on: The perception of self-produced speech: Is it accurate, and is it used?

Clarissa Selegrad[1], *Clara Lejeune*[2]

[1]University of Cologne, Germany
[2]University of Paris, France
cselegra@smail.uni-koeln.de, clara.lejeune@etu.u-paris.fr

Abstract

In this paper we propose a follow-up experiment on our first experiment in "The perception of self-produced speech: Is it accurate, and is it used?". Both experiments were developed in consideration of Motor Theory and the use of coarticulation cues in speech perception [1]. Our aim is to test if an advantage in hearing self-produced speech compared to other-produced stimuli can be found for the performance in a shadow production task. In this experiment we will add the condition of telling participants about the use of self-voice stimuli in the task. We expect to see the advantage of hearing self-produced stimuli, which we hope to find in our first experiment, enhanced by telling participants they will be hearing their own voice.
Index Terms: self-voice recognition, speech perception, shadow production, coarticulation, Motor Theory

1. Introduction

Through platforms as TikTok, YouTube or Twitch, we have seen a strong increase in video content production during the last years [2], [3]. Most likely, producers of said video content have gotten used to hearing their own recorded voice through the process of recording and editing their videos. Due to the additional bone conduction, self-voice is perceived differently from voices of other speakers, which are only perceived by air [4]. Therefore, speakers show a weak recognition of recorded self-voice [5] or a negative affective reaction when consciously listening to recorded self-voice. This could be the result of the psychological effect, that an unexpected event or sensation evokes a negative reaction [6]. It needs to be considered however, that the studies reporting these effects were conducted when only very few people had access to technology that allowed recording and playing self-speech. The invention and high distribution of technologies like smartphones mean that today a considerable portion of the population has the possibility to record and listen to their own voice. Thus, conscious recognition of recorded self-voice has most likely improved in recent years, especially among the growing community of social media content creators.

According to Motor Theory, listeners use their own production system to analyze speech for coarticulatory cues [1]. We argue that, as the production patterns of self-speech should be most familiar to participants, hearing self-produced stimuli gives an advantage in shadow productions. In a first experiment, detailed in a previous paper, we plan to test if hearing self-produced stimuli gives an advantage in a shadow production task over hearing other-produced stimuli. Additionally, we will examine if the conscious recognition of self-voice improves performance. In our follow-up experiment we will test if telling participants they will be hearing their own voice improves performance on shadow production for self-produced stimuli by encouraging an enhanced access to the motor system and self-voice processing.

2. Method

2.1. Participants

For our follow-up experiment, we will also test 60 to 80 speakers, 70 speakers would be ideal. However, the speakers will all need to have the same L1, either French or German, speak standard variety and not have been diagnosed with either a hearing impairment or a speech disorder. Gender should be evenly distributed among participants, so ideally there would be 35 male and 35 female speakers participating. All speakers will be between 18 and 30 years old.

There are three reasons for this limitation in age range. As in the first experiment, limiting the age to 30 (respectively 40) years should ensure that participants have not yet suffered any hearing loss [7] and at the same time, it is more likely that we will be able to find 70 speakers between 18 and 30 years who have experience with creating video content. Additionally, voices of the speakers should sound as similar as possible, which is one more reason for limiting the age to 30 years [8]. It is especially important that participants are used to hearing their own recorded voice, so activities as regularly creating video content of themselves for social media will be a requirement for participating in the study.

Before starting the experiment, possible participants will fill out a questionnaire. The questions will concern the participants' age, gender, L1 (including variety) experience with creating video content (i.e., hearing their own recorded voice), hearing or speech impairments and participation in our first experiment. Participants who meet the above detailed criteria will be accepted for the experiment. Subjects who have already participated in the first experiment will not be accepted to avoid a priming effect. Additionally, participants should not be close acquaintances (e.g., family members or close friends) with other participants, so that voice and speech patterns of other participants are unfamiliar to them [9], [10].

2.2. Stimuli

Stimuli production for this follow-up experiment will be almost identical to the first experiment. Therefore, it is also inspired by Fowler et al. 2003 [11]. As in the first experiment,

an aCa sequence was chosen to serve as stimulus. The same three consonants /p/, /t/ and /k/ will be used. Participants will be asked to produce these three sequences (/aka/, /apa/ and /aka/) with the first /a/ varying in length. Unlike in the first experiment, only three different lengths will be recorded (short (2s), middle (3.5s), long (5s) [11]) as this should suffice for the second experiment. These nine variations of aCa will be recorded by each speaker three times. This amounts to 27 stimuli by each speaker compared to 72 stimuli in the first experiment and in [11]. As in Experiment 1 and [11], participants will be asked to read from a screen in front of them. First, they will see vowel /a/ for a fixed period of time. Then, /pa/, /ta/ or /ka/ will appear. This way, speakers will not know which consonant will follow when they are producing the first vowel and they will give coarticulatory cues for the consonant only right before producing it. Participants will be recorded in a quiet room. Recordings will be digitized and balanced for volume, but not be edited further (splicing and adding together different recordings) as was done in [11].

2.3. Procedure

First, two groups will be formed, according to the participants' gender as reported in the pre-study questionnaire. Then, each of the two groups will be divided into five sub-groups with the same number of participants in every group. If we find 70 participants and achieve an equal female/male distribution, we will have five groups for the respective gender with seven participants in each group. Participants will be informed that a low degree of deception will be applied in the experiment [12].

The basic experiment design will be the same for all groups, but one differing testing condition will be added for each group. As in our first experiment, inspired by [11], participants will be given a choice task. They will be played the recorded aCa- stimuli and will be asked to shadow each stimulus they hear. There will be a one second pause between the stimuli. In the instruction for the participants, an emphasis will be put on shadowing the stimulus as closely as possible, but to also produce the "correct" stimulus (i.e., producing the same consonant that is in the stimulus). Participants will be asked to perform enough trial runs to get used to the task. In the trial run, participants will only hear other-produced stimuli. When a participant shadows the correct consonant three times in a row, they will start the experiment. Each participant will be asked to shadow 108 stimuli in total.

The first group (Group 1) will be told they will hear stimuli produced by themselves among stimuli produced by other participants. In the experiment 50% of the shadowed stimuli will be self-produced stimuli. This means that participants will hear each self-produced stimulus twice. From the pool of other-produced stimuli, every chosen stimulus will also be played twice.

The second group (Group 2) will have the same experiment design in terms of hearing other- and self-produced stimuli, but they will not be told their own voice will be among the stimuli.

The third group (Group 3) will be told that they will be hearing self-produced stimuli among others, as the first group. In the experiment they will however only shadow other-produced stimuli. As well as in Group 1, participants will hear every chosen stimulus twice.

The fourth group (Group 4) will shadow "blocks" of self-produced and other-produced stimuli. Before each block, they will be told if the following stimuli will be self- or other-produced. As in Group 1 and 2, half of the stimuli will be recorded self-voice, while other-produced stimuli will be chosen randomly. Every stimulus will be played twice as well. The fifth group (Group 5) will receive the same experiment design as the fourth group, but they will not be told if they are hearing self- or other-produced stimuli.

The division into groups according to gender can be explained by the fact that we want to test if self-voice recognition influences participants' shadow production performance. Male and female voices can be easily discriminated [13, 14]. Therefore, if e.g., a female participant hears a male voice, she will know she is not shadowing a self-produced stimulus, whereas with a stimulus produced by another female participant, it might be more difficult to decide if it is a self- or other-produced stimulus.

2.4. Data Collection

To determine the participants' reaction time, we will measure the difference between the consonant onset in the stimulus and the consonant onset in the participants' productions. Consonant onset will be determined by a fall in waveform amplitude [15].

Participants' might wait for consonant release in the stimulus to determine which consonant to produce. However, they would have to produce the consonant considerably later than consonant release in the stimulus. Fowler et al. determined that a consonant onset in the participants' production up to about 140ms after consonant release in the stimulus would not suffice to decide for the correct consonant based on the release in the stimulus [15]. Therefore, we will also consider a consonant onset in the shadow production up to 140ms after consonant release in the stimulus still informed by coarticulation cues in the stimulus consonant onset. Participant productions with an incorrectly shadowed consonant will be eliminated. However, because of the test runs, a low error rate is expected.

2.5. Potential Results

In this second study, we aim to test if explicitly telling participants their own recorded productions will be in the stimuli changes the outcome of the shadow production. Groups will be formed to test different experiment conditions. The purpose of Groups 1, 2 and 3 is to test if telling participants their own productions will be among the stimuli changes their performance. In the experiment design we tried to keep possible factors (i.e., gender, age, ability for recorded self-voice recognition through experience with video content production) as identical as possible.

For Groups 1 and 2, we added one changing condition: telling/ not telling participants, their own voice will be among the stimuli. We see several possible outcomes in the performance of Groups 1 and 2. Participants might perform better (i.e., have an accurate and faster shadow production) on self-voice and other-voice shadowing if being told their own productions will be among the stimuli. Another possibility is that Group 1 or Group 2 performs better only on self-voice stimuli. There might also be no difference in the performance between Group 1 and 2, on the self-voice and/or other-voice stimuli, which would mean telling participants about the self-voice stimuli does not make a difference. It is also possible either one or both groups perform worse on the self-voice stimuli due to the "shock effect" already described in our Experiment 1 paper. However, we hope to rule out this outcome by only accepting participants that are used to hearing recorded

self-voice. Group 3 was added to test if telling participants their own voice will be among the stimuli also improves performance on other-voice stimuli. It is possible that telling participants they will be hearing their own recorded voice enhances their focus on the task (shadowing stimuli) which leads to better performance in general, even if there actually are no self-voice stimuli in the task.

As in Groups 1 and 2, Groups 4 and 5 have one alternating condition: telling/not telling participants, they will be hearing their own recorded voice as stimulus. The difference between Groups 1 and 4 is that participants in the latter group will be told when they will be hearing self-produced stimuli and when not. To facilitate this, the stimuli for Groups 4 and 5 will be arranged in blocks of self- and other-productions instead of a random order. Several different results are also imaginable for Groups 4 and 5. There could be a better performance on the self-voice stimuli in Groups 4 and 5 or only in Group 4, compared to the performance on other-voice stimuli in the same group and/or compared to performance on self-voice stimuli in the first three groups. As we deem it possible that participants show a special focus when being told they will hear self-voice stimuli, there might also be a decrease in focus when participants know they will not be hearing their own voice which might lead to an inferior performance on other-voice stimuli. We also consider a learning effect in Groups 4 and 5 as participants will be hearing several self-voice stimuli in a row. Even if there is not a better performance on shadowing self-produced stimuli from the beginning, performance might improve because participants hear so many self-produced stimuli in a row that they have the possibility to learn about their own articulation patterns [16]. As we will divide participants into groups according to their gender, a difference in performance on the shadow production task would be obvious. However, we do not expect a difference in performance between genders.

Figure 1: *Potential average closure onset latencies (ms) between stimulus and shadow in told vs. untold condition (condition when participants have been told self-voice will be among stimuli vs. when they have not been told, Groups 1 and 2), based on [11].*

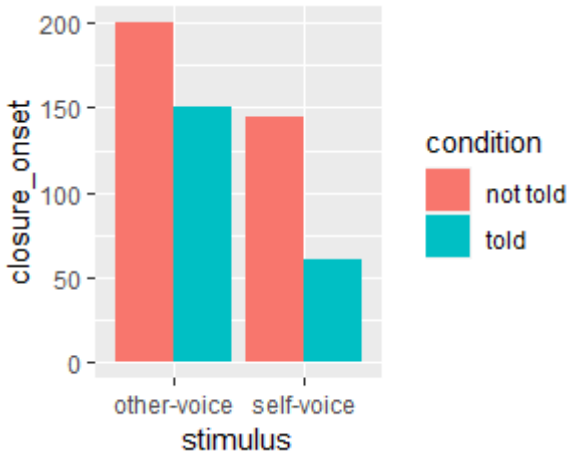
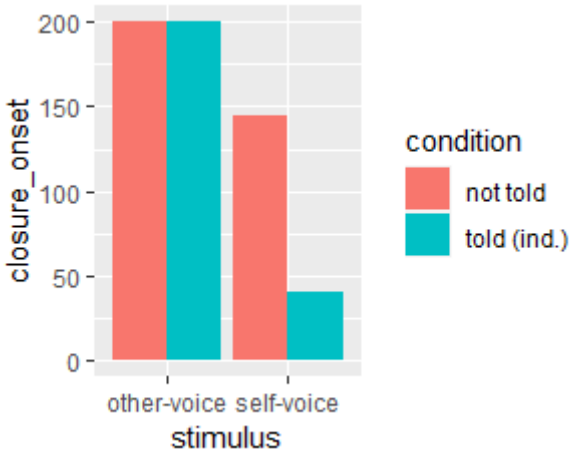


Figure 2: *Potential closure onset latencies (ms) for Groups 4 and 5, based on [11].*



3. Discussion

In the following we will discuss what implications follow from different experiment results and which results we deem most likely according to our hypothesis.

Our hypothesis is that telling participants their recorded aCa-productions are among the stimuli will result in a better performance on the shadow task. We expect an effect on self-voice stimuli as well as on other voice-stimuli, but stronger on self-voice stimuli. There are three considerations leading to our hypothesis.

As for our first experiment, our hypothesis is based on the Motor Theory, developed by Liberman et al. [17]. The existence of a link between perception and production and the assumption, that the production system is activated also when perceiving speech are two of the main claims of Motor Theory [17]. These claims are based on the discovery that listeners perceive acoustic signals matching single phonemes played by a machine at natural speech rate as one undistinguishable noise. This led to the realization, that speakers coarticulate phonemes and that listeners activate their own production system to decipher a natural speech signal they perceive [1], [18]. This means that listeners must be speakers and vice-versa [11]. How listeners compensate for coarticulation depends on their own production patterns. Listeners distinguish and categorize phonemes in perceived speech depending on how they produce these phonemes themselves [19]. They are better and faster at deciphering signals from other speakers when production patterns are similar to their own (e.g., in perceiving nasality) [20], [21], [22]. Listeners also use their own articulation apparatus to deduce missing phonemes in perceived speech [23]. Coarticulation makes it possible to produce comprehensible speech at a fast rate [1]. At the same time, listeners use coarticulation as cue to predict which phoneme will be produced next. Even if a phoneme cued through coarticulation is not the one expected in a context (e.g., of a word), listeners can correctly guess the next phoneme [24]. This means that coarticulation not only enables fast speech production but also supports listeners in perceiving speech. Listeners use their own production system to analyze perceived speech and especially coarticulation cues. The closer articulation patterns of a speaker are to their own, the faster and more accurate is listeners' perception. Thus, listeners must be

particularly capable of perceiving and analyzing and therefore shadowing self-produced speech.

