To be presented at the 20th International Congress of Phonetic Sciences, Prague, Czech Republic 2023

"KRIECH NICHT DA REIN!" - A NEW CORPUS OF NATURALISTIC MISPERCEPTION OF GERMAN MISHEARD SUNG SPEECH

Jessica Nieder[1] & Kevin Tang[2*]

[1]Department of Linguistics & [2]Department of English Language and Linguistics

Heinrich-Heine University Düsseldorf, Germany; [2]Dept. of Linguistics, University of Florida, USA

nieder@phil.hhu.de, kevin.tang@hhu.de

ABSTRACT

Naturally-occurring misperception [1] can help establish the ecological validity of laboratory findings of speech perception and generate new hypotheses. In this study, we report on a corpus of misheard German sung speech which contains instances of misperception reported by individuals. We validated the corpus by examining segmental confusions, and word mis-segmentation. Approximately 1,000 segment confusions were found. Our naturalistic segment confusions were significantly correlated with acoustic distances (r = 0.559) and with speech-in-noise-induced confusions in an experimental study (vowel: r = 0.364; consonant: r = 0.210). Our mis-segmentation patterns only partially confirmed the rhythmic segmentation hypothesis [2] and findings from previous studies. While boundaries inserted before strong syllables created content words following the preferred rhythmic properties of German, we find an unexpected amount of boundary deletion before strong syllables, resulting in nonce percepts which might reflect the expectation of listeners with neologisms in lyrics [3].

Keywords: German lyrics, mondegreen, slips of the ear, phonetic similarity, rhythmic segmentation.

1. INTRODUCTION

In conversations, listeners are faced with the task of identifying speech segments and segmenting the speech stream they receive into meaningful words. When communication operates smoothly, one can only assume that the listener's understanding is identical to the utterance of the speaker, but it is only when communication breaks down that we are able to disentangle what was said from what was received [4]. These naturally-occurring instances of speech misperception or *slips of the ear* have the potential of revealing our speech perceptual processes and have been the focus of corpus-based or experimental studies over the past years, e.g., [1, 5, 2, 6, 7, 8, 9]. Using a corpus of naturalistic misperception of spoken English [1] demonstrated how naturalistic misperception can help establish the ecological validity of experimental studies (e.g., [10, 11, 12]). However, compared to spoken speech, the misperception of sung speech is much less well researched. Listening to sung speech is a very different activity as compared to listening to a conversation due to words and melodies being processed interactively [13]. When listening to sung speech, the nature of the context is more complex; it involves not only the listener's immediate surrounding and the activity that the listener is doing while listening, but also the general expectation of the artist and the genre of the song containing the sung speech [14]. Moreover, sung speech is produced differently from spoken speech in terms of duration, vowel formants [15], intonation [16] and pronunciation. It is almost always masked with music, while spoken speech may be masked with different noise types.

In this study, we present a new corpus of naturalistic misperceptions of misheard song lyrics by focusing on misperceptions of sung speech in German by German native speakers. We demonstrate how such a naturalistic corpus can shed light on patterns in segmental confusions, and word mis-segmentation. We examine the patterns in the light of existing findings from experimental and corpus-based studies. Our data is openly available at https://osf.io/xajvf/ and https://osf.io/acqp6/.

1.1. Speech segmentation in German

For English, a stress-based language, it has been established that listeners make use of their knowledge of the opposition of strong, i.e., stressed, vs. weak, i.e., unstressed syllables for inserting or

*We gratefully acknowledge Gerrit Kentner for sharing his data [18], the data contributions of the students in the linguistic seminar, the research assistance of Lukas Ostrowski, and the feedback of Ulrike Domahs and our anonymous reviewers.

deleting word boundaries in speech segmentation [5, 2, 11]. As a result, the rhythmic segmentation hypothesis was proposed that predicts a deletion of boundaries before weak and an insertion of word boundaries before strong syllables for English mis-segmentation data [2]. In German, the preferred rhythmic pattern is trochaic. Polysyllabic words show at least one stressed syllable, while stress in monosyllabic words shows an opposition between grammatical vs. lexical words: monosyllables are stressed in lexical words but can be unstressed in grammatical words such as clitics [17]. From this it follows that the same pattern of insertions before strong and deletions before weak syllables as compared to the English data is expected [17, 18]. Due to the tendency of unstressed monosyllables being grammatical words in German, a boundary insertion before weak syllables is expected to lead to the creation of a grammatical word [18] in case of mis-segmentations.

