The perception of emotions in speech in two closely related languages

Lucia Cimrova

12924326

Supervisor: Prof. Paul Boersma

BA Thesis Linguistics

University of Amsterdam

27.06.2022

Table of Contents

Abstract

Most studies investigating emotion recognition have reported observable differences between culturally distant language speakers' perceptions of emotions. The authors noticed that native speakers reach higher accuracy in their native language than in a foreign language, describing this phenomenon as the in-group advantage. In this paper, we tried to determine whether speakers of two culturally and linguistically close languages, such as English and German, perceive emotions in the foreign language differently than in their native language. The results of our study turned out to be insignificant, as we expected (H1), but due to the small sample size, we were unable to draw conclusions. However, our methodology seems appropriate, and therefore we believe that it could serve helpful in future larger-scale experiments.

1. Introduction

Emotions are one of the typical characteristics of humans and animals. While these states of consciousness play an essential role in our daily lives, we still struggle with defining them. The American Psychological Association (APA) described emotions as "complex reaction patterns involving experimental, behavioural, and psychological elements". Feelings and moods arise from them (UWA.edu. 2019). Aside from experiencing emotions, people tend to express them in various ways. The following paper investigates the cross-linguistic variation in the perception of emotions in speech.

Previous research has shown that emotions can be recognized based on facial expression and the speaker's voice, regardless of the language they speak (Pell, et al., 2009). The same study also proved that native speakers have an advantage in recognizing emotions in their native language. The authors investigated Argentinian Spanish speakers' speech emotion recognition (SER) performance in English, German, and Arabic. According to the results, Spanish speakers score higher on stimuli in their native language than in the other languages. Language distance has been one of the factors influencing the outcome of the results, as Spanish is a Romance language, unlike German, English, and Arabic. Schuller (2018) shows in his study that emotional speech also varies across cultures. Several researchers have observed this phenomenon earlier, and found that people's perception of emotions depends on their cultural background and native language (Kamaruddin et al., 2012; Elfenbein & Ambady, 2002). The event has been characterized as the 'in-group advantage' (Elfenbein & Ambady, 2002; Althoff et al., 2016; Elfenbein, Laukka, 2021). The meta-analysis of 37 cross-cultural studies by Elfenbein & Laukka (2021) further described the suggested in-group advantage in their meta-analysis, where they compared several studies on the accuracy of emotion recognition. They found that emotions were perceived more accurately in studies where emotions were expressed and recognized by the members of the same culture. The study observes that increased cultural distance between two languages often results in decreased accuracy in speech emotion recognition. These findings support the claim that cultural and linguistic differences play an essential role in the production and the perception of emotions in speech.

As languages from the same language family tend to share many linguistic features (Harbert, 2006), we are interested in whether speakers of two languages from the same family perceive the emotions in each other's (foreign) language differently. This paper focuses on how language distance influences emotion recognition and examines the differences in the perception of emotional speech in two languages from the same language family (Germanic). We decided to focus on two relatively close languages, both culturally and linguistically, German and English, that are not mutually understandable.

To test how the perception of emotions in speech varies across two closely related languages, we decided to use two emotional speech databases, RAVDESS and EMO-DB. However, as they do not contain the same set of emotions, we only focus on the six that they share: happiness, sadness, disgust, anger, fear, and neutral. Both databases involve actors expressing emotions by adapting lexically neutral phrases' acoustic and prosodic features, either in American English (RAVDESS) or in the German language (EMO-DB). The experiment involves participants listening to multiple recordings of each emotion in both languages. After each recording, they are asked to choose which of the six emotions they heard. In the following sections, we compare the accuracy of German speakers listening to German stimuli versus English stimuli and vice versa. We only consider the version of American English used in RAVDESS and the version of German used in EMO-DB, both referred to as 'neutral'.

1.1. Differences and similarities between German and English

German and English belong to the Germanic language family, so they share several linguistic and phonological features. Such similarities involve, for instance, specific intonation patterns (e.g., stress), large vowel inventories including the length contrast, the presence of schwa, the absence of contrast in vowel nasalization, the contrast in voicing in plosives and fricatives (without gaps), and the Trochaic rhythm type. On the other hand, there is a difference in the word order. While English has a fixed SVO word order, German classifies as a language with two dominant word orders, as it also allows SOV. Moreover, according to WALS, German has twice as many cases (4) as English (2). Rounded front vowels, such as [y] and [ø], are present only in German, as well as uvular continuants. Although English lacks uvular consonants, it has a unique [θ] sound, unlike German. Both languages have complex syllable structures and make use of consonant clusters in both onset and coda (WALS, 2013). The (dis)similarities are summed up in Table 1. The bold features represent the contracts.

Table 1 The comparison of English and German, according to WALS

	ENGLISH	**GERMAN**
Vowel inventory	Large	Large
Rounded vowels	no	[y] and [ø]
Schwa	yes	yes
Nasalization contrast	no	no
Voicing contrast in fricatives and plosives	yes	yes
Uvular consonants	no	yes
Uncommon consonant	[θ]	no
Trochaic rhythm	yes	yes
Dominant word order	SVO	SVO & SOV
Number of cases	2	4
Syllable structure	complex	complex

English and German are referred to as stress-timed languages, unlike, for instance, Spanish, which is syllable-timed. The difference lies in the syllable intervals, where in stress-timed languages, the unstressed syllables balance out the stressed ones, which are generally longer. The syllables of a syllable-timed language have an approximately similar duration (Grabe, 1998). The falls in the intonation of short declarative sentences of English and German, which are also included in our stimuli, have been found to be almost identical. In fact, the tonal inventories of these two languages are relatively similar (ibid). Therefore, we presume that the participants' emotion recognition will be less biased by tonal variation than in previous studies (e.g., Pell, et al., 2009).

Scherer et al. (2001), who investigated the accuracy in German emotion recognition by comparing native speakers of eight languages, found that German speakers performed highest of all tested languages (German, English, Dutch, Italian, French, Spanish, Bahasa Indonesian). In addition, they found that there was a positive correlation between the performance and the closeness of the native language to the Germanic language family. After German speakers, the Dutch and the English scored with the highest accuracy rates, followed by speakers of the Romance languages. The non-Indo-European language speakers obtained the lowest score. Based on these findings, we believe that the smaller the distance between native and foreign languages, the smaller the difference between the score obtained in emotion recognition in the native and foreign stimuli; thus, the greater the extent of understanding emotions in the foreign language.