We also considered a possible effect from participants hearing and consciously or unconsciously [25] recognizing their own voice. Natural self-voice is perceived differently from other voices due to the additional bone conduction in the perception of self-voice, which leads to a negative affective reaction or listeners not recognizing their own recorded voice [4], [5], [6]. Thus, self-voice recognition is less good than recognition of one's own face [26]. At the same time, participants rated their own voice more attractive than other voices when not knowing they were hearing their own voice [27]. Familiar and unfamiliar voices as well as self-voice are processed differently [9]. The difference in perception might be one factor contributing to the difference in processing [28]. As can be seen, hearing (recorded) self-voice is different from hearing other voices. We hope to contradict the negative affective reaction reported in [6] or a "shock effect" by only accepting participants familiar with their own recorded voice. Due to eliminating the danger of a possible shock effect, we expect a better performance on the shadow task due to the different processing and the positive reaction on self-voice [27]. We deem it possible that an enhanced access to the processing of one's own voice and to the motor system – and therefore the advantage of decoding one's own coarticulation cues – can be achieved by telling participants they will be hearing their own voice.

Additionally, we consider a psychological effect. We expect that telling participants they will be hearing their own voice will increase their focus on the task. Information about oneself is processed preferentially [29]. As we tell participants in Groups 1, 3 and 4, that they will hear self-produced stimuli, this information may increase their focus in general. This might lead to a better performance, on self- and other-voice stimuli [30]. As participants are more focused, we deem a better performance on other-voice stimuli possible as well. As they expect to hear their own voice, they might pay more attention to every stimulus. Thus, participants in Group 3 should perform better than participants in Group 2 on other-voice stimuli. However, we expect a stronger effect on self-produced stimuli.

To summarize; as in Experiment 1, we expect a better performance on the shadow production task for self-voice stimuli. If results do not show a better performance on self-voice stimuli in Experiment 1, we examine if such an effect can be generated by telling participants they will hear self-voice stimuli and thus encouraging a stronger focus on the task and access to self-voice processing and the articulation system. If results in Experiment 1 indeed show a better performance on self-voice stimuli, we have the possibility to check if performance can be improved by telling participants about the self-voice stimuli in this follow-up experiment.

As not much (especially recent) research has been conducted on the perception of recorded self-voice to base our hypothesis on, we might receive results differing strongly from our expectations. We might find that performance in a shadow-production task is not better – or even worse (e.g., due to a shock effect) – when self-voice stimuli are used. If we also find no improvement in performance when we tell participants they will be hearing their own voice, this means we cannot encourage an access to the motor system by telling participants about the self-voice stimuli.

4. Conclusions

In this follow-up experiment, we plan to test if telling participants they will hear recorded self-speech as stimuli, improves performance in a shadow production task. We decided to only include participants used to hearing their recorded voice in the experiment to reduce the risk of a "shock effect" when participants hear and recognize their own voice. It would however be interesting to see if and how speakers unfamiliar with their recorded self-voice perform in a shadow production task. An option could be to change frequencies of the participants' recorded speech to match (or come close to) how speakers perceive their own natural speech, as was done in [31]. Not only could this reduce the risk of a shock effect, but also enable a more direct access to the participants' speech perception and speech processing. If unedited recorded self-voice is not recognized as self-voice and thus knowledge about articulation patterns is not accessed, editing recorded self-voice to match the speakers' own perception could be beneficial for self-voice recognition and thus change the processing.

Furthermore, we expect a psychological effect resulting from our experiment design. To examine if we can encourage access to the motor system and self-voice processing by informing participants about the use of self-voice stimuli, but make sure we are not simply measuring a psychological effect, we would have to eliminate said effect.

5. References

[1] A. M. Liberman, F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy, "Perception of the speech code," *Psychological review*, vol. 74, no. 6, pp. 431–461, 1967, doi: 10.1037/h0020279.

[2] F. El Afi and S. Ouiddad, "The Rise of Video-Game Live Streaming: Motivations and Forms of Viewer Engagement," in *Communications in Computer and Information Science*, vol. 1421, *HCI International 2021 - Posters*, C. Stephanidis, M. Antona, and S. Ntoa, Eds., Cham: Springer International Publishing, 2021, pp. 156–163.

[3] N. Choudhary, C. Gautam, and V. Arya, "Digital Marketing Challenge and Opportunity with Reference to Tik-Tok - A new rising Social Media Platform," *International Journal of Multidisciplinary Education Research*, vol. 9, no. 10, pp. 189–197, 2020.

[4] D. Maurer and T. Landis, "Role of bone conduction in the self-perception of speech," *Folia phoniatrica*, vol. 42, no. 5, pp. 226–229, 1990, doi: 10.1159/000266070.

[5] C. Rousey and P. S. Holzman, "Recognition of one's own voice," *Journal of Personality and Social Psychology*, vol. 6, no. 4, pp. 464–466, 1967, doi: 10.1037/h0024837.

[6] P. S. Holzman and C. Rousey, "The voice as a percept," *Journal of Personality and Social Psychology*, vol. 4, no. 1, pp. 79–86, 1966, doi: 10.1037/h0023518.

[7] J. C. Alvarado, V. Fuentes-Santamaría, M. C. Gabaldón-Ull, and J. M. Juiz, "Age-Related Hearing Loss Is Accelerated by Repeated Short-Duration Loud Sound Stimulation," *Frontiers in neuroscience*, vol. 13, pp. 1–14, 2019, doi: 10.3389/fnins.2019.00077.

[8] S. Rojas, E. Kefalianos, and A. Vogel, "How Does Our Voice Change as We Age? A Systematic Review and Meta-Analysis of Acoustic and Perceptual Voice Data From Healthy Adults Over 50 Years of Age," *Journal of speech, language, and hearing research : JSLHR*, vol.

63, no. 2, pp. 533–551, 2020, doi: 10.1044/2019_JSLHR-19-00099.
[9] K. Nakamura *et al.*, "Neural substrates for recognition of familiar voices: a PET study," *Neuropsychologia*, vol. 39, no. 10, pp. 1047–1054, 2001, doi: 10.1016/S0028-3932(01)00037-9.
[10] J. Plante-Hébert, V. J. Boucher, and B. Jemel, "The processing of intimately familiar and unfamiliar voices: Specific neural responses of speaker recognition and identification," *PloS one*, vol. 16, no. 4, e0250214, 2021, doi: 10.1371/journal.pone.0250214.
[11] C. A. Fowler, J. M. Brown, L. Sabadini, and J. Weihing, "Rapid access to speech gestures in perception: Evicence from choice and simple response time tasks," *Journal of Memory and Language*, vol. 49, no. 3, pp. 396–413, 2003. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2901126/pdf/nihms214014.pdf
[12] M. C.-T. Tai, "Deception and informed consent in social, behavioral, and educational research (SBER)," *Tzu Chi Medical Journal*, vol. 24, no. 4, pp. 218–222, 2012, doi: 10.1016/j.tcmj.2012.05.003.
[13] M. P. Gelfer and V. A. Mikos, "The relative contributions of speaking fundamental frequency and formant frequencies to gender identification based on isolated vowels," *Journal of voice : official journal of the Voice Foundation*, vol. 19, no. 4, pp. 544–554, 2005, doi: 10.1016/j.jvoice.2004.10.006.
[14] J. M. Hillenbrand and M. J. Clark, "The role of f (0) and formant frequencies in distinguishing the voices of men and women," *Attention, perception & psychophysics*, vol. 71, no. 5, pp. 1150–1166, 2009, doi: 10.3758/APP.71.5.1150.
[15] C. Vicenik, "An acoustic study of Georgian stop consonants," *Journal of the International Phonetic Association*, vol. 40, no. 1, pp. 59–92, 2010, doi: 10.1017/S0025100309990302.
[16] J. S. Magnuson, H. C. Nusbaum, R. Akahane-Yamada, and D. Saltzman, "Talker familiarity and the accommodation of talker variability," *Attention, Perception, & Psychophysics*, vol. 83, no. 4, pp. 1842–1860, 2021, doi: 10.3758/s13414-020-02203-y.
[17] B. Galantucci, C. A. Fowler, and M. T. Turvey, "The motor theory of speech perception reviewed," *Psychonomic bulletin & review*, vol. 13, no. 3, pp. 361–377, 2006, doi: 10.3758/bf03193857.
[18] D. H. Whalen, "The Motor Theory of Speech Perception," in *Oxford Research Encyclopedia of Linguistics*, D. H. Whalen, Ed.: Oxford University Press, 2019.
[19] J. Harrington, F. Kleber, and U. Reubold, "Compensation for coarticulation, /u/-fronting, and sound change in standard southern British: an acoustic and perceptual study," *The Journal of the Acoustical Society of America*, vol. 123, no. 5, pp. 2825–2835, 2008, doi: 10.1121/1.2897042.
[20] P. S. Beddor, A. W. Coetzee, W. Styler, K. B. McGowan, and J. E. Boland, "The time course of individuals' perception of coarticulatory information is linked to their production: Implications for sound change," *Language*, vol. 94, no. 4, pp. 931–968, 2018, doi: 10.1353/lan.2018.0051.
[21] L. Scarbel, D. Beautemps, J.-L. Schwartz, and M. Sato, "The shadow of a doubt? Evidence for perceptuo-motor linkage during auditory and audiovisual close-shadowing," *Front. Psychol.*, vol. 5, p. 568, 2014, doi: 10.3389/fpsyg.2014.00568.
[22] G. Zellou, "Individual differences in the production of nasal coarticulation and perceptual compensation," *Journal of Phonetics*, vol. 61, pp. 13–29, 2017, doi: 10.1016/j.wocn.2016.12.002.
[23] A. D'Ausilio, J. Jarmolowska, P. Busan, I. Bufalari, and L. Craighero, "Tongue corticospinal modulation during attended verbal stimuli: priming and coarticulation effects," *Neuropsychologia*, vol. 49, no. 13, pp. 3670–3676, 2011, doi: 10.1016/j.neuropsychologia.2011.09.022.
[24] C. E. Moore and E. Bergelson, "Listeners can use coarticulation cues to predict an upcoming novel word," *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 43, pp. 2890–2896, 2021. [Online]. Available: https://escholarship.org/uc/item/8h02h6tz#main
[25] M. Candini, E. Zamagni, A. Nuzzo, F. Ruotolo, T. Iachini, and F. Frassinetti, "Who is speaking? Implicit and explicit self and other voice recognition," *Brain and Cognition*, vol. 92, pp. 112–117, 2014, doi: 10.1016/j.bandc.2014.10.001.
[26] S. M. Hughes and S. E. Nicholson, "The processing of auditory and visual recognition of self-stimuli," *Consciousness and cognition*, vol. 19, no. 4, pp. 1124–1134, 2010, doi: 10.1016/j.concog.2010.03.001.
[27] S. M. Hughes and M. A. Harrison, "I like my voice better: self-enhancement bias in perceptions of voice attractiveness," *Perception*, vol. 42, no. 9, pp. 941–949, 2013, doi: 10.1068/p7526.
[28] M. Xu, F. Homae, R. Hashimoto, and H. Hagiwara, "Acoustic cues for the recognition of self-voice and other-voice," *Front. Psychol.*, vol. 4, p. 735, 2013, doi: 10.3389/fpsyg.2013.00735.
[29] P. Tacikowski and H. H. Ehrsson, "Preferential processing of self-relevant stimuli occurs mainly at the perceptual and conscious stages of information processing," *Consciousness and cognition*, vol. 41, pp. 139–149, 2016, doi: 10.1016/j.concog.2016.02.013.
[30] E. T. Higgins, "Knowledge activation: Accessibility, applicability, and salience," in *Social Psychology: Handbook of Basic Principles*, E. T. Higgins, Ed., New York: Guilford Press, 1996, pp. 133–168. [Online]. Available: https://www.researchgate.net/profile/e-higgins-2/publication/232462113_knowledge_activation_accessibility_applicability_and_salience
[31] J. T. Kaplan, L. Aziz-Zadeh, L. Q. Uddin, and M. Iacoboni, "The self across the senses: an fMRI study of self-face and self-voice recognition," *Social cognitive and affective neuroscience*, vol. 3, no. 3, pp. 218–223, 2008, doi: 10.1093/scan/nsn014.

Spirantization of the voiceless velar stop in parkinsonian speech - a purely physiological process?

Janine Schreen [1,2], Julia Marcus[1]

[1]*IfL Phonetics – University of Cologne*
[2]*Dept. of Neurology – University Hospital Cologne*

Abstract

This study aims to examine whether the spirantization of the voiceless velar stop in parkinsonian speech is not only caused by physiological processes, but may also be motivated by compensational aims. Therefore we recorded ten German and ten Spanish speakers who were diagnosed with PD and aged-matched healthy control groups. We found language dependent differences in the rate of spirantization of the velar plosive between the PD groups, indicating that the speakers differ in compensating the underlying dysarthric symptoms due to their respective language. These findings suggest that PD patients do not only underly pathological processes, but actively compensate their limitations in articulation to increase the rate of distinctiveness of their acoustic output.

Keywords: *speech production, spirantization, stop consonants, dysarthria*

1. Introduction

The underlying motor symptoms of PD, tremor, bradkinesia and rigidity, affect the speech of patients in different aspects. Speech gestures are slowed down and the active articulators often do not fully meet their targets. Not only the oral articulators, but also the glottal muscles are affected, which can result in extensive voicing of voiceless consonants [4,8]. The missing negative pressure due to the absent closure causes even more difficulties to produce the characteristic burst of stop consonants, which results in the acoustic impression that segments are "*merged together"* [4]. The tongue body and thus velar consonants and vowels seem to be earlier and stronger affected by the progressing disease.

Former research has pointed out that one charactersistic of parkinsonian speech is the spirantization of consonants, which can be lead back to incomplete closures [3,4]. In accordance with the early affectedness of the tongue body, spirantizations tend to occur more often in velar stops compared to the alveolar place of articulation. Nontheless, incomplete closures also take place in alveolar stops, but a different process tends to occur: if the closure is not completed, the stop is most likely articulated like an approximant ([t] -> [ɹ]). This is reported for English [7] and seems to be the case for German, too, though systematical analyses have not been done yet.