2. CORPUS

2.1. Compilation and annotation

For this study, a data set of 176 German misheard song lyrics was collected in a linguistics seminar at Heinrich-Heine University Düsseldorf and from three books containing a collection of German misheard sung speech [19, 20, 21][1]. To examine segmental confusions, we first obtained a phonetic transcription of the data using the CELEX pronunciation dictionary [22]. The transcriptions of the intended lyrics and the misheard lyrics were then segmentally aligned using an automatic alignment method (Pointwise-Mutual-Information-based Levenshtein distance, see [23]) that has been used to successfully align naturalistic misperceptions in English [1].

To examine word mis-segmentation, syllable stress was annotated using CELEX. A syllable was marked as strong if it had primary stress within the word it appears in, otherwise a syllable was marked as weak. Secondary stress was not marked because CELEX does not encode secondary stress and its existence is not well supported by acoustic evidence [24]. Subsequently, the data was coded manually for possible boundary misperception. Following [2, 18] we identified 1) **boundary insertions**, e.g., an additional word boundary in the perceived lyrics without a corresponding word boundary in the intended lyrics and 2) **boundary deletions**, e.g., a missing word boundary in the perceived lyrics that is available in the intended lyrics:

1. Griech - isch - er | Wein (original)
   Kriech | nicht | da | rein (perceived)
   'Greek wine/Do not crawl in there.'
2. Die | Crew | hat | da | noch | Fra - gen (original)
   Mir - ko | hat | da | noch | Fra - gen (perceived)
   'The crew/Mirko still has questions.'

69 data points were left for the mis-segmentation analysis.

3. STUDY I: SEGMENTAL CONFUSIONS

In the first study we focus on evaluating the segmental confusions found in misheard sung speech by comparing them with i) acoustic measurements and ii) speech-in-noise-induced segmental confusions.

3.1. Data for Validation

To validate the role of phonetic similarity in segmental confusions, we extracted the vowel frequency measurements of 16 vowels /a, ɪ, eː, ɛ, ɛː, ɪ, iː, ɔ, oː, ʊ, uː, ʏ, yː, œ, øː, ə/ of 69 male and 58 female speakers from [25]. The Hertz frequencies were converted to Bark scale [26] to better match with human perception. The Euclidean distance was computed for each pair of vowels using the Bark frequencies of the first two formants. The distance was computed for men and women separately before averaging to obtain a single set of distances.

To demonstrate the ecological validity of laboratory research, we extracted the confusion data from two speech-in-noise experiments (one vowel, one consonant) from [27]. Ten listeners with normal hearing were tested in a close-set forced choice phoneme-recognition task. Ten vowels /a, aː, ɛ, e, ɪ, i, ɔ, o, ʊ, u/ were presented in a CVC frame. 14 consonants /p, t, k, b, d, g, s, f, v, n, m, ʃ, ts, l/ were presented in a VCV frame. Five Signal-to-Noise ratios (SNRs) were examined (0, -5, -10, -15, and -20 dB). We focused on the confusion response rates for consonants and vowels at -15 dB SNR because they are least influenced by ceiling effects and have the largest variation across phonemes.

3.2. Results

995 segment confusions were found with 212 insertions, 252 deletions and 531 substitutions. Focusing on the vowel-vowel and consonant-consonant substitutions, two confusion matrices were tabulated and smoothed by adding 0.01 to all cells. The smoothed count matrices were transformed into confusion proportions. The proportion matrices were then converted into similarity matrices using the formula $S_{x,y} = (p_{x,y} +$

$p_{y,x})/2$, where $p_{x,y}$ is the proportion of times that the segment x was perceived as the segment y. The similarity matrices were then converted into distance matrices using $-ln(S_{x,y})$, a well-established metric for estimating the perceptual distance between two segments from similarity [28]. Similarly, the experimental confusion matrices were converted into distance matrices but using a more appropriate similarity metric $S_{x,y} = (p_{x,y} + p_{y,x})/(p_{x,x} + p_{y,y})$.

Pearson correlation was used to compare the global similarity of matrices. The statistical significance was evaluated using the Mantel test [29] with 10,000 permutations (upper-tailed) because sound distances are not completely independent. Our naturalistic segment distances were correlated with acoustic vowel distances (r = 0.559, p = 0.0001) and with the two sets of experimental-induced segment distances (vowel: r = 0.364, p = 0.0064; consonant: r = 0.210, p = 0.0245).

4. STUDY II: WORD MIS-SEGMENTATION

In the second study we focus on native word mis-segmentation and build on what has been shown by [18] for German in a cross-linguistic setting. In doing so, we test the predictions of the rhythmic segmentation hypothesis [2].