The following paper looks at native speakers of two closely related languages recognizing emotions in each other's languages. We composed an "alternative hypothesis" that there is, *at best, a small difference* between English and German speakers' perception of emotions in the foreign language (English for German and German for English) (H1). According to our prediction, in this study, the difference between correct answers in foreign and native stimuli will be similar in both English and German groups, unlike in previous studies on linguistically and culturally distant languages (e.g., Pell et al., 2009), due to the proximity of German and English (P1). The German speakers in our experiment are expected to show a tendency to perform slightly better on the foreign stimuli than the English participants, because of their common knowledge of English (unlike English speakers' knowledge of German) (P2). If this turns out true, it will be seen under the main effects of participant language, and it will be one of the causes of the, *at best, a small difference* between the two groups. In addition, we predict that sadness and neutral emotion will be confused with each other more than with any other emotion because of the lower and less variable pitch contour with which they are characterized (Busso, C., Lee, S. & Narayanan S. S., 2007). On the contrary, we expect the emotions with high activation levels, such as happiness and anger, to be recognized most accurately (ibid) (P3).

2. Methodology

2.1. Participants

The data is obtained from twenty participants, ten German and ten English speakers. At first, we preferred only participants who were highly proficient only in their native language, as it could potentially increase their performance in the perception task. However, we are aware that German speakers have higher knowledge of English than English speakers of German, especially within the tested generation (20-30 years old). Since most German native speakers in the required age group are already familiar with English, we decided to adjust the analysis and consider the knowledge of English as an element of natural variation. Nowadays, it would be almost impossible to find young German speakers without any knowledge of

English. Our experiment excluded Swiss speakers due to frequent exposure to multiple languages.

Ultimately, we ended up testing German speakers raised in Germany, most of whom had already mastered a certain level of English. The English group consisted of monolinguals of one of the multiple variants of English (Irish, British, American, Caribbean). All were recruited through personal connections to avoid dealing with substantial social, educational, and economic discrepancies. We decided rather test a smaller number of participants and be in control of their social status rather than to test a large number of participants and risk getting imprecise results caused by inconsistency. The anonymity has remained preserved. Every person who took the perception test was not younger than 20 nor older than 30 years old (i.e., born between 1992 and 2002).

Two crucially important factors were the place where the listeners grew up and which languages they were exposed to during their upbringing. Each participant was asked to answer three background questions in a short pre-experimental questionnaire.

2.2. Background questions

Our monolingual participants had to answer three questions before the beginning of the experiment. Depending on their native language, the questions were translated to either English or German, as well as the rest of the instruction. The answer to all three background questions was ‘Yes’/ ‘Ja’ or ‘No’/ ‘Nein’. The following questions were included in the survey:

1. English: Were you born between 1992 and 2002?
   German: Bist du zwischen 1992 und 2002 geboren?
2. English: Did you grow up in an English-speaking country?
   German: Bist du in Deutschland oder Österreich aufgewachsen?
3. English: Are you fluent in any other language than English?
   German: Sprichst du noch eine andere Sprache als Deutsch fließend?

If participants answer ‘Yes’/ ‘Ja’ to questions 1 and 2 and ‘No’/ ‘Nein’ to question 3, they proceeded to the perception task. If one of the first two questions was responded to ‘No’/

‘Nein’, or if the answer to the last question was ‘Yes’/ ‘Ja’, the participants were able to continue but they were excluded from the analysis later on.

2.3. Stimuli

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) consists of 7356 recordings collected from 24 professional actors of both genders with “neutral” North American accents. They express eight emotions (calm, happy, sad, angry, fearful, surprised, disgust, and neutral) through two “neutral” and lexically matched phrases (“Kids are talking by the door.” and “Dogs are sitting by the door.”). Each emotion is demonstrated in average and strong intensity, except for the neutral one. The database provides audio and video files, but we only used the audio recordings (e.g., Audio_Speech_Actors_01-24.zip, 215 MB) (Livingstone & Russo, 2018).

The Berlin Database of Emotional Speech (EMO-DB) contains 535 recorded utterances spoken by ten actors (five males, five females) in a “neutral” German accent, who express seven emotions (anger, boredom, anxiety, happiness, sadness, disgust, and neutral) in an explicit manner through 10 different phrases. Additionally, there are multiple versions of EMO-DB. The first was created between 1997 and 1999 (*Berlin Database of Emotional Speech*). We used the version published by Kaggle updated in 2020 (https://www.kaggle.com/). Therefore, we consider the quality of the recordings of both databases to be relatively similar.

The following experiment involves a mix of recordings from both databases. Each participant listened to 36 recordings in each language (72 in total), with a short break in the middle. They were presented with 6 emotions expressed by 8 speakers (4 males and 4 females) through all types of sentences from both databases. To ensure that the participants focused their attention on the emotional characteristics as much as possible, we used only voices demonstrating the strong intensity in RAVDESS (EMO-DB contains only explicitly expressed emotions). Below are listed all phrases used in the databases. In the German cases, we provided them with English translations.

Phrases:

RAVDESS (*The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)*):

a) Kids are talking by the door.

b) Dogs are sitting by the door.

EMO-DB (*Berlin Database of Emotional Speech*):

a) Der Lappen liegt auf dem Eisschrank.
The tablecloth is lying on the fridge.

b) Das will sie am Mittwoch abgeben.
She will hand it in on Wednesday.

c) Heute abend könnte ich es ihm sagen.
Tonight I could tell him.

d) Das schwarze Blatt Papier befindet sich da oben neben dem Holzstück.
The black sheet of paper is located up there besides the piece of timber.

e) In sieben Stunden wird es soweit sein. [sic]
In seven hours it will be.

f) Was sind denn das für Tpten, die da unter dem Tisch stehen?
What about the bags standing there under the table?

g) Sie haben es gerade hochgetragen und jetzt gehen sie wieder runter.
They just carried it upstairs and now they are going down again.

h) An den Wochenenden bin ich jetzt immer nach Hause gefahren und habe Agnes besucht.
Currently at the weekends I always went home and saw Agnes.

i) Ich will das eben wegbringen und dann mit Karl was trinken gehen.
I will just discard this and then go for a drink with Carl.

j) Die wird auf dem Platz sein, wo wir sie immer hinlegen.
It will be in the place where we always store it.

2.4. Human perception experiment

As all participants are treated as monolinguals, we created two versions of the experiment so that they could read the instructions translated into their language. The perception task was programmed with the help of ir. Dirk Jan Vet (https://www.fon.hum.uva.nl/dirk/). He created a link connected to an online form, so the participants were able to complete the task individually using their computers and headphones. They did not need to look for and download either of the databases, and they could easily answer the questions by clicking on buttons (see Figures 1 and 2).

There were two rounds with a break in between. The first round contained 36 recordings in one language and the second round 36 recordings in the other language (72 in total). Half of the German-speaking participants started with the English stimuli, and the other half with the German stimuli. The English speakers were also split in half, and the two groups started with different stimuli. In this way, we wanted to control for possible within-participant divergence in responses caused by the lack of attention towards the end of the experiment: figures 1 and 2 show previews of the options presented to the participants after each recording.