This phenomenon of different acoustic outcome for the stops can be explained by physiological means, as the tongue has more space to move in the front of the mouth than in the back. Additionally the tongue body is moving slower than the tongue tip, thus an incomplete closure rather causes a narrow gap between tongue and velum, causing fritction. Regarding the distinctiveness of the speech sounds in both languages, this process can also be seen as compensational from a phonological point of view. In English the velar fricative does not occur and in German it has an allophonic distribution, so that the replacement of the plosive with its fricative counterpart is not causing major misunderstandings in listeners. On the other hand this would not be the case for the alveolar articulation place, so it is replaced with the approximant [ɹ] to achieve sufficient discriminability [6].

To find more support for this compensational approach, a language, in which the velar fricative is not complementary distributed and in which both the voiced and unvoiced velar stop occur, has to be examined. Spanish is one of the languages to fulfil these criteria, although it has to be taken into account that there are various dialectal differences, so that the actual realization of the phonemes might differ. In standard european Spanish the velar fricative [x] as well as the unvoiced velar stop [k] can occur both in CV and VCV condition. Comparing speech data from Spanish and German speaking patients could thus provide more information about the nature of the spirantization of the velar voiceless stop in parkinsonian speech.

2. Method

2.1 Participants and recordings

A total of 40 speakers were recorded. The German speakers who were diagnosed with PD (N= 10; 7 male, 3 female, aged 63) and a healthy control group (N=10; 7 male, 3 female, aged 68) were recorded at the University Hospital of Cologne. The Spanish speakers who were diagnosed with PD (N=10; 6 male, 4 female, aged 65) and a healthy control group (N=10; 6 male, 4 female, aged 69) were recorded at the UCM Madrid, Department of Psychology.

The study design consisted of two tasks, which included the reading of a part of a chapter from Harry Potter and the Goblet of Fire in the respective language and a semi-structured interview. Both tasks endured 10 minutes, so that for each participant a total of 20 minutes of speech was recorded. Speech data was recorded at 44,1 kHz with a microphone headset and analyzed with praat.

2.2 Data

All voiceless velar stops in VCV condition were selected and measured. To avoid the influence from adjacent fricatives, stops in VC condition were not considered. Stops that were realized as approximants were excluded. In a next step the spirantization phase was measured and the percentage was calculated. All plosives which showed a spirantization under 10% or above 90% were excluded from the dataset and classified as non-spirantisized or fully-spirantisized respectively, so that only semi-spirantisized stops were analyzed. After the selection the number of items for each participant ranged from 250 to 401.

Table 1: Grade of Spirantization for all groups

Group	non-spirantisized	semi-spirantisized	fully-spirantisized
DE-PD	20,1%	66,9%	13%
DE-Con	68,7%	31,3%	-
ES-PD	34,4%	62,7%	2,9%
ES-Con	71,2%	28,8%	-

Table 1 shows the percentage of non-spirantisized, semi-spirantisized and fully-spirantisized velar stops. It is noticeable that for the German PD group fully spirantisized stops appear to occur more often compared to the other groups. However, for both PD-groups semi-spirantisized velar stops are most frequent.

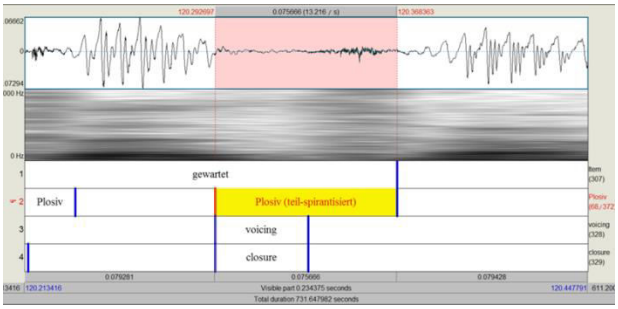


Figure 1: semi-spirantisized voiceless velar stop from the German dataset

Figure 1 shows an example of a semi-spirantisized velar plosive. The spirantization can occur at the beginning of the closure phase, but also, as in the example, during the burst.

3. Results

To examine whether the means of the three conditions *language (German and Spanish), health condition (PD and control) and task (reading and interview)* were significantly different, an ANOVA was done. *Health condition* and *language* have shown to have a strong effect on the data (p = 9.59e-212 and p = 0.000145), whereas *task* had no significant effect (p = 0.468), so it was not considered in the further analysis.

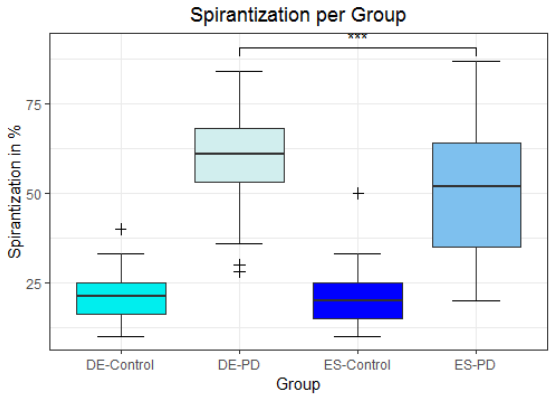


Figure 2: Spirantization in % per group

Figure 2 shows the distribution of the data. The data from both healthy control groups are equally distributed. Both PD groups have significantly higher medians (see Table 2), with the German PD group having overall higher values, and a wider distribution of the data compared to the healthy control groups. The Spanish PD group has the

widest distribution of data, as represented in the boxplot. Table 2 shows that there are significant differences between all groups except for the comparison of both of the healthy control groups.

Table 2: group comparisons

group1	group2	p	p.signif
DE-Control	DE-PD	1,46E-174	***
DE-Control	ES-Control	0,528	n.s.
DE-PD	ES-Control	2,77E-178	***
DE-Control	ES-PD	1,24E-115	***
DE-PD	ES-PD	3,56E-21	***
ES-Control	ES-PD	1,68E-119	***

Figure 3 shows the distribution of the data for the PD subgroups. The means for both groups are similarly high (Table 3), it is shown that the data for the German group is quite centered around 60%, wehereas there are two main centerpoints at 30% and 60% of spirantization rate for the Spanish group.

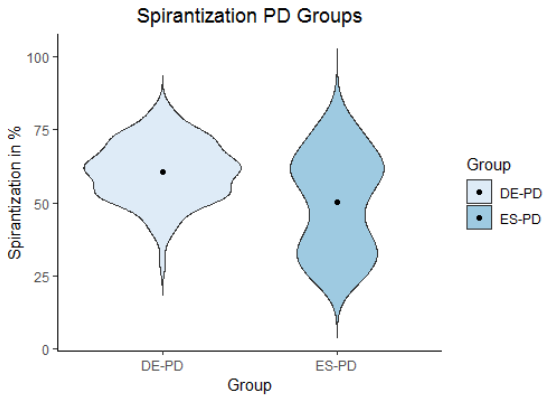


Figure 3: distribution of the data for both PD groups

Table 3: mean and standard deviation for all groups

Group	mean	sd
DE-Con	20,945	5,648
DE-PD	60,74	10,295
ES-Con	20,265	6,173
ES-PD	50,26	17,003

These findings can be interpreted as showing a greater interindividual variability within the German PD Group and on the other hand a greater intraindividual variability within the Spanish PD group. In the German PD group each speaker produced spirantization rates from lower to higher values. Spanish PD speakers instead showed predispositions for either higher or lower values.

4. Discussion

In general, the data indicates that the spiranitzation of the voiceless velar stop is a phenomenon that frequently appears in parkinsonian speech, independant from the language background. The fact that it also occurs – at a lower frequency – in healthy adults could be lead back to an effect of aging, thus physiological processes. Spirantization is caused by incomplete closures or articulatory undershoot [1,3,4], which means that the articulatory target is not fully met or (in the case of semi-spirantisized segments) reached with a delay or might even be interrupted.

An alternative hypothesis could also be that the spiranitzation is caused by some form of artculatory overshoot. Patients could compensate the incomplete closure by using more pressure, which results in aspiration of the stop, because the closure can not fully be realized. German patients could make further use of the aspiration by compensating the lenition of the stop, which is also a frequently reported symptom of hypokonetic dysarthria [4,8,9], as aspiration is a characteristic feature of german stops. Additionaly, by aspirating the stop, the length of the stop increases so that it can rather be perceived as a voiceless stop by listerners. A comparison between PD and healthy controls concerning the main duration of the stops could give further information on that thought, as hyperarticulation can be associated with shorter segment duration. Previous research has already provided different results for speech rate and velocity though, which indicates that patients generally differ in that aspect [4].

On the other hand the Spanish PD speakers in this study have shown more variabilty in the spirantitation of the voiceless velar stop. Some patients produced spirantization of the velar stop on a higher, others on a lower rate. This could either be explained by different severity of dysarthria, or by different compensational strategies of the speakers. As already mentioned in the introduction, there are many dialectal differences in the realization of the spanish phonemes and althogh the speakers lived in the same region, they could differ in their way of realizing the analyzed consonants. The velar region is a quite frequent place of articulation in Spanish,

as both voiceless and voiced stops and fricatives plus (for some dialects) the approximant occur. The process of lenition is a general characteristic phenomenon of the Spanish language and might also be influenced by sociocultural aspects [4]. It is conceivable, that the varition shown in the data is caused by the fact that the speakers already differ at a higher rate in their way of articulation. However, in comparison the Spanish subgroups differ significantly in the rate of spirantization of the velar plosive, with low rates in the healthy control group and higher to lower rates in the PD group, so that the spirantization seems to be linked to the dysarthric symptoms. Further research on the nature of the different realizations of the velar consonants in Spanish in connection with hypokinetic dysarthria and possible compensational strategies is required to further explore this aspect.

Because the stops that have been realized as approximants have been excluded from the analysis, we can not provide any information on a possible accumulation of approximantly realized stops in Spanish parkinsonian speech compared to the German subgroup. A systematic analysis of the realization of alveolar and velar stops in Spanish parkinsonian speech could thus provide more evidence concerning the compensational approach. But taken into account the variation of velar articulation place and the process of lenition, it is imaginable that the distribution of approximant realization is similar to the distribution of the spirantization, as the velar approximant is an allophone of the voiced velar stop in certain Spanish dialects [2,5].

To conclude, the analyzed data has shown a correlation between language background and spirantization of the velar voiceless stop in parkinsonian speech. Further research is required to find more evidence for the claim that speakers activeley compensate their impairment of articulation to achieve higher distinctiveness of their speech sounds. Furthermore the exploartion of the nature of spirantization concerning articulatory under- and overshoot could deliver more insight into dysarthric speech.

5. Acknowledgements

All data has been made up by the author for a student's project. No prticipants were recorded in this study.

References

[1] Ackermann, H., & Ziegler, W. (1991). Articulatory deficits in Parkinsonian dysarthria: An acoustic analysis. *Journal of Neurology, Neurosurgery and Psychiatry*, *54*(12), 1093–1098.

[2] Carrasco, P., Hualde, J. I., & Simonet, M. (2012). Dialectal differences in Spanish voiced obstruent allophony: Costa Rican versus Iberian Spanish. *Phonetica*, *69*(3), 149–179.

[3] Chenausky, K., MacAuslan, J., & Goldhor, R. (2011). Acoustic analysis of PD speech. *Parkinson's Disease*, *2011* (January). https://doi.org/10.4061/2011/435232

[4] Duffy, J. R. (2020). Motor Speech Disorders: Substrates, Differential Diagnosis, and Management. Elsevier Health Sciences. 4th edition.

[5] Hernández-Campoy, Juan & Villena Ponsoda, Juan. (2009). Standardness and nonstandardness in Spain: dialect attrition and revitalization of regional dialects of Spanish. International Journal of The Sociology of Language. 2009. 181-214.

[6] Lindblom, B. (1990). Explaining Phonetic Variation: A Sketch of the H&H Theory. In: Hardcastle, W.J., Marchal, A. (eds) *Speech Production and Speech Modelling*. NATO ASI Series, vol 55. Springer, Dordrecht.

[7] Logemann, J. A., Fisher, H. B. (1981). Vocal tract control in Parkinson's disease: phonetic feature analysis of misarticulations. *The Journal of speech and hearing disorders*, *46*(4), 348-352.

[8] Weismer, G. (1984) Articulatory characteristics of Parkinsonian dysarthria: Segmental and phrase-level timing, spirantization, and glottal-supraglottal coordination. In: McNeil, M., Rosenbek, J., Aronson, A. (Eds), In M. McNeil, J. Rosenbeck, & A. Aronson (Eds.), The dysarthrias: Physiology, acoustics, perception, management. San Diego, CA: College Hill Press, 101–130.

[9] Whitfield, J. A., Reif, A., & Goberman, A. M. (2018). Voicing contrast of stop consonant production in the speech of individuals with Parkinson disease ON and OFF dopaminergic medication. Clinical Linguistics and Phonetics, 32(7), 587–594.

The perception of self-produced speech: Is it accurate, and is it used?

Clara Lejeune[1], *Clarissa Selegrad*[2]

[1]University of Paris, France
[2]University of Cologne, Germany
clara.lejeune@etu.u-paris.fr, cselegra@smail.uni-koeln.de

Abstract

While it has been repeatedly shown that people dislike listening to a recording of their own voice, the few studies on whether people can recognize it without being told (and if it happens consciously or not) seem to contradict each other. Moreover, only little research has been done to explore whether a recognition effect could constitute an advantage in production and/or perceptual tasks. We will thus test people on their ability to quickly retrieve coarticulatory cues through a task where they will have to shadow either a recording of their own voice or someone else's without being told of the speaker's identity. We predict better performance on self-produced speech in general, so we will test French versus German natives. The role of familiarity will also be tested by comparing a group used to self-recording (e.g., through social media) and one which is not. The results could have broad implications in terms of self-awareness and speech perception cues, but also within the more practical fields of language acquisition, language disorders, and speech recognition.

Index Terms: self-produced speech, speech perception, self-recognition, coarticulation cues, Motor Theory

1. Introduction

Many studies suggest bad or a lack of recognition of self-produced speech, such as [1] which was one of firsts to show people's wrong expectations and negative reaction(s) towards a recording of their own voice. Such results can simply be explained by the fact that we hear our own voice through bone conduction (i.e., our skull) as well as through the air when we speak, while a recording of our own voice will only capture air conduction, similarly to as other people hear us (cf. [2], cited by most of the articles here).

Yet, while [3] coincided with this since the subjects did not recognise their voice among others', they still rated it as more attractive than others', and more attractive than the other subjects' judgment of it. This suggests an unconscious recognition, probably due to self-familiarity and memory (cf. 3.1), thus a potentially different processing of our voice, as studies on verbal learning by children like [4] hints at (cf. 3.2).