4.1. Data for Validation

To validate our findings we compare our data to four previous studies on mis-segmentation: [2] present natural and laboratory-induced mis-segmentations of continuous speech on English data and confirm the validity of the rythmic segmentation hypothesis for English native misperception (henceforth, ENG). Using experimental evidence, [7] investigate mis-segmentation in Dutch and confirm the predictions of the rhythm segmentation hypothesis for their data (henceforth, DUT). Focusing on mis-segmentation of English song lyrics by German native speakers (henceforth, ENG♪-GER♪), [18] shows that the rhythm segmentation hypothesis holds for non-native song perception. [30] presents experimental data for non-native (English-German) and native misperception (German-German). We took the German-German data from the supplementary materials of [30] and coded it for boundary type and syllable strength. This left us with 19 observations (henceforth, GER).

4.2. Results

Overall, we find more boundary deletions than insertions (42 vs. 27), a finding that is also confirmed by [18] for German in a cross-linguistic setting (ENG♪-GER♪). Table 1 displays the results of this study in comparison to the results of the aforementioned previous studies [2, 18, 7, 30].

Focusing on boundary insertions, more insertions before strong syllables are reported for ENG and DUT [2, 7]. For ENG♪-GER♪, the same amount of boundary insertions before strong as compared to weak syllables is found [18]. Interestingly, for German, this study and GER [30] shows the opposite pattern: we find more insertions before weak than before strong syllables. Focusing on boundary deletions, for ENG and ENG♪-GER♪ more deletions before weak than before strong syllables are reported [2, 18]. For DUT, roughly the same amount of boundary deletions before weak as compared to strong syllables is found [7]. For German, this study and GER show the opposite pattern, i.e., more boundary deletions before strong than before weak syllables [30] are observed.

Inspecting the boundary deletions in our data reveals that in 20 of 39 cases the deletion before a strong syllable created a nonce word, e.g., "*Flaggenhof*" for "*Flaggen hoch*" (nonce word vs. 'raise flags'). In addition, all deletions before weak syllables in our data created nonce words (3 of 3 cases). Compared to the deletion data, only 2 of the 27 cases of insertions created nonce words.

Omitting the nonces, and thus taking into account existing percepts only, leaves us with a total of 25 insertions (4 before strong syllables, 21 before weak syllables) and a total of 19 deletions (19 before strong). However, the general pattern of more insertions before weak and more deletions before strong syllables, contrary to the predictions of the rhythmic segmentation hypothesis, remains the same: $\chi^2 = 27.258$, df = 1, $p < .001$.

Another prediction of the rhythmic segmentation hypothesis states that an insertion before a weak syllable leads this weak syllable to likely be a grammatical word [18]. We observe the following pattern in our data of existing percepts that is displayed in Table 2.

In case of boundary insertions before strong syllables, our data supports the rhythmic segmentation hypothesis: If a boundary is inserted before a strong syllable, this strong syllable became a lexical word, e.g., "*Geduld ist ungesund*" (original) vs. "*Der Tod ist ungesund*" (perceived) 'Patience/The death is unhealthy'. However, if a word boundary is inserted before a weak syllable, the data does not support the rhythmic segmentation hypothesis: in only 3 cases of a total of 21 cases the weak syllable became a grammatical word,

Table 1: All boundary insertions and deletions before strong vs. weak syllables in our data as compared to data from four other studies. Chi-square test of independence for the relationship between boundary type and syllable strength: GER♪ (this study): $\chi^2 = 39.368$, df = 1, $p < .001$; GER: $\chi^2 = 0.30536$, df = 1, $p = 0.5805$; ENG: $\chi^2 = 22.484$, df = 1, $p < .001$; DUT: $\chi^2 = 16.208$, df = 1, $p < .001$; ENG♪-GER♪: $\chi^2 = 8.2073$, df = 1, $p < .005$.

	Insertion					Deletion				
	GER♪	GER	ENG	DUT	ENG♪-GER♪	**GER♪**	GER	ENG	DUT	ENG♪-GER♪
Strong	**4**	4	90	101	22	**39**	1	68	72	19
Weak	**23**	14	45	36	22	**3**	0	107	73	63

Table 2: Boundary insertions before weak vs. strong syllables leading to grammatical vs. lexical words within the group of existing percepts only.