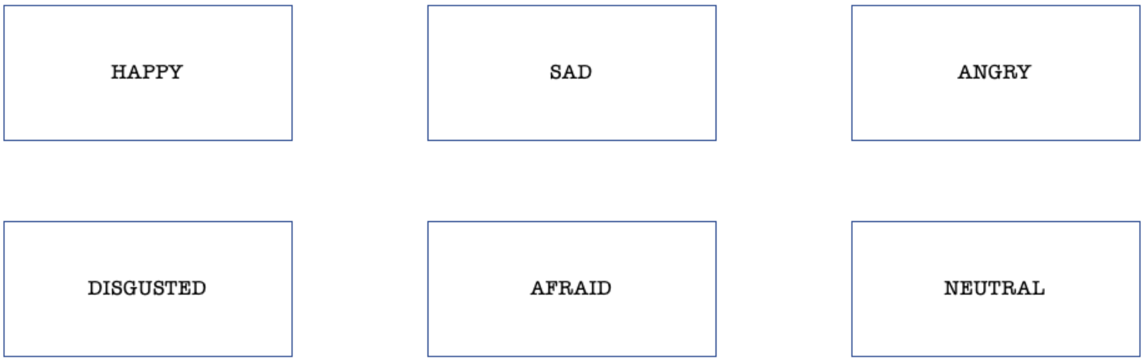


Figure 1 Answer options in English

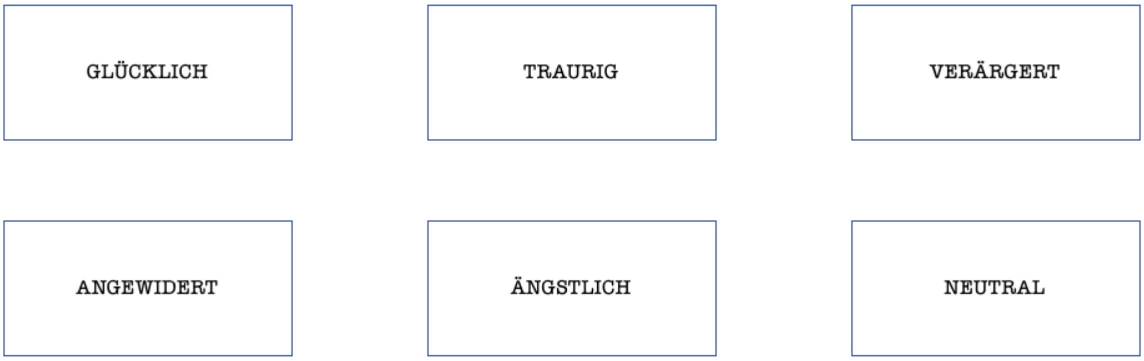


Figure 2 Answer options in German

3. Results

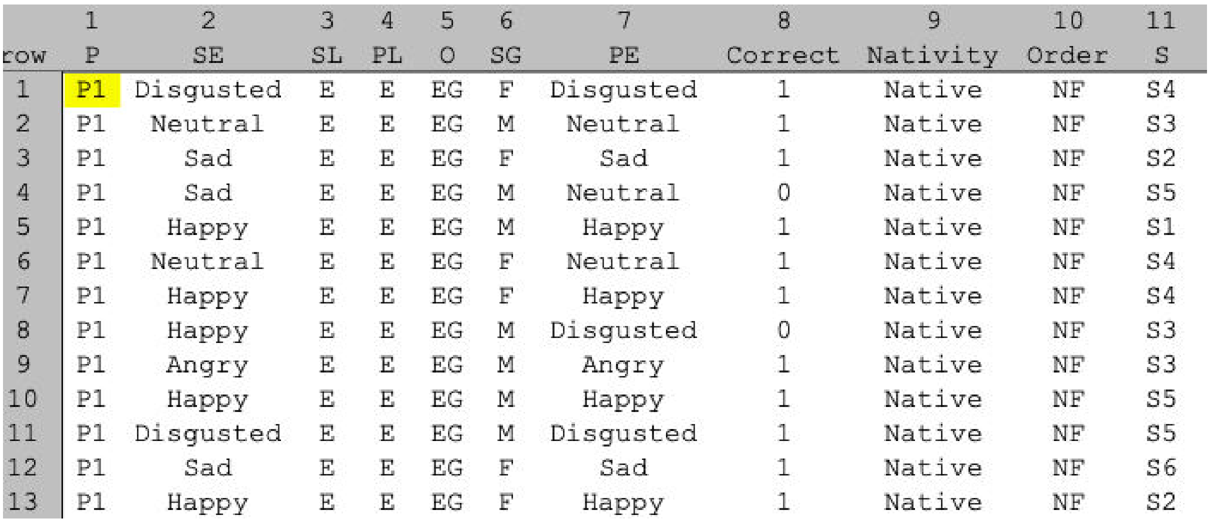
3.1 Data Processing

The data collection took approximately two weeks. The answers were reported by English native speakers of the British (English and Irish) and American (USA and Caribbean) variants and German native speakers only from Germany. Before approaching the participants, Prof. Paul Boersma adjusted the volumes of all recordings in Praat to an absolute peak of 0.99. Each participant was asked to use their own laptop and headphones. The expected time was 10 minutes, but most participants completed the experiment within 7-8 minutes. The individual results were downloaded from the FileGator server. ir. Dirk Jan Vet ensured the files contained the seven columns relevant to the RQ: P, SE, PE, SL, PL, O, and SG. The data processing consisted of extracting these columns and analysis in R Studio (see next section).

The subject number (P) was assigned to each participant before the start of the experiment, as they had to indicate it in the first step. The numbers were based on 'mod 4' (1=4, 2=5, 3=6, 4=7, etc.) because there were four types of participant groups: English speakers starting with English (1), English speakers starting with German (2), German speakers starting with English (3), and German speakers starting with German stimuli (4). The language of the stimuli represented the speaker language (SL) and a within-participant variable.

SE was one of the six emotions that the actor in the recording intended to express (speaker emotion). On the contrary, PE (participant emotion) was the emotion that each subject reported. Out of these two components, we defined the dependent variable named "Correct" with the binary value of 1/0, depending on whether SE and PE matched or not.

The language of the speakers was indicated as SL and one of the participants as PL. There were only two options: E or G. The letter O was used as the indicator of the order with two possible options: EG and GE, depending on which language the participant started with. Based on this, we defined Order as the counterbalancing predictor with also two values FN/NF (foreign/native). With the help of order, we came upon "Nativity", the main predictor that would answer our research question. PL together with SG (speaker gender) served as control predictors (see Figure 3). The speaker number was denoted as S. We added this variable to control for the random effects of the speakers, as some of them might have been easier to understand than others.