Moreover, some studies completely challenge the theory of a bad perception of (the recording of) our own voice and/or speech, such as [5] or [6], the latter showing good recognition of one's voice when there is a delay of twelve seconds between the stimulus and the answer, or [7] supporting better self-recognition by participants who regularly listened to their recorded voice (in this case, radio hosts). This latter study suggests that thanks to technology (e.g., audio messages and/or self-recorded videos) our generation might be better at recognising one's voice, perhaps processing it too.

Our experiment aims at testing one's ability to recognise their own voice and use it during a perception and production task. Participants will have to shadow an unpredictable syllable (among three known possibilities) as quickly as possible without being told if the stimulus was a recording of their own voice or not, thus anticipate as much as possible using coarticulatory cues.

A significant difference in speech onset time (SOT, further explained in 2.4.) when the stimulus is one's own voice and when it is not would suggest a conscious or unconscious self-recognition. However, people might show an initial shock as in [1] and [4] causing a lag in SOT which will confirm a recognition but not necessarily show if they can use it if they were simply surprised to hear their voice (they will not be told anything about stimuli types). In such a case, it will be useful to add a second experiment where we tell them when their voice will be the stimuli (cf. discussion in 3.2).

We to predict a positive effect of self-recorded stimuli on speech perception, here through the faster retrieval of coarticulatory cues, whatever one's native language (German versus French natives), partly by extending Motor Theory (MT). Indeed, since MT postulates a similar encoding between production and perception through speech gestures which are retrieved by our motor system [8], it is hypothesised that self-produced speech (gestures) will be better anticipated thanks to self-memory thus familiarity (stronger than acoustic differences because of bone conduction), cf. 3.1.

In a nutshell, we predict participants will process their recorded voice more efficiently, perhaps even recognise it consciously, especially for those with self-recording experience, thus shadow the stimuli they produced themselves more quickly than the ones produced by the other participants i.e., unknown voices. However, as suggested before, it is also possible that participants recognise their own voice without being able to use it to perceive speech in a more efficient way, as it is possible that they do not recognise it at all, since it might not seem relevant from an evolutionary perspective for instance. Implications are further discussed in 3.

2. Method

2.1. Participants

We will test 60 to 80 speakers aged between 18 and 40, separately based on their native language (L1): French or German. Each L1 will have a control and a test group. The control group will be made of speakers without experience with self-recording (through videos, audio messages…) i.e., listening to their recorded voice. The test groups will be made

	French natives	German natives	Total
Age (Average (Lowest/Highest))	… (… / …)	… (… / …)	… (… / …)
Gender (Female/Male [a])	… / …	… / …	… / …
Do you regularly record audios or videos of yourself? (Yes/No)	… / …	… / …	… / …
If yes, do you share them with your close ones (friends, family) and/or publicly online? (Close/Online/None)	… / … / …	… / … / …	… / … / …
Do you regularly communicate via audio messages? (Yes/No)	… / …	… / …	… / …
If yes, do you communicate with your close ones and/or acquaintances and strangers? (Close/Acq&Strg)	… / … / …	… / … / …	… / … / …

Table 1: *Summary of participants' background information*

[a] i.e., the numbers of participants fitting one of the two or three categories, e.g., 33/27 for gender means 33 women and 27 men.

of speakers who regularly record themselves (in an audio and/or visual manner) thus are exposed to their recorded voice more frequently, something predicted to enhance performance.

The selection of participants will be achieved through an online questionnaire with two key questions: Do you regularly record videos and/or audios of yourself? Do you regularly communicate via audio messages? Two other questions on the targets of such videos and messages will be added to hint at a study on communication rather than speech perception, i.e., articulation or intonation differences based on people's way of communicating and with whom. Table 1 summarises the information collected on participants for the selection.

At the end of the experiment, participants will be asked to judge if their hearing is (i) better than average, (ii) average, or (iii) worse than average, and asked if they have ever been diagnosed a language disorder (cf. 3.2). They will also be asked if they recognised their voice among the stimuli.

The age limit was fixed at 40 for two concurring reasons: First, although the beginning of age-related hearing loss is often situated around 60 years old, repeated exposure to loud sounds seems to accelerate this 'natural' onset [9]. Moreover, it might be difficult to find as many older people familiar with self-recording as younger ones, the latter who also may have been using technology and social media from an earlier age.

2.2. Stimuli

Most of the stimuli and experiment is inspired by Fowler et al. (2013) [10] and kept the same in order to compare results. The main difference is that our stimuli will consist of all the recordings made by participants since self-recognition is the main interest here while [10] tested and actually showed quick coarticulation retrieval, i.e., anticipation of unknown syllables almost as well as known syllables.

Participants will always produce a sequence made of: /a/ (chosen in order to be common to French and German) which will vary in eight durations, and the syllable made of a voiceless stop consonant (either /p/, /t/ or /k/) and the vowel /a/ again. The general stimuli template will be referred to as aCa. Each aCa type will be pronounced three times, making it 72 stimuli produced per participant, all as in [10] again.

2.3. Procedure

The recording sessions will happen much as in [10]: Each participant will pronounce the vowel /a/ in eight different durations and produce the syllable which appears on the screen, so that they will not be aware of the upcoming syllable thus not provide too many coarticulatory cues in their first vowel. The syllables will be either /pa/, /ta/ or /ka/, and will be recorded in a soundproof booth one to two weeks (if possible) before the experiment. This waiting period will give us time to collect enough data to start testing participants but will be interesting to keep short since [5] shows a small decrease in recognition performance after only one to two weeks.

The stimuli will be organised in semi-randomised lists, so that each participants will hear one stimulus after the time in a random order while making sure that they hear their own voice thirty to forty times, i.e., circa 25% of the total number of 144 stimuli, but never more than two times in a row (as in [10, for each session of Experiment 1]).

Each participant will know that the consonant will be either /p/, /t/ or /k/ but never which one in advance, and they will not be told either if their own recordings will be used as part of their stimuli list. It will thus be crucial as well to make sure that participants do not know or speak to each other before the experiment, i.e., that only their own voice can sound (at least) familiar.

The main instruction will be: "Shadow the first vowel then the right syllable as quickly as possible", thus we will first have a trial session to make sure that participants adjust their speech onset time to find a good balance between speed and accuracy. None of their own voice recordings will be used as stimuli during this phase. As soon as they are correct three times in a row without taking too much time (based on [10], more than a 200 ms latency will already be abnormal, in which case their results will be considered outliers thus probably discarded), then the test session can begin.

To recap, participants will have to shadow a sequence aCa without knowing which of the three consonants between /p/, /k/ or /t/ will appear, thus try to anticipate the upcoming consonant as quickly as possible using coarticulatory cues, especially since the first /a/ will vary in duration.

2.4. Data analysis

We will measure speech onset time, i.e., the time between consonant onset of the model production and that of the shadow production. The onset of the consonant will be placed at the end of vowel periodicity and/or amplitude (which are

often visible around the same time), i.e., at the left edge of consonant closure since this measure has been proven to be more relevant than (from) consonant release to consonant release in [10]: Indeed, in experiment 3, their participants already perceived the upcoming consonant in the previous vowel's late F2 (i.e., second formants), something also visible in their productions. Moreover, Fowler et al. consider that they started shadowing the Ca syllables too quickly to reflect the wait for consonant release perception.

3. Possible results and discussion

3.1. Hypothetical results and interpretations

Our measure being speech onset time (SOT) and having three variables (two independent ones i.e., self-recording experience and native language, and one dependent i.e., stimulus speaker), there can be many different results, thus leading to different theoretical and practical implications as well as follow-up analyses and experiments.

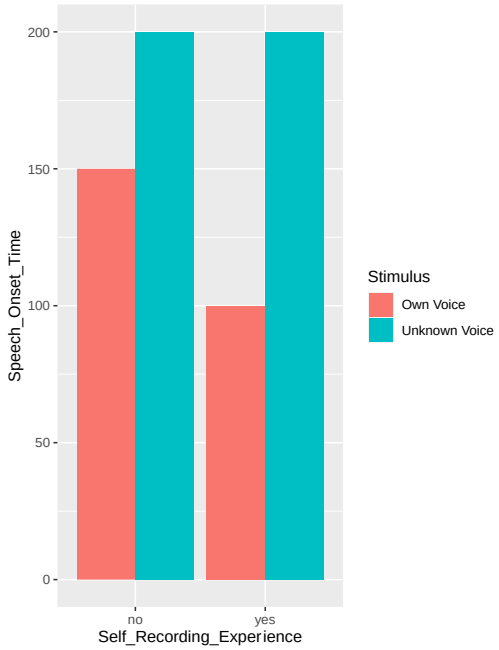
As previously mentioned, we predict no SOT difference based on a participant's native language, but one based on self-recording experience for the stimuli of their own voice only, i.e., a significant help in speech perception (thus shadowing speed) for those familiar with listening to their recorded voice. Yet, we also predict that every participant should recognise their voice (at least unconsciously) thus be quicker on the stimuli in their own voice versus an unknown voice. Finally, we predict no significant SOT difference between native language and self-recording experience groups on stimuli in an unknown voice. Figure 1 shows the expected results (language group is not included because it is not predicted as a factor, and it thus makes the graph clearer).

Yet, while we consider a difference between the two language groups very unlikely (since French and German cultures seem to be as advanced in terms of technology and to use social media as much), there is a stronger possibility that self-produced stimuli is not an advantage for speech perception or that people do not recognise their voice at all.

First, if participants consciously or unconsciously recognise their voice, they may be shocked and/or have a negative reaction towards it as in [1] and [4]. In [1], participants were warned of their upcoming voice and the study showed different expectations from reality. In [4], even though only 40% of the participants consciously recognised their voice, there was an overall negative effect of self-produced stimuli on word learning. Our participants might thus be too diverted by their own voice to concentrate on the task instruction (which is already demanding and far from conversations) thus show a lag. Ideas to compensate these potential results are discussed in the next section.

Finally, it is also possible that participants cannot recognise their voice at all, thus not use it to enhance perceptual performance. Indeed, it is possible that a recording of one's voice is not perceived as a previous personal but 'foreign' speech gesture (because of bone conduction), i.e., that our internal and external speech productions (thus perception) are not mapped with our recorded voice speech gestures. This would suggest that people with self-recording experience have learned to do so, thus it contributes to self-memory and familiarity (through our motor system or not). An intrinsic inability to recognise one's voice could be supported by an evolutionary angle since it might not seem a particular useful survival skill, although it might be becoming one.

Figure 1: *Predicted speech onset time (ms) for shadow production of own versus unknown voice by participants with versus without self-recording experience (based on [10]).*



The reason why it is predicted that participants will recognise their own voice whatever their self-recording experience (although it is predicted to help as well), thus that they may also use it to enhance their perception arises partly from an extension of gestural theories, in particular MT. This latter was first developed by Liberman et al. in 1985 [7] to explain why speech perception was so automatic and robust whereas the acoustic output from speech production seems to lack invariants: For them, our motor system is used during speech perception in order to retrieve speech gestures. Since then, many studies as [7] seem to support such a link between production and perception. Fowler's Direct Realism is very similar to MT in that it also predicts the retrieval of speech perception cues thanks to production structure. Thus, if speech perception and production possess the same currency thus help each other, and especially if they are retrieved by our motor system, it seems logical for us to assume that additional exposure (through internal and external production and perception) would make the processing of our own voice, whether it is recorded or not (the currency being abstract) [7].

Our special familiarity and memory of ourselves and specifically our own actions have indeed already been shown in other domains such as self-face recognition [10] although we do not perceive our face exactly the same as others do. Moreover, self-recognition has also been suggested to be more accurate with the advancement of recording technology and the popularity of audio-visual social media, perhaps making it now useful to be able to discriminate one's recorded voice from someone else's [10, 11].

These intrinsic ability (tested through the group without experience with self-recording on their own versus an unknown voice) and environmental influence (tested through the group with experience with self-recording on their own versus an unknown voice) could be the two main factors for the good self-produced speech perception results we predict.

3.2. Discussion

Although this study could have profound theoretical and even practical implications, it faces some limits.

First of all, the perception of self-produced speech has not been studied much, especially recently, making it difficult to make specific predictions and test even more specific phenomena (e.g., answer delay, block progression).

Indeed, it has already been mentioned that studies seem to contradict each other. One of the most complex aspect of our experiment is the expectation of a quicker SOT when one hears a stimulus in their voice while potentially expecting a lag in SOT in the same condition because of conscious or unconscious self-recognition, especially for participants without self-recording experience [1].

If such a lag is found, a follow-up study could be conducted where we tell participants when their voice is going to be the stimulus, whether it is going to be or not. If the participants have a quicker SOT when their own voice is the stimulus and they are told so, it will confirm a recognition in Experiment 1 and show that such a recognition helps the retrieval of coarticulatory cues, i.e., enhance speech perception. However, if participants show the same SOT with the two types of stimuli when they're told it will be their voice or not (whether it really is or not), it will suggest that the unconscious or conscious recognition found through the lag does not help participants perceive their own speech better than other people's (and perhaps makes it worse as in [4]). It is also possible that participants show a quicker SOT when they were told that their voice would be the next stimulus even when they were lied to. This would imply a psychological effect causing them to pay more attention to their own speech thus perceive it more efficiently.

Such a study could also be conducted if no SOT difference is found between the shadow production of self-produced versus unknown voice stimuli. Indeed, if telling them their voice is going to be the stimulus ends up shortening their SOT, it could imply that recognition of one's recorded voice is difficult but that the access of coarticulatory cues in one's voice is easier when one is self-aware. However, it is also possible to find the psychological effect explained just above.

Moreover, since studies as [13] show more inter-individual than intra-individual variance in speech production, it could be interesting to compare participants' model and shadow productions when their voice is the model and when it isn't to perhaps find individual regularity, i.e., a potential further explanation if participants are good at voice self-recognition and/or anticipation of their own speech. Formants could be computer as in [13] and/or VOT, i.e., voice onset time, since it is a duration measure thus might be more reliable than formants or pitch for an automatic computation. To our knowledge, no such study has been done before, probably because production is deemed as too variant, so we cannot strongly predict the findings of more similarities between one's own model and shadow (of own's model).

Studies reduplicating and completing this one will be needed to understand the key mechanisms of self-perception, self-awareness and self-recognition of one's own voice. Using the same experimental design could also have important practical implications, especially in the fields of speech recognition, language remediation and second language acquisition. I will briefly mention the last two.