	Grammatical	Lexical
Strong	0	4
Weak	3	18

e.g., "*lasset uns gemeinsam*" (original) - "*lasst uns gemein sein*" (perceived) 'Let's [...] together/Let's be mean'. In the remaining 18 cases, the weak syllable became a lexical word, e.g., "*Dies Kind soll unverletzet sein*" (original) - "*Dies Kind soll unser letztes sein*" (perceived) 'This child should be unharmed/our last'. While in these cases the weak syllable became a lexical word, contrary to the prediction, the syllable before the word boundary involves the creation of grammatical words in 7 of 11 cases, as in the following example:

3. wir | fah-ren | auf | Feu - er - rä - dern | Rich - tung | Zu - kunft (original)
   wir | fah - ren | auf | eu - ern | Rä - dern | Rich - tung | Zu - kunft (perceived)
   'we are riding on flaming bikes/your bikes into the future'

Here, the former weak syllable of the nonce word *Feuerrädern* in the original lyrics becomes a strong syllable in the onset of the new lexical word and the resulting percept is a correct grammatical phrase consisting of a grammatical word and a lexical word: *euern Rädern*.

5. DISCUSSION

Study I suggested that vowel confusions from misheard sung speech is strongly influenced by phonetics (r = 0.559). This is a surprising finding, since the confusion matrix was extracted regardless of any phonological environments and words. However, vowel and consonant confusions were only weakly correlated with experimental confusions (rs = 0.210-0.364). This could be due to how the experimental conditions differ greatly from the naturalistic one, such as the lack of music, and that segments were spoken (not sung) and were presented in VCV or CVC frames in isolation. Despite all the potential top-down influences and the influences of music on sung speech, listeners still heavily rely on phonetic/acoustic similarity during the processing of sung speech.

Study II demonstrated that the predictions of the rhythmic segmentation hypothesis cannot explain the pattern in our data. This raises the question as to why German native misperception behaves so differently from the data in previous studies. Against the predictions of the rhythmic segmentation hypothesis, we find more boundary deletions before strong than before weak syllables, often leading to a creation of new word forms or nonces. While the nature of these nonces remains a topic for further studies, a tentative explanation for their occurrence could lie in the importance of affective signals in perception, as has been shown by [14]. The nonces appear to be more humorous than the original lyrics, leading the listener to perceiving a more amusing nonce word instead of the original lyrics. Another reason for the amount of nonces lies in the nature of song lyrics. The production of song lyrics often use innovations such as neologisms for stylistic reasons or to express certain emotions [3]. The familiarity of the listeners in our study with nonces in song lyrics might have lead them to create neologisms themselves when misperceiving the original lyrics. These results open the path for further research on the listener's expectations in speech perception.

Taken together, this study reported on a new corpus of naturally-occurring misperception of sung speech. Despite its relatively small size and all the inherent differences with listening to spoken speech and sung speech, we demonstrated the role of both bottom-up (phonetics) and top-down (lexical expectation) factors in the processing of sung speech. This reinforces the idea of how examining our everyday perceptual errors has the potential to establish the ecological validity of laboratory findings of speech perception and generate new hypotheses [1, 5, 2, 6, 7].

6. REFERENCES

[1] K. Tang, "Naturalistic speech misperception," Ph.D. dissertation, University College London, 2015.

[2] A. Cutler and S. Butterfield, "Rhythmic cues to speech segmentation: Evidence from juncture misperception," *Journal of Memory and Language*, vol. 31, no. 2, pp. 218–236, 1992.

[3] V. Werner, "Love is all around: a corpus-based study of pop lyrics," *Corpora*, vol. 7, no. 1, pp. 19–50, 2012.

[4] J. Laver, "The production of speech," in *New Horizons in Linguistics*, J. Lyons, Ed. Harmondsworth: Penguin, 1970, pp. 53–75.

[5] K. Tang and A. Nevins, "Measuring segmental and lexical trends in a corpus of naturalistic speech," in *Proceedings of the 43rd Meeting of the North East Linguistic Society*, H.-L. Huang, E. Poole, and A. Rysling, Eds., vol. 2. GLSA (Graduate Linguistics Student Association), 2014, pp. 153–166.

[6] Z. Bond, *Slips of the Ear: Errors in the Perception of Casual Conversation*. Leiden, The Netherlands: Brill, 1999.

[7] J. Vroomen, M. Van Zon, and B. de Gelder, "Cues to speech segmentation: Evidence from juncture misperceptions and word spotting," *Memory & Cognition*, vol. 24, no. 6, pp. 744–755, 1996.

[8] M. A. Tóth, "A microscopic analysis of consistent word misperceptions," Ph.D. dissertation, Universidad del Paıś Vasco, 2017.