	1	2	3	4	5	6	7	8	9	10	11
row	P	SE	SL	PL	O	SG	PE	Correct	Nativity	Order	S
1	P1	Disgusted	E	E	EG	F	Disgusted	1	Native	NF	S4
2	P1	Neutral	E	E	EG	M	Neutral	1	Native	NF	S3
3	P1	Sad	E	E	EG	F	Sad	1	Native	NF	S2
4	P1	Sad	E	E	EG	M	Neutral	0	Native	NF	S5
5	P1	Happy	E	E	EG	M	Happy	1	Native	NF	S1
6	P1	Neutral	E	E	EG	F	Neutral	1	Native	NF	S4
7	P1	Happy	E	E	EG	F	Happy	1	Native	NF	S4
8	P1	Happy	E	E	EG	M	Disgusted	0	Native	NF	S3
9	P1	Angry	E	E	EG	M	Angry	1	Native	NF	S3
10	P1	Happy	E	E	EG	M	Happy	1	Native	NF	S5
11	P1	Disgusted	E	E	EG	M	Disgusted	1	Native	NF	S5
12	P1	Sad	E	E	EG	F	Sad	1	Native	NF	S6
13	P1	Happy	E	E	EG	F	Happy	1	Native	NF	S2

Figure 3 Preview of processed data in Praat

3.2 Statistical results

Each of the 20 participants submitted 72 answers, which gave us 1440 as the total number of responses. The code for the statistical analysis was created by Prof. Paul Boersma and run through R Studio. Since the response is binary, the most appropriate model for our methodology was the generalized linear mixed-effects model (lme4: glmer) (see Figure 4). The output reports one error message: “boundary (singular) fit: see ?isSingular”, but as it does not affect the analysis we continued despite that.

```
nativity.contrast <- cbind (c(-0.5, +0.5))   # Foreign = -0.5; Native = +0.5
colnames (nativity.contrast) <- c("-F+N")
contrasts (table$Nativity) <- nativity.contrast
contrasts (table$Nativity)

language.contrast <- cbind (c(-0.5, +0.5))   # E = -0.5; G = +0.5
colnames (language.contrast) <- c("-E+G")
contrasts (table$PL) <- language.contrast
contrasts (table$PL)

gender.contrast <- cbind (c(+0.5, -0.5))   # F = +0.5; M = -0.5
colnames (gender.contrast) <- c("-M+F")
contrasts (table$SG) <- gender.contrast
contrasts (table$SG)

order.contrast <- cbind (c(+0.5, -0.5))   # FN = +0.5; NF = -0.5
colnames (order.contrast) <- c("-NF+FN")
contrasts (table$Order) <- order.contrast
contrasts (table$Order)

model <- lme4::glmer (Correct ~ Nativity * (SG * PL + Order) + (Nativity * SG | P) +
(Nativity * (PL + Order) | S), data=table, family=binomial, control =
lme4::glmerControl(optCtrl = list(maxfun = 1e5)))
summary (model)
```

Figure 4 R Studio lme4:: glmer model code

To address the question of the role of native language in emotion recognition, we first needed to determine the main predictors: Nativity (nativity.contrast), PL (language.contrast), SG (gender.contrast), and Order (order.contrast). Each value of a binary variable carries a plus or a minus marker. The plus marker specifies the predicted direction (e.g., +G = German-speaking listeners expected to perform better (P2)). The random effects of P are included in the model for Nativity and SG, and the random effects of S for all within-speaker predictors (PL, Order, Nativity) (see Figure 4). Because of that, a generalization is realizable from both samples (P and S). Since we used the restricted form of items (neutral and long enough sentences), we consider our analysis ungeneralizable to other types of sentences. Throughout the following sections, the 95% confidence interval is used to establish the reliability of the results.

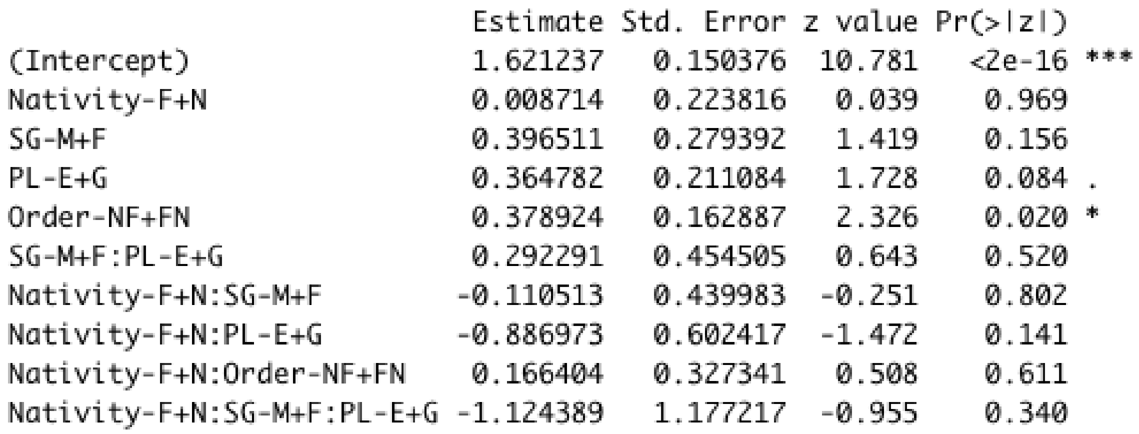
3.2.1. Main effects

All the predictors described below are categorical and binary. Figure 5 illustrates the fixed effects as the result of the statistical analysis. In general, the z-values of all parameters except 'Order' (z-value = 2.33) are inside of the critical range of 95% confidence interval (i.e., <-1.96 or >1.96). The effects of SG and PL on the number of correct answers have not been found. The only predictor with a reliable impact on the dependent variable was the 'Order'.

Figure 6 displays the odds ratios for all parameters, as well as the lower and upper limits of the confidence intervals. The estimate of the Intercept (1.62) suggests that, theoretically, if none of the predictors (x) influenced the dependent variable 'Correct' (y), the response of the participants would be correct. In descriptive terms, based on the odds ratio of the Intercept (OR [5.05]; CI [3.74, 6.82] Pr(>|z|) = <2e-16), the listeners were more than five times more likely to score correctly than incorrectly ($e^{1.62}$). Therefore, we assume they were not guessing.