First, if our study does show good recognition and/or processing of self-produced stimuli, it will be interesting to also test it with different types of production, for example in one's second language, especially since it has been linked to a less negative reaction to one's own voice, as mentioned in [3]. However, [4] does not suggest that people with phonological impairment (PI) have much more difficulty recognising their voice although it supports a bad representation of their production since recognition was not helped by self-produced deviant stimuli over typical ones. Nevertheless, there was a slightly lower score and more variance in voice self-recognition among the children with a PI, implying the enhancement of it could be crucial as part of their treatment.

While this implies that second language learners might recognise their voice while not being aware of their errors or accent, our study might support the importance of exposure to self-produced stimuli in self-recognition and more importantly in speech perception, in our case the retrieval of coarticulatory cues, i.e., phonetic details. Thus, if the group with more self-recording experience is indeed better and quicker at the shadow task, it suggests that giving more exposure of their own voice to people with language disorders and second (or foreign in the case of immigration) language learners could help improve not only self-recognition, but also attention to acoustic details in their production such as coarticulation, thus better the perception of their errors, which could help a better mapping between their production to the 'standard' one. In fact, the shadow task itself might be a good training.

Finally, while [4] suggest that verbal short-term memory is weaker when the stimuli are made of one's own voice, the authors attribute this effect to the already mentioned negative reaction one has towards their voice because of wrong expectations. Thus, again, we predict that repeated exposure to one's voice might alleviate this negative effect, which will be supported if we do not find a lagging effect in participants with self-recording experience in our experiment. If we do, then our idea of a follow-up experiment where we tell people that their voice will be the stimuli will be crucial to see if this deletes the negative effect of one's recorded voice and perhaps help language learning.

4. Conclusion

We plan on conducting an experiment to check if our hypothesis that one's own voice is better perceived than other people's holds or not, hypothesis partly based on an extension of Motor Theory (and gestural theories in general since they all postulate the retrieval of structure in production to perceive speech). To this end, French and German participants with versus without experience with self-recording will have to shadow an unknown syllable, i.e., anticipate speech using coarticulatory cues, which had been pronounced by them or someone unknown. Thus, our hypothesis will be supported if their speech onset time is quicker when the stimulus is their own voice (without being told is) and when the participant has experience in self-recording.

5. Acknowledgements

We would like to thank the doctors Doris Mücke, Ioana Chitoran and Simon Rössig for their useful courses and feedback, and for kindly providing the template files (especially D. Rössig for the bar chart script).

6. References

[1] Holzman, P. S., & Rousey, C. (1966). The voice as a percept. *Journal of Personality and Social Psychology, 4(1),* 79–86.

[2] Maurer, D., & Landis, T. (1990). Role of bone conduction in the self-perception of speech. *Folia Phoniatrica*, 42, 226-229.

[3] Hughes, S. M., Harrison, MA. (2013). I like My Voice Better: Self-Enhancement Bias in Perceptions of Voice Attractiveness. *Perception, 42(9),* 941-949.

[4] Daryadar, M., & Raghibi, M. (2015). The Effect of Listening to Recordings of One's Voice on Attentional Bias and Auditory Verbal Learning. *International Journal of Psychological Studies, 7(2*), 155-163.

[5] Strömbergsson, S. (2013). Children's recognition of their own recorded voice: influence of age and phonological impairment. *Clinical Linguistics & Phonetics*, *27(1)*, 33–45.

[6] Olivos, G. M.D. (1967). Response Delay, Psychophysiologic Activation, and Recognition of One's Own Voice. *Psychosomatic Medicine, 29(5),* 433-440.

[7] Rousey, C., & Holzman, P. S. (1967). Recognition of one's own voice. *Journal of Personality and Social Psychology, 6(4, Pt.1*), 464–466.

[8] Galantucci, ., Fowler, C. A., Turvey., (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review, 13 (3),* 361-377.

[9] Alvarado, J. C., Fuentes-Santamaría, V., Gabaldón-Ull, M. C., & Juiz, J. M. (2019). Age-Related Hearing Loss Is Accelerated by Repeated Short-Duration Loud Sound Stimulation. *Frontiers in neuroscience*, *13*, 77.

[10] Fowler, C. A., Brown, J. M., Sabadini, L., & Weihing, J. (2003). Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks. *Journal of memory and language*, *49*(3), 396–413.

[11] Chakraborty, A., & Chakrabarti, B. (2018). Looking at My Own Face: Visual Processing Strategies in Self-Other Face Recognition. *Frontiers in psychology*, *9*, 121.

[12] Hughes, S. M., & Nicholson, S. E. (2010). The processing of auditory and visual recognition of self-stimuli. *Consciousness and Cognition, 19(4),* 1124-1134.

[13] Whalen D.H., Chen, W., Tiede, M. K., & Nam, H. (2018). Variability of articulator positions and formants across nine English vowels, *Journal of Phonetics*, Volume 68, 1-14.

Coarticulatory nasalization and the influence of gender

Alicia Janz, Christina Stedtler[1]*, Andrea Fung*[2]

[1]University of Cologne
[2]Université de Paris
ajanz1@smail.uni-koeln.de

Abstract

Speakers use coarticulatory information on vowels to anticipate the following sounds. Based on their own phonological representations they have expectations about the amount and duration of coarticulatory information. Previous research conducted on the topic of coarticulation and coarticulatory nasalization only considered findings from cis-men and cis-women. We conducted a hypothetical study, including trans-women, and -men in our group of participants investigating the question of how nasal coarticulation differs in these speaker groups. In our study we might have found lower nasalization values for trans- men, and -women compared to cis-men, and -women, and higher standard variation. For anticipatory nasal airflow we might have found no significant differences between cis-male, trans-male, and cis-female participant, and significantly longer anticipatory nasal airflow and overall duration of nasalization for trans-female participants.

These results are in line with the assumption hypothesis that speaker's perception of speech, along with social and attentional factors shapes the way words and sounds are mentally represented.

Keywords: gender, coarticulation, acoustic nasalization, anticipatory nasal airflow

1. Introduction

It is widely known that listeners use coarticulatory information on vowels to anticipate the following sounds which was shown e.g. in studies on [u]-fronting [1] and nasalization [2].
Many languages lexically contrast between nasalized and non-nasalized vowels. German is not one of those languages and the fact that vowel nasalization still happens as part of coarticulatory processes is not surprising. What is surprising, however, is that speakers within the same speech-community differ substantially in their use of nasalization. Several studies have found these differences between speakers of different speech communities, supporting the widespread view that production and perception are somehow linked.
In 2002, Pierrhumbert [3] put forward a model in which she proposes that all speech perceptually experienced by a speaker, is classified as instances of word or sound categories, and retained in long-term memory along with phonetic detail. Social, and attentional factors can influence which representations are weighed more heavily and therefore, stored as part of the phonological representations in long-term memory [4]. In this sense it is tempting to imagine that positive attitudes towards a speaker's own speech community shape their phonological representations towards the speech productions used in their community.

Some studies on coarticulatory nasalization found supporting evidence for this model in communities of male vs. female speakers. Tamminga & Zellou [5] for example, showed that women tend to use less nasalization compared to men. In syllables containing a nasal consonant. Khwaileh [2] also found women to use longer anticipatory nasal airflows in high vowel contexts compared to men (target words consisted of vowel /i/ preceded and followed by a nasal consonant).

Both the above-mentioned studies worked with cis-male and cis-female subjects, meaning that all participants identified with the biological sex attributed to them at birth. Therefore, very little is known about participants that do not identify with the sex attributed to them at birth, like trans- or non-binary persons. Investigating coarticulation in these speaker groups could potentially shed light on the question of whether the way we speak is only influenced by biological sex or also, or even more, by social gender.

In our study we, therefore, investigate the question of how participants of different genders, including cis- and transgender male and female, differ in their use of coarticulatory cues, specifically nasalization. We primarily focus on the questions of how gender affects the duration of nasal coarticulation in general and specifically if gender influences the duration of anticipatory nasal airflow.

1.1. General predictions

Previous studies investigating nasal coarticulation in cis-men, and -women found generally less nasal coarticulation for women compared to men. Therefore, we also predict to find generally less nasal coarticulation in cis-women when compared to cis-men and longer anticipatory nasal airflow for cis-women compared to cis-men.

There is a general lack of knowledge about coarticulation in trans-speakers. To avoid bias based on prejudice, we expect to find no differences between the general amount of nasal airflow and anticipatory nasal airflow when comparing trans-men to cis-men and trans-women to cis-women. We do, however, expect to find greater variability in the trans groups because participants differ substantially in their age and stage of transitioning.

2. Method

To elicit whether there are substantial differences in coarticulation patterns of speakers with different gender, we conducted a production study.

2.1. Participants

In our study we tested 80 individuals aged between 21 and 27 (mean age 24.7). All participants were students at the university of Cologne and monolingual German speakers. Additionally, they all lived in the Cologne area for at least ten years with a mean duration of 15.1 years. All subjects received monetary compensation for taking part in this study. Subjects reported no hearing or speaking impairments prior to the study.

Before the experiment all subjects additionally filled out a questionnaire [7] about gender identification and previous language experience. Participants were then divided into four groups: cis-male, trans-male, cis-female and trans-female. A detailed distribution on gender and age of all participants is shown in table 1.

Table 1: *number and mean age (standard deviations in brackets) of all participants by group.*

Gender:	Trans female	Cis female	Trans male	Cis male
Number of participants	18	28	12	22
Mean age	25.6 (3)	23(1,7)	25.8(2,5)	24.2(1,6)

2.2. Stimuli

Our stimuli consisted of ten sets of minimal triplets with a CVC-CVC structure containing either nasal (N) or nonnasal (C) consonants in either matching (C_C-C_C, C_N-C_N, N_N-N_N) or unmatching (C_C-C_N) context. An additional 40 filler words were embedded in each trial. All words were produced twice in a carrier phrase.

Even though previous studies observed gender related differences in anticipatory nasal airflow high vowels, we chose to include target words with non-high vowels only, to avoid any possible formant interference of the vowel formants with the nasal formant, which could have a negative effect on automatic nasality detection.

2.3. Procedure

Participants were asked to produce target words within a carrier phrase with a CVC-CVC sequence containing different vowels. One trial consisted of a training phase (6 sentences), followed by four sets of 25 sentences.

Participants were seated in a sound attenuated booth with headphones (Sennheiser DT 990 Pro) facing a digital screen. Recordings were carried out using head mounted condenser microphones (AKG C 417 PP) and interface (Focusrite Scarlett 2i2) on a MacBook Pro using Adobe Audition 2020 (44kHz, 32Bits, mono). The entire experiment was conducted in a single session, lasting about 1 hour.

All sentences were embedded in a story-like context. Participants listened to the first part of a story and were then asked to read out loud the second part of the story presented on the screen. Each story was accompanied by a short, animated video. Figure 1 shows a screenshot of the training phase.

Figure 1: *screenshot of the training phase without target word and corresponding story picture.*

2.4. Data analysis

Nasality detection poses considerable obstacles when done without using precise, physical methods like measuring the nasal airflow using a face mask. Using these physical methods, on the other hand, might lead to highly unnaturalistic speech.

We therefore chose to measure acoustic nasality using automatic nasal formant detection in *Praat* [5] and Chen's [6] acoustic measure of nasalization A1-P0. This relative measure determines difference in amplitude between low frequency nasal peak (P0) and the amplitude of the vowel's F1 harmonic (A1). With increased nasality, the nasal formant peaks increase and the amplitude of the oral formants (especially F1) decreases.

For reasons of comparability, we normalized all segment durations across speakers. Nasality was always measured at vowel-midpoint.

As acoustic measures of nasality are indirect measurements, they must be looked at relative to other measures like vowel formants.

Segment boundaries were placed automatically and handcorrected by a trained phonetician. The duration of nasalization and anticipatory nasal airflow was then analyzed in proportion to the normalized duration of the segment and syllable. The amount of nasalization was measured at the highest nasal peak (P0), usually at the midpoint of the nasal consonant.

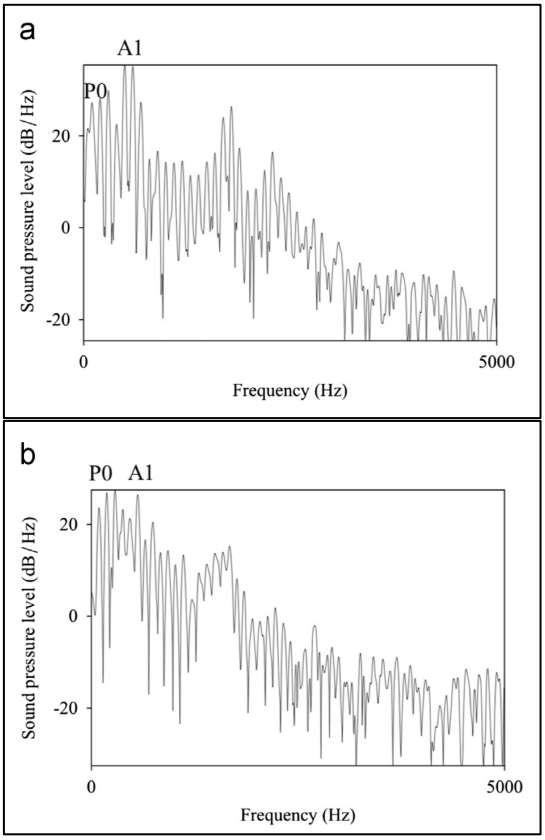


Figure 2: *Example spectra for oral and nasal vowels from the words "bad" (a), and "mad" (b). Two measures of nasality (A1 and P0) are displayed in the upper left corner indicating nasality peaks.* (Adapted from: [8])

3. Possible results

The following section presents hypothetical results. Predictions were made based on the literature presented in the Introduction of this paper as well as the authors imagination.

3.1. Amount of nasalization

The results for the general amount of nasality throughout conditions are displayed in Figure 2.

We found cis-male participants to have the overall strongest nasality values, and smallest variance. Trans-men produced slightly weaker nasalization with no significant difference of the mean values when compared to the cis-male group. We did, however, find a significantly higher mean variation.

For cis-female participants we found slightly less nasalization when compared to cis-male and trans-male participants and no significant differences in standard variation. Trans-female participants showed significantly lower nasalization when compared to the cis-male group and significantly wider spreading when compared to all other groups.