[9] R. Marxer, J. Barker, M. Cooke, and M. L. Garcia Lecumberri, "A corpus of noise-induced word misperceptions for english," *The Journal of the Acoustical Society of America*, vol. 140, no. 5, pp. EL458–EL463, 2016.

[10] G. A. Miller and P. E. Nicely, "An analysis of perceptual confusions among some English consonants," *The Journal of the Acoustical Society of America*, vol. 27, pp. 338–352, 1955.

[11] M. S. Vitevitch, "Naturalistic and experimental analyses of word frequency and neighborhood density effects in slips of the ear," *Language and Speech*, vol. 45, no. 4, pp. 407–434, 2002.

[12] S. A. Phatak and J. B. Allen, "Consonant and vowel confusions in speech-weighted noise," *The Journal of the Acoustical Society of America*, vol. 121, no. 4, pp. 2312–2326, 2007.

[13] S. S. Asaridou and J. M. McQueen, "Speech and music shape the listening brain: evidence for shared domain-general mechanisms," *Frontiers in Psychology*, vol. 4, p. 321, 2013.

[14] C. Beck, B. Kardatzki, and T. Ethofer, "Mondegreens and Soramimi as a Method to Induce Misperceptions of Speech Content â Influence of Familiarity, Wittiness, and Language Competence," *PLoS ONE*, vol. 9, no. 1, 2014.

[15] J. Sundberg, "Formant structure and articulation of spoken and sung vowels," *Folia Phoniatrica et Logopaedica*, vol. 22, no. 1, pp. 28–48, 1970.

[16] R. J. Zatorre and S. R. Baum, "Musical melody and speech intonation: Singing a different tune," *PLoS Biology*, vol. 10, no. 7, p. e1001372, 2012.

[17] P. Eisenberg, "Syllabische Struktur und Wortakzent. Prinzipien der Prosodik deutscher Wörter," *Zeitschrift für Sprachwissenschaft*, vol. 10, no. 1, pp. 37–64, 1991.

[18] G. Kentner, "Rhythmic segmentation in auditory illusions - evidence from cross-linguistic mondegreens," in *Proceedings of 18th ICPhS*. Glasgow: ICPhS, 2015.

[19] A. Hacke and M. Sowa, *Der weisse Neger Wumbaba: kleines Handbuch des Verhörens*, ser. Die Wumbaba-Trilogie. Kunstmann, 2004.

[20] ——, *Der weisse Neger Wumbaba kehrt zurück: zweites Handbuch des Verhörens*, ser. Die Wumbaba-Trilogie. Kunstmann, 2007.

[21] ——, *Wumbabas Vermächtnis: Drittes Handbuch des Verhörens*. Kunstmann, 2012.

[22] H. R. Baayen, R. Piepenbrock, and L. Gulikers, *The CELEX Lexical Database. Release 2 (CD-ROM)*. Philadelphia, Pennsylvania: Linguistic Data Consortium, University of Pennsylvania, 1995.

[23] M. Wieling, J. Prokić, and J. Nerbonne, "Evaluating the pairwise string alignment of pronunciations," in *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education*. Association for Computational Linguistics, 2009, pp. 26–34.

[24] F. Kleber and N. Klipphahn, "An acoustic investigation of secondary stress in German," Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel, Tech. Rep. 37, 2006.

[25] W. F. Sendlmeier and J. Seebode, "Formantkarten des deutschen Vokalsystems," TU Berlin, Institut für Sprache und Kommunikation, Tech. Rep., 2006.

[26] H. Traunmüller, "Analytical expressions for the tonotopic sensory scale," *The Journal of the Acoustical Society of America*, vol. 88, no. 1, pp. 97–100, 1990.

[27] T. Jürgens and T. Brand, "Microscopic prediction of speech recognition for listeners with normal hearing in noise using an auditory model," *The Journal of the Acoustical Society of America*, vol. 126, no. 5, pp. 2635–2648, 2009.

[28] R. N. Shepard, "Toward a universal law of generalization for psychological science," *Science*, vol. 237, no. 4820, pp. 1317–1323, 1987.

[29] N. Mantel, "The detection of disease clustering and a generalized regression approach," *Cancer Research*, vol. 27, no. 2 Part 1, pp. 209–220, 1967.

[30] B. Voss, *Slips of the Ear: Investigations Into the Speech Perception Behaviour of German Speakers of English*, ser. Tübinger Beiträge zur Linguistik. G. Narr, 1984.

[1] We do not approve of the racist word in the titles.