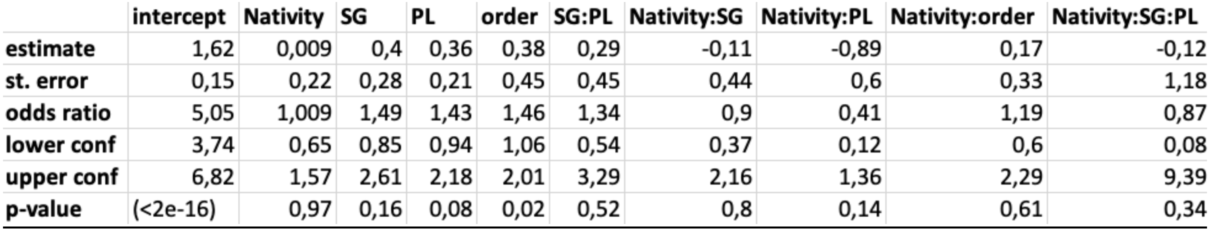
Since the correct answer in the native language was chosen only 1.009 times more likely than in the foreign language, the general Nativity effect is not significant (OR [1.009]; CI [0.65, 1.57]; Pr(>|z|) = 0.97). As a consequence of including random effects of speakers in the analysis, SG did not turn out to have a significant impact on listeners' perception either. However, female speakers seem to be almost 1.5 times more likely associated with the correct perception of emotions than male speakers (OR [1.49]; CI [0.85, 2.61]; Pr(>|z|) = 0.16). Again, due to the wide confidence interval, this result is not generalizable. From the perspective of the listeners, there is an observable tendency for German native speakers to score correctly more often than the English, as we expected (P2). Specifically, the average German listener is estimated to have 1.43 times higher odds of scoring correctly than an average English listener (OR [1.43]; CI [0.94, 2.18]; Pr(>|z|) = 0.08). The confidence interval is not narrow, so this claim cannot be generalized, but since the p-value is not far from the significance border, the results propose a tendency in the direction of our prediction (P2).

There are substantial differences between participants starting with their native and the foreign language (OR [1.46]; CI [1.06, 2.01]; Pr(>|z|) = 0.02). Generally, the listeners who are first presented with a foreign language are almost 1.5 times more successful than the participants starting with their native language. However, this is most probably a result of another variable(s) not included in our study.



```
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                   1.621237   0.150376  10.781   <2e-16 ***
Nativity-F+N                  0.008714   0.223816   0.039    0.969
SG-M+F                        0.396511   0.279392   1.419    0.156
PL-E+G                        0.364782   0.211084   1.728    0.084 .
Order-NF+FN                   0.378924   0.162887   2.326    0.020 *
SG-M+F:PL-E+G                 0.292291   0.454505   0.643    0.520
Nativity-F+N:SG-M+F          -0.110513   0.439983  -0.251    0.802
Nativity-F+N:PL-E+G          -0.886973   0.602417  -1.472    0.141
Nativity-F+N:Order-NF+FN      0.166404   0.327341   0.508    0.611
Nativity-F+N:SG-M+F:PL-E+G -1.124389   1.177217  -0.955    0.340
```

Figure 5 Results of the generalized linear mixed effects model in R Studio



	intercept	Nativity	SG	PL	order	SG:PL	Nativity:SG	Nativity:PL	Nativity:order	Nativity:SG:PL
estimate	1,62	0,009	0,4	0,36	0,38	0,29	-0,11	-0,89	0,17	-0,12
st. error	0,15	0,22	0,28	0,21	0,45	0,45	0,44	0,6	0,33	1,18
odds ratio	5,05	1,009	1,49	1,43	1,46	1,34	0,9	0,41	1,19	0,87
lower conf	3,74	0,65	0,85	0,94	1,06	0,54	0,37	0,12	0,6	0,08
upper conf	6,82	1,57	2,61	2,18	2,01	3,29	2,16	1,36	2,29	9,39
p-value	(<2e-16)	0,97	0,16	0,08	0,02	0,52	0,8	0,14	0,61	0,34

Figure 6 Confidence interval and odds ratio per variable and per interaction

3.2.2. Interaction effects

To find the answer to our research question, it is essential to determine the effect of interaction between PL and Nativity and four other interactions (SG:PL, Nativity:SG, Nativity:order, Nativity:SG:PL). While some covariates seem to influence the *number of correct answers*, we did not find any interaction effects of these predictors on *the dependent variable*. Since the Nativity:PL interaction's confidence interval crosses 1 (OR [0.41]; CI [0.12, 1.36]; Pr(>|z|) = 0.14), there is insufficient evidence to conclude that the German and English groups score similarly on Native and Foreign languages.

Whether the interaction between order and Nativity influences the number of correct answers stays ambiguous (OR [1.19]; CI [0.6, 2.29]; Pr(>|z|) = 0.61), furthermore, due to the wide confidence interval (OR [1.34]; CI [0.54, 3.29]; Pr(>|z|) = 0.52), we also cannot generalize whether or not native speakers of English and German perceive emotions expressed by the two genders differently. The effect of the interaction between Nativity and SG (OR [0.9]; CI [0.37, 2.16]; Pr(>|z|) = 0.8) is also ungeneralizable. Similarly, the interaction between the covariates and the explanatory variable did not turn out to be significant (OR = 0.87; CI [0.08, 9.39]; Pr(>|z|) = 0.34). Due to the largest-scale confidence interval and relatively low odds ratio, we cannot observe any tendencies. The interpretation of our results remains unclear, and our research question stands open.

Figure 7 illustrates the predicted probabilities of 'Correct' per SG and PL. While the y-axis represents the probability percentage of responding correctly, the four points on the x-axis show the two levels of nativity per each group of native speakers. It is almost impossible to predict the likelihood of German native speakers choosing the correct emotion when listening to a male voice. The likelihood of reaching an accurate score when listening to speakers of a particular gender regardless of the language and 'Nativity' cannot be concluded either, due to the small number of participants. However, an increase in sample size could solve this problem.