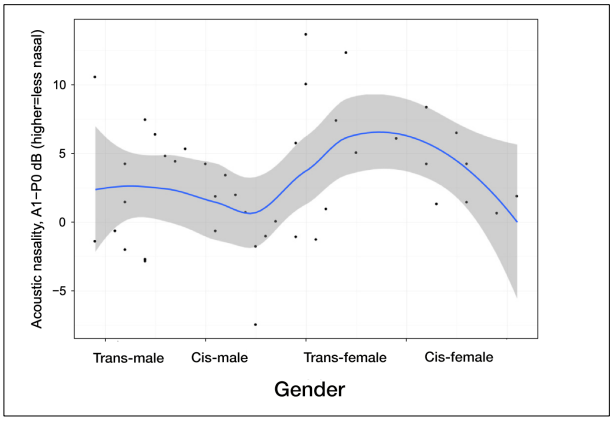


Figure 2: *acoustic nasality (A1-P0) by individual speaker for all groups and conditions, fit with a loess curve. Lower values indicate more nasality, higher values indicate less nasality.* (Adapted from: [8])

3.2. Anticipatory nasal airflow

In vowel context followed by a nasal consonant (CVN), we additionally measured the anticipatory nasal airflow, meaning the duration of nasality preceding the nasal consonant.

In our group of participants, we found cis-male participants to produce the shortest anticipatory nasal airflow with low variance among the group. When compared to trans-men, we found no differences in duration of anticipatory nasal airflow or variance.

For the cis-female group we found significantly longer anticipatory nasal airflow compared to cis- and trans-men, with lower variance compared to all other groups. Finally, trans-female participants used significantly longer anticipatory nasal airflow compared to cis-female, cis-male, and trans-male participants, resulting in the longest overall durations of nasality. Furthermore, the standard variance in this group was significantly higher compared to all other groups, resulting in a much wider spreading of values.

4. Discussion of possible results/ Conclusion

Motor theorists promote the view that coarticulation serves two purposes in speech perception. According to them it is used to efficiently understand the intended utterance (appropriate coarticulation leads to faster, more predictive responses), and it is used to make linguistic decisions about what was said [12]. Vowels preceded by a nasal coda only (CVN), for example, have less nasalization than vowels flanked by both an onset and a coda nasal consonant (NVN) [10]. This coarticulatory information in the speech signal therefore helps speakers to understand an utterance.

Individual representations of the amount of appropriate coarticulation, however, are likely to vary between speakers from different speech communities of the same language. The expectations a listener has about the extent to which particular segments should overlap are based on their language specific phonetic representation [14.]

In our study we found gender related differences in nasal coarticulation between trans-, and cis-female participants,

but not between trans-male, and cis-male participants. Results for the male group, as well as the differences in nasal coarticulation between male and female participants overall, support the assumption that social factors and attitudes shape a speaker's mental representation of words and sounds [3]. Regarding the differences between trans-, and cis-female participants there several possible explanations. It is possible, for example, that speakers of the trans-group either overcompensate by using extreme amounts of nasality and longer durations of anticipatory nasal airflow to suggest social affiliation with female speech-communities. Another explanation could lead in the opposite direction, which would be creating boundaries to all other speaker groups and suggest affiliation with trans-female speakers only. However, it is important to note that the trans-female results were far from homogeneous, making both (and more) explanations possible at the same time.

Furthermore, it is important to mention that biological factors could also play a role in nasalization and subjects from both groups were at differing stages in their transition biologically, but also psychologically.

Previous research has shown that the amount of coarticulatory information has an impact on intelligibility as well as listeners' reaction times [11]. If the perception and production of coarticulation was a learned phenomenon, shaped by the speech an individual is in contact with and has positive attitudes towards, it is likely to assume that speech within an individual's own speech community is the most intelligible. Speech from outside an individual's own speech community, however, would be less intelligible and lead to slower reaction times in discrimination tasks.

If individuals from a certain speech community are perceived to use inappropriate coarticulatory information, this could lead to negative associations towards members of that speech community by speakers from other speech communities.

5. Outlook

Our possible results provided a first insight on coarticulatory nasalization produced by trans-male, and -female participants. We found substantial differences between nasalization durations of trans-female participants when compared to cis-male, cis-female, and trans-male participants. It is yet to be determined which possible social factors and attitudes (possibly even intentionality) could be responsible for these differences.

Furthermore, the question of how these different productions impact the perception of coarticulatory nasalization remains unanswered. We therefore, propose to conduct a perception study including discrimination tasks.

6. References

[1]. Oh, E. (2010). Speaker gender and the degree of coarticulation. *언어*, *35*(3), 743-766.

[2]. Khwaileh, F. A. (2011). *Temporal and aerodynamic aspects of velopharyngeal coarticulation: Effects of age, gender and vowel height*. The University of Tennessee Health Science Center

[3]. Pierrehumbert, J. (2002). Word-specific phonetics. *Laboratory phonology*, *7*(1), 101-140.

[4]. Zellou, G. (2017). Individual differences in the production of nasal coarticulation and perceptual compensation. *Journal of Phonetics*, *61*, 13-29.

[5]. Tamminga, M., & Zellou, G. (2015). Cross-dialectal differences in nasal coarticulation in American English. In *ICPhS*.

[6]. Janz, A. (2022). *Imaginary questionnaire on language experience and gender identification.*

[7]. Boersma, Paul & Weenink, David (2022). Praat: doing phonetics by computer [Computer program]. Version 6.2.09, retrieved 15 February 2022 from http://www.praat.org/

[8]. Zellou G., & Tamminga M. (2014). *Nasal coarticulation changes over time in Philadelphia English.* In *Journal of Phonetics 47 (2014) 18-35*

[9]. Chen, M. Y. (1997). Acoustic correlates of English and French nasalized vowels. *The Journal of the Acoustical Society of America*, *102*(4), 2360-2370

[10]. Cohn, A. C. (1990). *Phonetic and phonological rules of nasalization* Ph.D. Dissertation UCLA

[11]. Chen, M. Y. (1997). Acoustic correlates of English and French nasalized vowels. *The Journal of the Acoustical Society of America*, *102*(4), 2360-2370

[12]. Martin, J. G., & Bunnell, H. T. (1981). Perception of anticipatory coarticulation effects. *The Journal of the Acoustical Society of America*, *69*(2), 559-567.

[13]. Kawasaki, H. (1986). Phonetic explanation for phonological universals: The case of distinctive vowel nasalization. *Experimental phonology*, 81-103.

[14]. Beddor, P. S., & Krakow, R. A. (1999*). Perception of coarticulaoty nasalization by speakers of English and Thai: Evidence for partial compensation.* The Journal of the Acoustical Society of America, 106(5), 2868-2887

Perceptual assimilation of German voicing by Mandarin speakers

Karoline Marliani[1], *Caihong Weng* [2], *Ruoxuan Li* [2]

[1]Universität zu Köln
[2]Université de Paris
kmarlian@smail.uni-koeln.de

Abstract

This paper investigates the assimilation of German plosives to the native perception categories of Mandarin speakers. Mandarin distinguishes between aspirated and unaspirated plosives, both of which voiceless. German meanwhile employs voiced and voiceless plosives.

Newer research suggests that, despite the different voicing, the distinguishing factor of the plosives is aspiration. Grasping this point, this paper aims to observe how 40 Mandarin speakers perceive two different German speaker's plosives compared to their native plosive categories, using the AXB task.

The assumed results suggest that the probands mainly interpret the plosives relying on their native categories. There would be a small difference in perception between the two German speakers, making the one with southern origin easier to perceive.

This would support the idea that German plosives are, in fact, categorized better by aspiration than by voicing – and that this distinction is stronger in southern German speakers.

Index Terms: speech perception, plosives, Mandarin, German

1. Introduction

Non-native speech sounds are often assimilated to native categories. This means e.g., that a native German speaker might use the German representation of an /a/ even when speaking English, where the /a/ is produced slightly differently. Flege (1995: 237) assumes that such mispronunciations, leading to accent, are partly caused by the perception of the speaker, which, with progressing age, tunes itself to the native language. A widespread example includes the /r/ and /l/ sounds which young Japanese children can still perceive, but adults fail to.

Based on the influence the native language shows on perception, Best (1995) invented the Perceptual Assimilation Model (PAM). According to it, non-native speech sounds can be classified as excellent or poor examples of the native categories.

Assimilation of two new speech sounds with non-native contrasts into the native system is excellent when these phones are categorized as different first language (L1) phonemes: Two-Category assimilation (TC). Category Goodness difference (CG) takes place, when the two non-native sounds assimilate to the same native phonemes but show different degrees of fitness. Assimilation is poor when two contrasting phones are categorized as the same phoneme of L1: Single-Category assimilation (SC). This paper will examine perceptual assimilation between Mandarin and German.

Furthermore, an influence of the origin of the German speakers on the perception results will be investigated, contributing to the topic if voicing or spreading of the glottis should be the distinguishing factor between German plosives (as suggested e.g., by Jessen & Ringen, 2002, Solska, 2021).

2. Theoretical Background

2.1. Mandarin Plosives

Phonologically, all Mandarin plosive consonants are voiceless (see Lin, 2007: 22). The conventional feature used to contrast different categories of plosives is aspiration. For example, /pʰ/, /tʰ/, and /kʰ/ are contrasted with unaspirated /p/, /t/, and /k/ in the Mandarin consonant inventory (see for example Lin 2007). However, depending on dialect and position in the syllable, the unaspirated /p/, /t/, /k/ are sometimes represented as voiced /b/, /d/, /g/ in spoken Mandarin (Duanmu, 2007).

On this point, it should be remarked that the number of fluent Standard Chinese (or Mandarin) speakers is assumed to be at about 20 per cent of the Chinese population (Duanmu, 2007: 8). Since the regions speak several strongly different dialects, Mandarin is the connecting language between them, but in many cases not their mother tongue.

2.2. German Plosives

The German plosives, comparatively, are different from the Mandarin ones. The plosives /p/, /t/, and /k/ are voiceless while /b/, /d/, and /g/ are mostly assumed to be voiced in the German consonantal system (Shih & Möbius, 1998).

In contrast, researchers such as Jessen and Ringen (2002) assume that instead of voicing the distinguishing factor should be the spreading of the glottis, which is, in essence, similar to the distinguishing factor in Mandarin, aspiration.

The voicing, meanwhile, is assumed to be a dialectal influence. Especially for the southern regions of Germany, in Austria and Switzerland, all plosives tend to be produced voicelessly (see Solska, 2021). However, as the following examples show, aspiration is also not always produced together with the assumedly voiceless or aspirated plosives /p/, /t/, and /k/. Despite this, the voiceless plosives /p/, /t/, and /k/ in German all share a number of properties.

They are aspirated in word initial position (e.g. Panne [pʰanə]), but unaspirated after /ʃ/ or /s/ (e.g. /p/ and /t/ occur in word-initial clusters starting with /ʃ/: speisen [ʃpa̯ɪzən], steigen [ʃta̯ɪgən]; /k/ occurs in word-initial clusters starting with /s/: Skizze [skɪt͡sə]). There is also free variation between aspirated and unaspirated plosives in final position, but voiceless plosives may only be unreleased when followed by another consonant, never word-finally (e.g. Hut [huːtʰ]). The examples are taken from Lyons (2013).

Since these variants are automatic and predictable, they can all be collected into allophones of the voiceless plosives. /b/, /d/, and /g/ meanwhile, are often produced voiceless as well, such as word-initial, word-final, and after voiceless consonants.

If we recount, these are mostly the same places as the ones where the voiceless plosives are aspirated. This might indicate that a balanced system between voicing and aspiration is used to distinguish between the phonemes, greatly dependent on the positioning in syllable and word.

Since aspiration is seemingly fitting in more cases, however, this paper will focus foremostly on aspiration. The following tokens of the German speakers will focus on the contrast between aspirated and unaspirated as well, placing the contrasting plosives in word-initial position. A study that integrates plosives in word-initial clusters starting with /s/ or /ʃ/ could yield opposing results to the one made here, since voicing could be the distinguishing factor for plosives in this position. Hence, the results concerning the difference between the two German speakers and the following analysis should be viewed with caution.

2.3. Hypotheses

Our question now is with the assimilation of the German plosives /p/ and /b/ inside the Mandarin language system. Since Mandarin (phonologically) does not know voiced plosives, Single-Category assimilation, Category Goodness Difference assimilation, and Two-Category assimilation (Best et al., 2001) are all possible. For simplicity, the Category Goodness difference will be integrated into the Single-Category assimilation for this paper, as it is difficult to observe the goodness of fit with a forced-choice perception test.

In Single-Category assimilation, /p/ in German would be assimilated as [p] in Mandarin, while the non-native phoneme /b/ would be assimilated as [p] as well. Yet, because the German /p/ is produced aspirated in most cases, turning it into [pʰ], it could also be assimilated as the aspirated plosive [pʰ] in Mandarin. In this case, the German /b/ would still be assimilated to [p]. Such assimilation would be supported if /b/ is produced voiceless and the distinction between German /p/ and /b/ is mainly in aspiration. The latter example would be Two-Category assimilation.

Our hypothesis is therefore based on the PAM, in which TC assimilation would be preferred over SC assimilation in the phonetic space. This means that Mandarin L1 speakers would perceive the contrast [b]-[pʰ] (representing TC assimilation) better than the contrast [b]-[p] (representing SC assimilation), which is in line with the Best's PAM: TC assimilation > CG assimilation > SC assimilation. The Category Goodness of fit could later be investigated if the two German phonemes fall into one perceived phoneme for Mandarin native speakers.

Our second, minor objective is to observe if there are strong differences in the perception of the two different German speakers. Since the first hypothesis assumes the stance that an aspiration contrast will be perceived as two different phonemes, this second hypothesis predicts no major differences between the perception of the two speakers. A result like this would speak for the possibility to distinguish German word-initial plosives through aspiration, regardless of speaker origin.

3. Methodology

Two adult native German speakers, one of northern German origin and one of southern, will produce five tokens of each of the four target consonants in CV nonsense syllable pairs. The examples for this will be written in Latin letters, not following the IPA. The target syllables will be <pa>, <ba>, <pu>, and <bu>. Accumulated, this will result in 40 tokens for the following perception task. As preventive measure for errors, the speakers are allowed to reproduce the syllables if they want to. To prevent labelers from accidentally favoring one dialect over the other, only tokens with significant production errors will be excluded afterwards.

For this, we will recruit 40 similarly aged native speakers of Mandarin Chinese who had no experience with German or any other relevant languages (e.g. English and French, which employ similar phonemes). Using The AXB task (for comparison of different perception tasks see e.g. Gerrits & Schouten 2004), the probands will have to perceive the contrasts between the phonemes. For each token, there will a Mandarin Chinese word corresponding to aspirated [pʰ] and a Mandarin Chinese word corresponding to unaspirated [p] played respectively before and after the target syllable, and therefore acting as A and B of the task. It has to be emphasized that this forced choice test assumes that there are no other Mandarin sound categories to which the German phonemes in question are likely to assimilate. The response data of about 1600 choices will be submitted to an ANOVA for the analysis.

4. Results

As figures 1 and 2 show, the assumed results show strong trends to assimilate the German /p/ with the Mandarin voiceless aspirated bilabial plosive [pʰ] and the German /b/ with the Mandarin voiceless unaspirated bilabial plosive [p]. Both results would be verified by an ANOVA analysis. In figure 1, the amount of participants who chose to categorize the German /b/ as Mandarin aspirated [pʰ] is low and can be assumed to occur due to being unused to hearing a different language or erroreous judgement. Additionally, these deviating choices were spread over several different speakers.

As seen in figure 2, Mandarin speakers categorize the German /p/ mainly as [pʰ]. However, the rate at which they

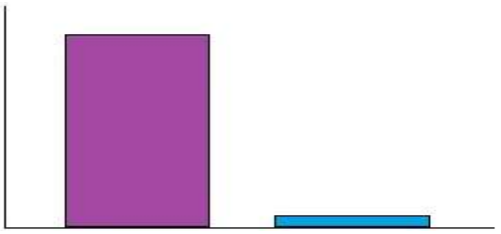


Figure 1 - German /b/ perceived as Mandarin [p] (violet) vs. Mandarin [pʰ] (blue)

categoize the tokens as [p] is higher than the choice for [pʰ] in figure 1. This might be explained through a lack of significant aspiration in some tokens, which could lead to confusion in the Mandarin speakers. Yet, since most German /p/ in word-initial position are produced with aspiration and the Mandarin speakers slowly get accustomed to the contrast between /p/ and /b/ as the experiment continues, the as [p] perceived tokens are in the minority.

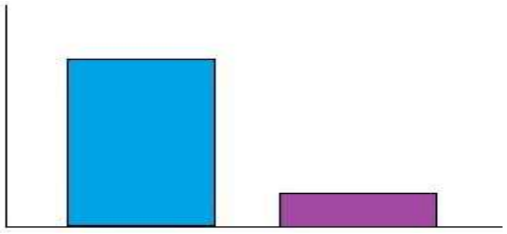


Figure 3 - German /p/ perceived as Mandarin [pʰ] (blue) vs Mandarin [p] (violet)

The results shown in figures 1 and 2 are magnified for the German speaker with a southern dialect compared to the one with northern dialect. As seen in figure 3, the tokens of the speaker with southern origin (the two bars on the right side) are almost always categorization as Mandarin [p] for German /b/ (only 2 tokens as [pʰ] by 2 different speakers), whereas German /p/ was more categorized as [pʰ] compared to the tokens of the northern German speaker (the two bars on the left).

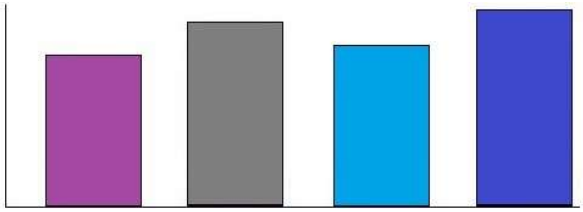


Figure 2 - from left to right: northern German speaker's /p/ perceived as [pʰ] (violet) and /b/ perceived as [p] (grey), southern German speaker's /p/ perceived as [pʰ] (light blue) and /b/ perceived as [p]

The distinction, however, would not be significant in an ANOVA analysis. Especially so, since there are only two speakers as an example, which is not a representative number compared to the total of German speakers.

It is to assume that the voicing of plosives in the northern dialect amplifies the difference form the Mandarin native plosive, thus hindering the categorization.

5. Discussion

If the assimilation results are consistent with our predictions based on the PAM (as shown in the last section), Mandarin speakers will be able to categorize the German /p/ and /b/ as [pʰ] and [p] respectively.

The native Mandarin speakers thus perceive the non-native phones relying on their own categories and the distinction between aspirated and unaspirated as two different phonemes. In the PAM, this is called Two-Category assimilation. From our assumption, the categorization of the phones would proceed like shown above regardless of the dialect of the speaker.

The possibility of Two-Category assimilation is only given if the perceived phones resemble two different categories of Standard Mandarin Chinese. If both German /p/ and /b/ were similar to the same category, the Mandarin speakers would perceive them as the same phoneme. For Best (1995), this would mean a poor assimilation (or Single-Category assimilation). The Mandarin listeners would have trouble learning the correct pronunciation of the language, since their native categories do not differentiate between them.

Luckily for the Chinese population, this differentiation is given with the expected results. Mandarin speakers should be able to perceive the contrast of German plosives in word-initial position rather well and thus learn it more easily than a contrast that is non-native to them.

An example for such a contrast might be the /p/ in word-initial clusters starting with /s/ or /ʃ/. As explained in section 2, /p/ is assumed to be unaspirated in this position. In this case, Mandarin speakers might associate it with the phoneme [p], which was reserved for the German /b/ in our experiment. The Mandarin speakers might thus get confused how to produce the German /p/ properly.

Another issue arises with /b/ behind vocals or voiced consonants. In these positions /b/ might differentiate from /p/ only by voicing, not by aspiration. /b/ in these positions is usually voiced while, to the author's knowledge, there has yet to be researched a clear tendency for the aspiration of /p/ in this position. With the speculation in place that where aspiration occurs in /p/, /b/ is voiceless while where /b/ is voiced, /p/ might be produced without aspiration, it is possible that the native Mandarin speakers have trouble to differentiate the two phonemes in these positions.

After all, the same phone [p] seems to be associated with /p/ or /b/, depending on the position in the sentence. Additionally, in cases where the voicing contrast might be the main contrast between /p/ and /b/, this contrast might be hard to perceive for Mandarin speakers. Depending only on the native perception categories, /b/ and /p/ in such positions could be assimilated by Single-Category assimilation by native Mandarin speakers. The best result for such a contrast might be to end in Category Goodness assimilation, which is still easier to perceive than Single-Category assimilation.

On this subject, further exploration has to be made. Following the imaginary results, /p/ and /b/ seem to be easily distinguished in word-initial position, but the overall perception of /p/ and /b/ might be more difficult depending on position.

Another factor to make perception of contrast between German /p/ and /b/ harder, might be dialect. As shown in the results, the difference for perceiving the dialect of the northern German speaker compared to the southern German speaker is expected to be portrayed by the slightly lesser rate of 'correct' categorization mainly for the northern German /b/. If a Category Goodness difference were to be investigated, the southern German speaker's tokens might be a better fit than the ones from the northern speaker.

This result was expected, because the southern German speakers, as well as speakers with origin in Austria and Switzerland, tend to produce more voiceless plosives. The slope assumed here is that in Switzerland plosives are devoiced the strongest, followed by Austria, and finally Germany – in which the southern part still devoices more strongly than the northern (see Solska 2021: 87). Only considering the plosives, the easiest speaker of German to perceive for Mandarin speakers is most likely one with origin in Switzerland. However, how Solska (2021: 88) emphasizes, the devoicing can not be fully explained through the speakers, but strongly depends on context as well.

Should the results of the experiment not be similar to the one drawn above, then it is to assume that the contrast of the German plosives to the native Mandarin categories is bigger than imagined for the mandarin L1 speakers, even in word-initial position. This might also indicate that the German speakers produced a contrast that is relying on voicing rather than aspiration.

6. Conclusion

The imaginary results for the experiment suggest that Mandarin speakers perceive the German plosive contrasts of /p/ and /b/ rather well. They assimilate it in two categories of their native language. This indicates that the categories of Mandarin and German plosives are similar – at least in word-initial positions.

A study for the perception of German plosives in different phonological contexts could lead to different results. It is strongly suggested for further studies to include contexts where the plosive /b/ is voiced to observe the reaction of the Mandarin speakers.

Single-Category assimilation is a possible outcome for such an experiment. More likely, however, would be a Category Goodness difference, since German /p/ without aspiration in some contexts would be more similar to Mandarin [p] than German /b/. Both results would be problematic for the Mandarin speaker's learning of German, since the similarity to Mandarin [p] is stronger for different German phonemes in different contexts. Hence, it could lead to confusion in the learning process. In this case, the written representations of German might add to the confusion, since <p> and <b> are produced with a big variety of different phonetic features, depending on speaker origin and context.

Additionally, it can be hypothesized that German plosives are indeed categorized more strongly by aspiration than voicing – at least in word-initial position. Further exploration with more German speakers might, however, find a system that is more fitting for the variability in plosives than the distinction by voicing or aspiration alone.

The hypothesis that the rather voiceless productions of southern German speakers are better perceived as northern German speakers yielded no significant results. It is suggested to investigate this again with a bigger variety of German speakers, also including those with origin in Austria or Switzerland.

7. Acknowledgements

The authors would like to thank the organizers of the Paris-Cologne Joint seminar on variability in speech production and perception for their help and the organizing committees of the past INTERSPEECH conferences for providing the template files.

8. References

[1] C. T. Best, *A direct realist view of cross-language speech perception. Speech perception and linguistic experience*, pp. 171–20. 1995.

[2] C. T. Best, G.W. McRoberts, and E. Goodell, "Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system," *Journal of the Acoustical Society of America*, vol. 109, no. 2, pp. 775–794. 2001.

[3] S. Duanmu, *The phonology of standard chinese.* Oxford: OUP Oxford. 2007.

[4] E. Gerrits, M. E. H. Schouten, "Categorical perception depends on the discrimination task," *Perception & Psychophysics* vol. 66, no. 3, pp. 363-376. 2004.

[5] M. Jessen, C. Ringen, "Laryngeal features in German," *Phonology* vol. 19, no. 2. 2002. DOI: 10.1017/S0952675702004311 P

[6] Y.-H Lin, *The sounds of chinese,* with audio cd (Vol. 1). Cambridge: Cambridge University Press. 2007.

[7] E. Lyons, *English and German Consonant Systems Compared. Phonemic and Phonetic Contrasts*, Munich: GRIN Verlag. 2013. https://www.grin.com/document/496677

[8] C. Shih, and B. Möbius. "Contextual Effects on Voicing Profiles of German and Mandarin Consonants," *Proceedings of the Third International Workshop on Speech Synthesis* (Jenolan Caves, Australia), pp. 81-86. 1998.

[9] T. Solska, *Die Realisierung der Plosive in den Nationalvarietäten des Deutschen in Deutschland, Österreich und in der Schweiz.* 2021. https://doi.org/10.14712/18059635.2021.1.4

The role of iconic gestures for speech recognition in infancy

Lena Pfannholzer

Cologne University, Germany
pfannhol@smail.uni-koeln.de

Abstract

The comprehensiveness of speech depends on both the auditory cues as well as visual cues like mouth movements and gestures. In order to understand the process of speech recognition it seems pivotal to not only investigate the roles of each of these visual cues, but also the connection of them. While much research has already been done on the relationship of speech, mouth movements, and gestures in adult communication, little is known about the early development of this connection. Especially in the field of iconic gestures there is still a lot to investigate. This study aims at contributing a part to closing this gap. In two different experiments, evidence is found for the ability of 2-year-old children to not only understand the abstract meaning of different iconic gestures, but also to distinguish between iconic gestures that are congruent and gestures that are incongruent with the accompanying speech. These findings pave the way for future research to investigate the early development of the connection between the visual cues of mouth movements and gestures together and the auditory cues of spoken words.

Index Terms: speech recognition, mouth movements, iconic gestures, multimodality, communicative development

1. Introduction

Human communication is multimodal: We produce and perceive auditory cues via speech and visual cues via gestures (including head movements) and mouth movements. The two modalities of communication have already been investigated on a large scale, though only a few studies have dealt with the connection of both, e.g., Drijvers & Özyürek (2017) or Krason et al. (2021). Drijvers & Özyürek (2017), for instance, found that the best recognition of clear as well as degraded speech was achieved when both visual cues, mouth movements and gestures, were present - iconic gestures being the even bigger factor for comprehensiveness. In another study, Krason et al. (2021) were able to support the findings that "iconic gestures have a pivotal role in face-to-face communication" (10). They also found more evidence for the complementary use of speech and gestures in speech recognition including a dynamic process of weighing auditory and visual cues during communication. These findings show that the investigation of the role of visual cues, especially that of gestures, is indispensable for understanding the process of speech recognition.

However, in the context of communicative development, the various studies to be found have either focused on lip movements (e.g., Dodd, 1979; Kuhl & Meltzoff, 1982) or gestures (e.g., Behne et al., 2014). Since the field of research on the informativeness of gestures is still fairly young by comparison, there is still a lot to investigate. While there has already been a lot of work concerning young children's use of deictic gestures such as pointing, little is known about their use of other kinds of gestures, for instance iconic gestures (Behne et al., 2014: 2049).

The proposed study aims at contributing a part to fill this gap, in order to pave the way for future studies that can investigate the connection of both mouth movements and as gestures in early language development.

1.1. Lip movements

The role of lip movements for speech recognition in infancy has been of long interest for researchers of linguistics as well as psychology, since findings in this area of research may deliver strong evidence for the motor theory of speech perception.

In her study Dodd (1979) was able to support the view that young infants are already aware of the congruence or incongruence of lip movements and speech. For her experiment, she presented twelve 10- to 16-week-olds with nursery rhymes that were shown to the infants with the lip movements and speech being in- and out-of-synchrony (with 400 ms delay). The setup is illustrated in the following figure.

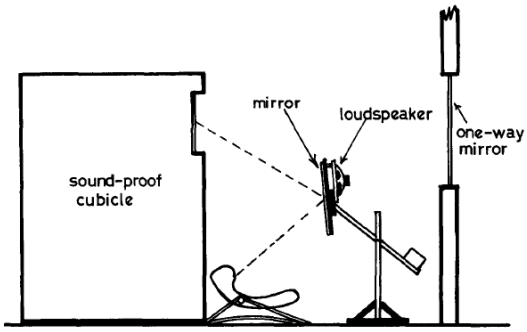


FIG. 1. Arrangement of apparatus.

Fig. 1: Experimental setup in Dodd (1979: 481).

The attentiveness of the subjects was assessed by two independent observers. The results of this study showed that infants consistently preferred to look at the matching stimuli rather than the mismatched stimuli.