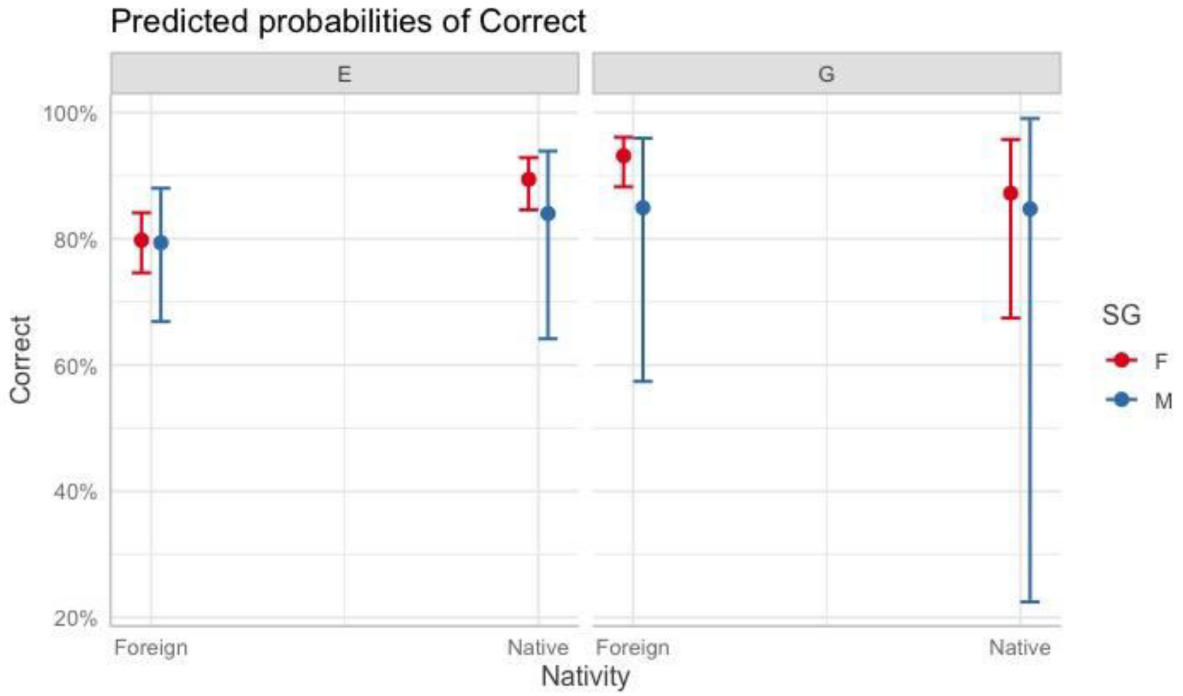
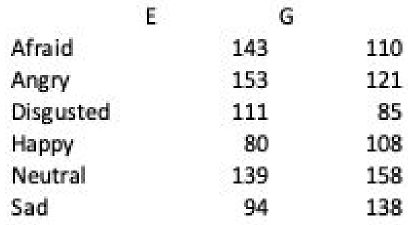


Figure 7 Predicted probabilities of "Correct"

3.3 Emotion recognition

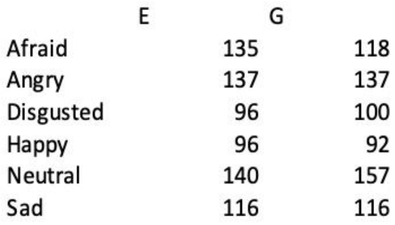
Figures 8 and 9 were created using R Studio's xtabs function. In Figure 8, the emotions reported by participants (PE) are listed on the left, and the speaker language (SL) above. The stimuli consisted of all six emotions expressed three times. The numbers reflect the recognition accuracy of all participants when listening to the corresponding language. The higher the number, the more successfully the listeners guessed the correct emotion. The neutral emotion in German was generally most evident. Still, the performance in each emotion could have been influenced by other factors, such as 'Order' or the speaker's ability to express the emotion.

Alternatively, Figure 9 illustrates the aptitude of the native speakers of English and German to recognize emotions in speech regardless of SL. The neutral emotion was the easiest to understand for both groups of listeners. The sum of correct answers per emotion in Figure 8 corresponds to the sum of correct answers in the same emotion in Figure 9, representing the total number of correct responses per emotion.



	E	G
Afraid	143	110
Angry	153	121
Disgusted	111	85
Happy	80	108
Neutral	139	158
Sad	94	138

Figure 8 Participant emotion per speaker language (PE-SL)



	E	G
Afraid	135	118
Angry	137	137
Disgusted	96	100
Happy	96	92
Neutral	140	157
Sad	116	116

Figure 9 Participant emotion per participant language (PE-PL)

Figure 10 (below) represents the recognition accuracy in more detail, specifically by including 'Nativity'. Under the x-axis, the first letter of each group denotes the language the group started with, and the second letter represents the participant language (e.g., EE = English native speakers starting with English). Each of the six tested emotions is depicted in a different colour. Every emotion was played six times in both languages and listened to by all four groups. Each group consisted of 5 subjects (6 emotions x 6 times x 2 languages x (4 groups x 5 subjects)), resulting in a total number of 1440 submitted answers. Every group reported 360 responses (5 participants x 72 recordings), and every emotion was answered 60 times within each subject group (360 answers/ 6 emotions, or 5 participants x 6 emotions x 2 languages). For example, the dark blue column in EE illustrates that the 5 subjects of this group selected the emotion 'Afraid' 58 times correctly, and only two times incorrectly when they confused it with the 'Angry' emotion. According to the chart in Figure 10, the 'Angry' and ‘Neutral’ emotions were selected accurately most often: every 227 times out of 240 trials (95%). ‘Afraid’ scored second with 201 correct answers (84%), and ‘Sad’ followed with 184 successful recognitions (77%). The least recognizable were the ‘Disgusted’ and ‘Happy’ emotions, as both obtained only 166 correct reactions (69%).

English native speakers found it most challenging to recognize ‘Disgusted’ and reacted correctly only 76 times out of 120 trials (63%). The easiest was for them ‘Angry’, in which they scored with almost 93% accuracy. On the other hand, the German-speaking participants reported the correct answer for ‘Neutral’ as much as 118 out of 120 times (98%), unlike for ‘Happy’, in which they scored lowest (69%).

A more detailed analysis of the two most precisely perceived emotions, ‘Neutral’ and ‘Angry’, revealed that the emotions most often confused with ‘Neutral’ was ‘Sad’, and with ‘Angry’, it was ‘Disgusted’. The listeners who misjudged ‘Neutral’ selected ‘Sad’ 80% of the time, and in the case of ‘Angry’, they chose ‘Disgusted’ 58% of the time. However, the misjudgement occurred relatively rarely, in both cases, only 5% of the time. On the contrary, in the cases of the misjudgement of ‘Sad’, in 43% of the trials, the emotion was reported as ‘Neutral’ instead, and in 46% as ‘Afraid’. As for the least accurately recognized emotions, ‘Disgusted’ was perceived as a different emotion in 31% of the trials, where it was reported as ‘Angry’ 35% of the time and as ‘Sad’ 32% of the time.

In general, German native speakers reached the accuracy in 83% of the trials, while English native speakers in only almost 80% of the cases. The best overall score was gained by the German-speaking group starting with English. The 5 participants recognized the correct emotions 303 times out of 360 trials (84%). At the same time, it was the only group that accurately reported the intended emotion 60 out of 60 times, and not once, but twice (‘Angry’ and ‘Neutral’). As a result of this remarkable performance among German-speaking listeners, the above-mentioned statistical results in Figures 5 and 6 portray a positive p-value in the case of PL, although still insignificant.