In another experiment by Kuhl & Meltzoff (1982) a few years later, 18- to 20-week-old infants were shown two videos each which displayed the same face, one repeating /a/ as in *pop* and the other repeating /i/ as in *peep*. At the same time, they were listening to an audio track with the respective sounds. The subjects' attentiveness to either of the two faces (only one of which displayed the matching sound) was observed with an eye-tracking device. The results of this study also seem to indicate that children so young might already

have knowledge of the relationship between auditory cues and articulation.

1.2. Iconic gestures

It is still an open question at which point iconic gestures become understandable and producible in the communicative development of infants. Behne et al. (2014) state that it is even unclear "whether 2-year-old children are already capable of this" (2050). This unclarity might stem from the abstract meaning of iconic gestures in comparison with deictic gestures, for instance. The majority of studies on gesture-speech integration accept McNeill's rather broad definition of iconic gestures:

> "[…] they are free to show only what is relevant, and also are unable to show anything else. For this reason, iconic gestures, together with the accompanying speech, offer a privileged view of thought. They are the closest look at the ideas of another person that we, the observers, can get."
>
> (McNeill, 1992: 133)

McNeill (1992) states that iconic gestures allow the observer to read someone else's mind. In this sense, no concrete meaning is conveyed, but very abstract additional information to what is being said, e.g., information on actions, shapes or sizes of objects, or spatial relations.

Fig. 2: Iconic gesture showing the form of a sandwich.

There is no consensus on the importance of iconic gestures for speech recognition. However, in most studies they are attributed relevance to the understanding of the message (cf. Kandana Arachchige, 2021).

Considering that the studies in this genre are primarily focused on very young infants or adults, it seems promising to explore the gap between the age groups. It is clear that children learn to understand the connection between lip movements and speech already at a very early age, but can we say the same about iconic gestures?

2. Methods

In order to approach an answer to this question and to ensure comparability with other studies in this field, the following two-step approach is proposed: Firstly, an experiment is designed that allows to test infants' abilities to derive information from iconic gestures. In a second experiment, those iconic gestures that proved to be understandable by the subjects in Experiment 1, are presented in- and out-of-synchrony to the accompanying speech in a similar design as in Dodd (1979) to test infants' abilities to distinguish between matches and mismatches of speech and gestures.

The respective hypotheses to be tested are the following:

H1: 18- to 28-month-old children can consistently derive information from iconic gestures.

H2: 18- to 28-month-old children favor iconic gestures that match the accompanying speech which would indicate that they are able to distinguish between congruent and incongruent gestures.

2.1. Experiment 1

40 18- to 28-month-old native speakers of English will be presented with ten different setups. Each time, either a single physical object or two objects in a specific spatial relation are placed in front of them. The setups are the following: (big) ball, marble; big box, small box; big pyramid, small pyramid; marble on top of/next to/inside/behind a box.

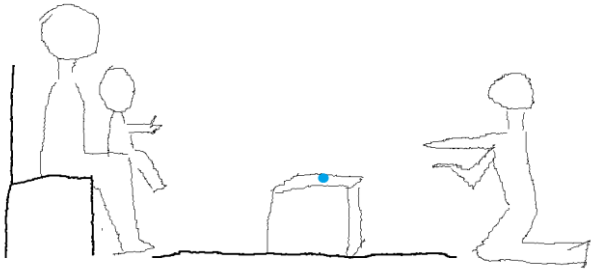


Fig. 3: One of the setups for Experiment 1 (marble on top of a box).

In a face-to-face interaction a narrator will ask the infants to choose the correct iconic gesture to describe the scene while substituting the target word with the corresponding iconic gesture. The questions will be of the following form: Is this ball like this [*painting a round shape*] or like this [*painting a pointed shape*]? Is the marble here *[painting the spatial relation of behind*] or here [*painting the spatial relation of inside*]?

Correct answers will be rewarded with a point, and points will be deducted for incorrect answers. This means, that each subject will be ranked on a scale from -10 to 10, that represents their ability to derive the correct information from the ten different gestures. Furthermore, the iconic gestures will be rated by their comprehensibility and ranked on a scale from -40 to 40 (representing the number of participants).

2.2. Experiment 2

The same participants, excluding those who produced too many errors and ranked lower than 0, will be presented with a similar setup as in Kuhl & Meltzhoff's (1982) experiment. The children will be seated in front of two screens displaying the same narrator showing two different gestures, only one of those matching the story that is being told. To ensure the independence of the results from the mouth movements, the narrator's head will not be visible in the video.

The stories contain the iconic gestures that were proven to be understandable in Experiment 1 and will be told in the way of a fairytale with the gestures being incorporated in the story as naturally as possible. The attentiveness of the subjects to the screens will be observed with an eye-tracking device. Furthermore, the overall attentiveness to the story will be analyzed by two independent observers who will be using

attention buttons (buttons that are pressed whenever the subject loses/gains back obvious interest in the story).

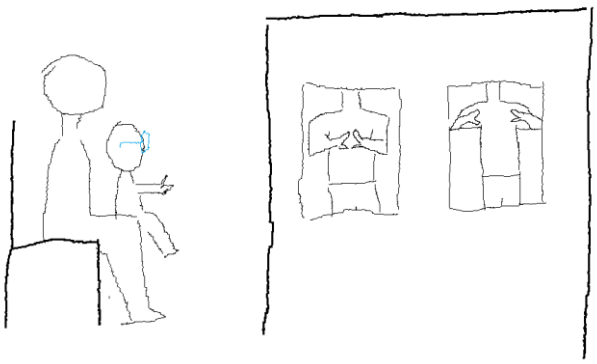


Fig. 4: Setup for Experiment 2.

In order to ensure the significance of the results, each subject will be listening to and watching three different fairytales: One where the gestures on both screens will always be matched with the speech, one that is accompanied by only mismatched gestures, and a last one where the congruence of the visual stimuli with the accompanying speech will be varied. In all three setups, the children's attentiveness to the story as well as the specific focus on one of the screens will be analyzed and compared with each other for each fairytale.

3. Results

The possible results of the study are presented and discussed in the following two chapters.

3.1. Experiment 1

As shown in Figure 5, the majority of the subjects were able to match the correct iconic gesture with the scene that was presented. Only six out of 40 infants produced a negative score, while 22 infants scored five points or higher.

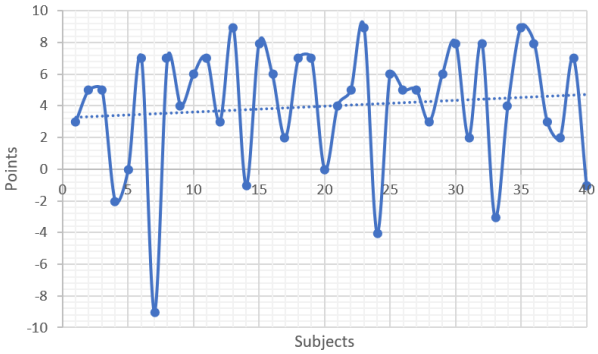


Fig. 5: Representation of the rating for each subject's ability to derive information from iconic gestures.

The overall tendency shows very clearly that the participants in this study have consistent understanding of the ten iconic gestures that were chosen for this experiment. Out of those gestures, however, two (gestures 2 and 8) seemed to cause comprehension problems, as can be seen in Figure 6.

The other eight iconic gestures were matched fairly consistently with the correct setup.

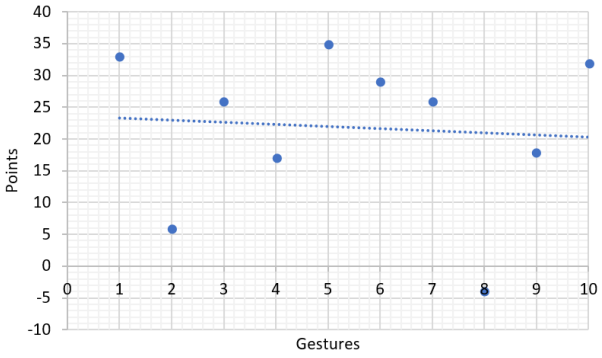


Fig. 6: Representation of the rating for the comprehensiveness of each gesture.

The two gestures that caused the most problems were the one for the marble/small ball on the one hand, and the gesture for the spatial relation of *next to* on the other. One can only speculate about the reasons for this ambiguity. A peculiarity of the *marble gesture* was that, in contrast to all other gestures, it was usually only executed with one hand. For this reason, it can be suggested that the gesture was not as salient as the other one. It also might have been confused with the iconic gesture for 'okay' (see Fig. 7).

Fig. 7: Iconic gesture for 'okay'.

Currently, the data does not allow for a satisfactory explanation for the incomprehensibility of the *next to-gesture*. A second study with a similar design seems necessary to support these results and find reasons for the inconsistencies. Regardless, the six subjects that reached a negative score, as well as the two gestures that scored below 15, have been excluded from the second experiment.

3.2. Experiment 2

Taking into account the results of Experiment 1, the 34 remaining subjects were each presented with three stories (on three different days) containing the eight iconic gestures that were proven to be understandable. In the first round, two different male narrators were displayed on the two screens, both of which always executed the correct gesture to the respective scene. As expected, no significant tendency for the one or the other screen could be observed, as exemplified for Subjects 1 and 2 in Figure 8.

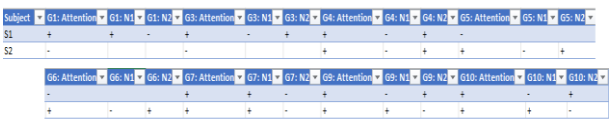


Fig. 8: Representation of attention for Subjects 1 and 2 for each gesture (all of them matches). Next to the overall attention, the specific focus on either Narrator 1 or 2 is listed.

The second story only contained mismatched gestures and was also performed by two different narrators. Like in the first setup, no tendency for the subject to look at either of the two screens could be observed. The results showed very clearly though that the overall attentiveness was significantly lower than in the first round (see Fig. 9). The findings go along with the widely held opinion that iconic gestures provide relevant information that support speech recognition.

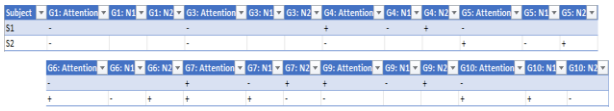


Fig. 9: Representation of attention for Subjects 1 and 2 for each gesture (all of them mismatches). Next to the overall attention, the specific focus on either Narrator 1 or 2 is listed.

In the last and crucial round, the congruence of the iconic gestures with the speech was varied. For each scene that was accompanied by an iconic gesture, one matching and one mismatching gesture was presented by the same narrator on each screen. Like in the first two rounds, the overall attentiveness of the subjects was analyzed together with their attentiveness to either the correct or the incorrect iconic gesture. An extract of the data can be seen in Figure 10.

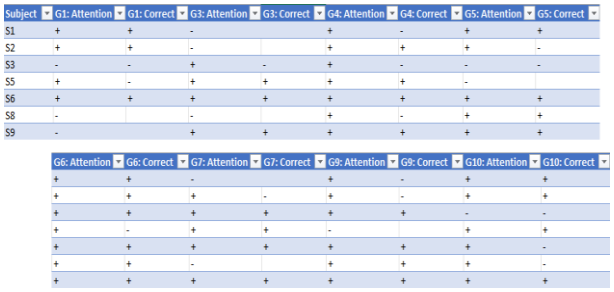


Fig. 10: Representation of attention for Subjects 1, 2, 3, 5, 6, 8, and 9 for each gesture (each with one match and one mismatch). Next to the overall attention, the specific focus on either the correct or incorrect gesture is listed.

The comparison of the different setups of Experiment 2 shows that the overall attentiveness in round one (only matched gestures) and round three (one matched and one mismatched gesture) was more or less consistently on the same level. The specific focus on either of the screens is therefore well comparable. While in round one no tendency for either of the screens could be observed, the analysis of the data from the third round shows that in roughly 74 % of the cases the subjects paid attention to the gesture that was congruent with the accompanying speech. This result can be interpreted as an indicator for an overall preference for 18- to 28-months old children to look at matched iconic gestures rather than mismatched ones. Furthermore, these findings seem to strongly support the view that children at that age are already able to derive the abstract meaning of iconic gestures and have the same understanding of the connection between speech and iconic gestures as between speech and lip movements.

4. Conclusions

The results of Experiment 1 and 2 provide strong evidence for both my hypotheses. To further explore and prove the early development of the understanding of iconic gestures, future research should conduct similar studies with children of different age groups. Above all, it seems necessary to increase the number of gestures investigated, since in the scope of this study only a few iconic gestures could be looked at.

Moreover, it seems promising to investigate the connection of the different visual cues (gestures and mouth movements) together with the auditory cues, to learn more about the process of speech recognition in general.

5. Acknowledgements

I would like to thank Nicholas Griffen (Université Paris Cité) for the enjoyable collaboration in the development of the idea for this study. Special thanks go to Doris Mücke (University of Cologne) and Simon Rössig (University of Cologne) for their guidance and patience in the preparation and writing of this paper. Last, but not least, I want to express my gratitude to RuPaul who taught me, that if I can't love myself, how in the hell am I gonna love somebody else? Ramen.

Hürth, April 2, 2022

6. References

[1] Behne, T., Carpenter, M., & Tomasello, M. "Young Children Create Iconic Gestures to Inform Others", *Developmental Psychology* 50(8), 2049–2060, 2014.

[2] Dodd, B. "Lip Reading in Infants: Attention to Speech Presented in- and out-of-Synchrony", *Cognitive Psychology* 11, 478–484, 1979.

[3] Drijvers, L. & Özyürek, A. "Visual Context Enhanced: The Joint Contribution of Iconic Gestures and Visible Speech to Degraded Speech Comprehension", *Speech, Language, and Hearing Research* 60(1), pp. 212–222, 2017.

[4] Kandana Arachchige, K., Simoes Loureiro, I., Blekic, W., Rossignol, M. & Lefebvre, L. "The role of iconic gestures in speech comprehension: an overview of various methodologies", *Frontiers in Psychology*, 12, 2021.

[5] Krason, A., Fenton, R., Varley, R., & Vigliocco, G. "The role of iconic gestures and mouth movements in face-to-face communication", *Psychonomic Bulletin & Review*, 2021.

[6] Kuhl, P. & Meltzhoff, A. "The Bimodal Perception of Speech in Infancy", *Science* 218, 1138–1141, 1982.

[7] McNeill, D. *Hand and mind: What gestures reveal about thought*, University of Chicago press: Chicago, 1992.