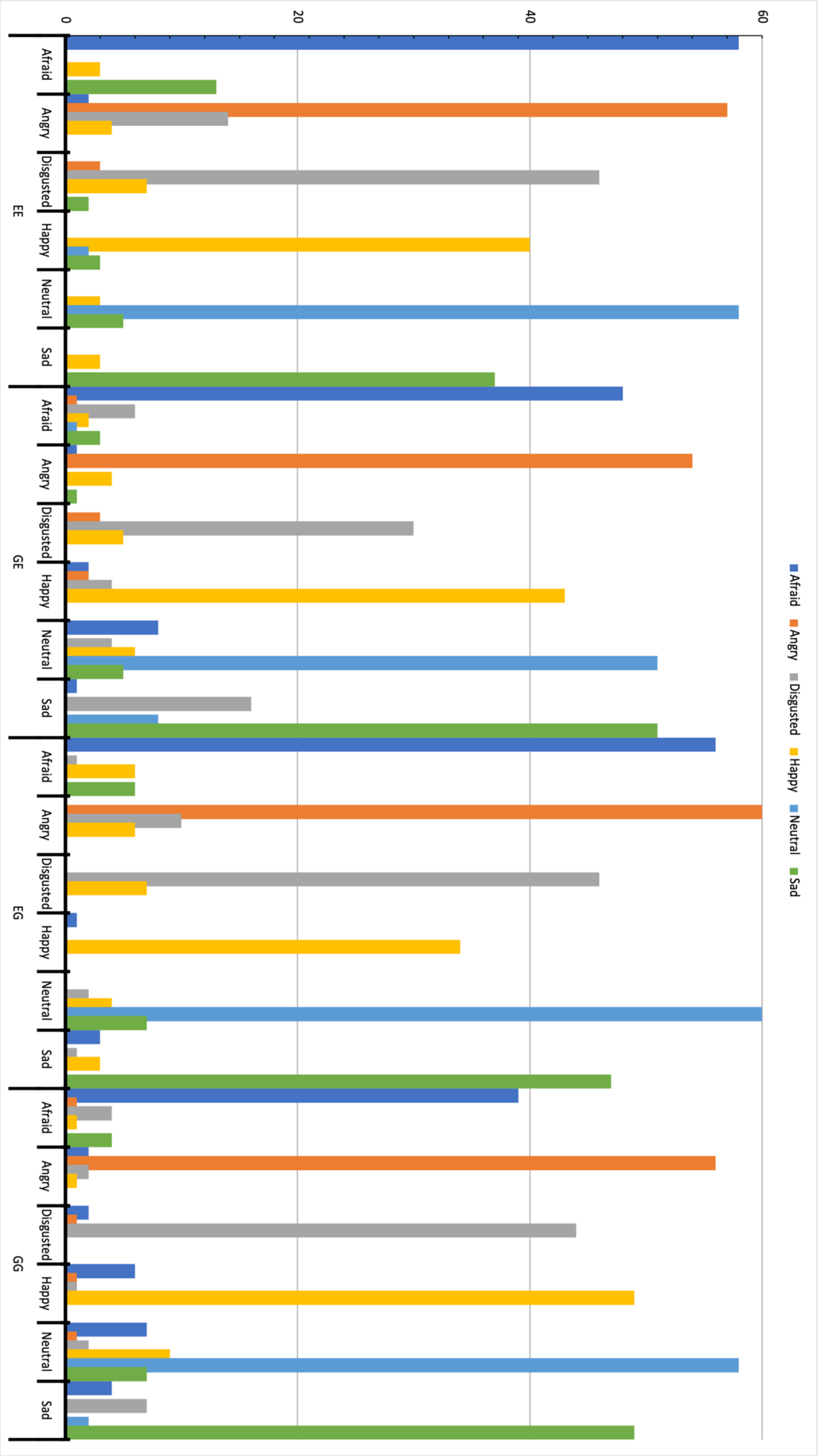


Figure 10 Reversed confusion matrix: Recognition accuracy per emotion per participant group

4. Discussion

This experiment was conducted to determine the effect of the previously found in-group advantage on speech emotion recognition (Elfenbein & Ambady, 2002; Althoff et al., 2016; Elfenbein, Laukka, 2021). We tried to prove that the advantage of a native language becomes minimal or disappears when recognition happens in a closely related language to the native. The results of this experiment can be interpreted in several ways. As presented in section 3, the presence of the Nativity effect in our German-English sample has not been found. However, the confidence intervals were too large-scale to draw conclusions. Although we did expect the absence of the in-group advantage on native stimuli (i.e., a negligible difference in the score of correct answers) (H1), the outcome does not confirm our prediction, according to which the interaction between PL and Nativity is significant (OR [0.41]; CI [0.12, 1.36]; Pr(>|z|) = 0.14). We believe that the substantial reason behind the insignificance of the data was caused by the inconsistency of answers among the twenty participants. The only generalizable result of a main effect is the one of 'order' (OR [1.46]; CI [1.06, 2.01]; Pr(>|z|) = 0.02). It is unclear why participants who started with their native language performed worse than those beginning with the foreign language; however, the topic remains open to future research. We assume that this output was caused by another variable(s) not included in our study, such as more effort put into the experiment when starting with foreign stimuli. The effect of interaction between ‘order’ and ‘Nativity’ on ‘correctness of answers’ also remains a question (OR [1.19]; CI [0.6, 2.29]; Pr(>|z|) = 0.61).

The German and the English native speakers' perceptions of emotions in their native and the foreign languages cannot be considered similar due to the large-scale confidence interval (PL: OR [1.43]; CI [0.94, 2.18]; Pr(>|z|) = 0.08). However, the English participants tended to be affected by Nativity more than the German participants, which points in the direction of our expectation about German native speakers performing correctly on the foreign stimuli more often than the English participants (P2). Despite this observable tendency, the size of the confidence interval does not allow us to generalize this phenomenon to the whole population.

Another surprising phenomenon observed in this study was the relatively low confusion rates between the neutral and the sad emotions. In fact, 'Neutral' was selected accurately 95% of the time, which suggests that the actors of both databases conveyed the emotional content in these cases most clearly. Supposing that 'Angry' and 'Happy' have similarly high activation levels observed through the pitch contour, it does not seem to play an essential role in our research, as 'Happy' belonged to the least often correctly recognised emotions, while 'Angry' the opposite. Since the statistical results of the emotion recognition (see 3.2.1. Emotion recognition) do not claim 'Happy' to have the highest, and 'Sad' and 'Neutral' to have the lowest recognition rates, our study does not support the findings of Busso, Lee, & Narayanan (2007).

However, we found certain limitations that could have potentially influenced the study's outcome. Firstly, as in every online experiment, the control over participants was insufficient, and we relied on their comprehension of the instructions. Secondly, the statistical results were partially affected by the variation between the databases and their restricted sets of emotions. One of the differences lies in the diversity of the actors in the recordings and their skills. As the databases were created independently, they also contain non-identical neutral phrases, and four out of six emotions were negative. Although the volumes of the recordings were modified to make them less inconsistent, there were noticeable differences in silences across the sentences (especially in the German recordings). Moreover, we encountered several unavoidable natural effects, such as participants' background or speakers' accents. Even though both databases consist of 'neutral' accents, we believe the neutral accent is undefinable. In addition, the perception could have been influenced by phonetic cues, which some listeners could have had better than others. Finally, participants' current exposure to other languages was not controlled, although based on the answers to the background questions numbers 2 and 3, we presume they were exposed to German or English most of their lives.

Despite all limitations, the listeners recognised the emotions from the recording accurately most of the time, which indicates that although there were several differences between the databases, the recordings were comparably understandable, and the task was straightforward. However, it does not mean there is no space for improvement. Further research could enhance the analysis by extending the scope of the current research, more specifically by collecting data from more participants listening to more sentences by more speakers. Additionally, in the future, it would be interesting to look at the differences between the perception of positive and negative emotions. The emotion set in this paper was chosen from the shared series of emotions in two different databases. For this reason, we ended up analysing the majority of negative emotions, which could have also influenced the outcome. It would be also possible to extend the analysis by including emotions as predictors. We see potential in studying two linguistically related but culturally distant languages, such as Dutch and Afrikaans, to determine the effects of culture on the perception of emotions in speech. Furthermore, we believe natural speech would also bring different results than artificial recordings.

5. Conclusion

Unlike previous studies on the perception of emotions in speech that involved languages from different families, our study focused on the same phenomenon in a contrasting way. The paper's ultimate goal was to investigate the in-group advantage found in speech emotion recognition. We tested English and German native speakers' perception of emotions in each language and compared their results. As English and German are two closely related languages culturally and linguistically, we expected to find *at best, a small difference* between English and German speakers' perception of emotions in the foreign language (English for German and German for English) (H1).

To determine whether the stimuli's language affects the listeners' performance in recognition, we extracted six emotions from already existing online databases RAVDESS and EMO-DB. Both databases contain recordings of an artificial speech by professional actors. The participants were tested through an online experiment, where each of them listened to 36 recordings of short neutral sentences in each language (72 in total). They were asked to select one of the following emotions they heard in each recording: 'Angry', 'Afraid', 'Happy', 'Neutral', 'Disgusted', 'Sad'. The results of this study do not provide evidence for or against the hypothesis (H1). We are not able to generalize the outcome due to large-scale confidence intervals caused by the small-scope sample. Yet, we observed some tendencies that future research could build upon. For instance, not all emotions with supposedly higher pitch contour were recognised more accurately. On the contrary, 'Neutral' was perceived with the highest accuracy. When analysing the effect of Nativity, we noticed that the English native speakers tended to be influenced by it more than the German participants. The difference between correctly selected answers in English group’s native language and the foreign language was greater than in the German-speaking group. Another observation was the proneness of the participants to score higher when starting with foreign and continuing with their native language. In fact, even though the order of the languages did not affect the correctness of the answers, we consider this observation interesting. A larger sample could uncover the possibly existent impact on the overall performance of the participants.

Nevertheless, we believe that the close relatedness of English and German could be one of the causes of the mixed results of our experiment.

In conclusion, language distance and cultural closeness play an essential role in recognising emotions in speech. By further investigation of these effects, we can reach a better understanding of how our background affects the way we communicate.

References

Althoff, J., Laukka, P., Elfenbein, H. A., Thingujam, N. S., Rockstuhl, T., Iraki, F. K., Chui, W. (2016). *The expression and erecognition of emotions in the voice across five nations: A lens model analysis based on acoustic features*. Journal of Personality and Social Psychology, Vol. 111, No. 5., 608-705.

Berlin Database of Emotional Speech. Retrieved from (http://emodb.bilderbar.info/docu/)

Busso, C., Lee, S. & Narayanan S. S. (2007). *Using neutral speech models for emotional speech analysis*. University of Southern California. Los Angeles.

Elfenfbein, H. A. & Ambady, N. (2002). *On the universality and cultural specificity of emotion recognition: A meta-analysis.* Washington University of St. Louis. Psychological Bulletin, Vol. 128, No. 2, 203-235.

Elfenbein, H. A. & Laukka, P. (2021). *Cross-cultural emotion recognition and in-group advantage in vocal expressions: A Meta-Analysis.* Emotion Review, Vol. 13, No. 1, 3-11

EMO-DB Dataset. Retrieved from (https://www.kaggle.com/piyushagni5/berlin-database-of emotional-speech-emodb)

Grabe, E. (1998). *Comparative intonational phonology: English and German.* Radboud University Nijmegen, Nijmegen.

Harbert, W. (2006). *The Germanic languages*. Cambridge University Press.

Kamaruddin, N., Wahab, A., & Quek, C. (2012). *Cultural dependency analysis for understanding speech emotion*. Expert Systems with Applications, 39(5), 5115-5133.

Livingstone S. R., Russo F. A. (2018). *The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English*. PLoS ONE13(5): e0196391.

Mesquita, B., Frijda, N. H., & Scherer, K. R. (1997). Culture and emotion. In J. E. Berry, P. B. Dasen, & T. S. Saraswathi (Eds.), *Handbook of cross-cultural psychology: Vol. 2. Basic processes and developmental psychology* (pp. 255-297). Boston: Allyn & Bacon.

Pell, M.D., Monetta, L., Paulmann, S. et al. (2009). *Recognizing emotions in a foreign language*. https://link.springer.com/article/10.1007/s10919-008-0065-7

Scherer, K. R., Banse, R., & Wallbott, H. (2001). *Emotion inferences from vocal expression correlate across languages and cultures*. Journal of Cross-Cultural Psychology, 32, 76–92.

Schuller, B. W. (2018). *Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends*. Retrieved from https://cacm.acm.org/magazines/2018/5/227191 speech-emotion-recognition/fulltext

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). Retrieved from (https://zenodo.org/record/1188976#.YrRwtuxBzX8)

The science of emotion: exploring the basics of emotional psychology. (2019). Retrieved from https://online.uwa.edu/news/emotional-psychology/

WALS online (2013). In Dryer M. S., Haspelmath M. (Eds.). Leipzig: Max Planck institute for evolutionary anthropology. Retrieved from https://wals.info/

Wiese, R. (1996). *The phonology of German.* Oxford University Press