LISTENER-ORIENTED CONSEQUENCES OF PREDICTABILITY-BASED ACOUSTIC ADJUSTMENT

Omnia Ibrahim, Ivan Yuen, Wei Xue, Bistra Andreeva, Bernd Möbius

Department of Language Science and Technology, Saarland University, Germany
omnia@lst.uni-saarland.de

Abstract: This paper investigated whether predictability-based adjustments in production have listener-oriented consequences in perception. By manipulating the acoustic features of a target syllable in different predictability contexts in German, we tested 40 listeners' perceptual preference for the manipulation. Four source words underwent acoustic modifications on the target syllable. Our results revealed a general preference for the original (unmodified) version over the modified one. However, listeners generally favored the unmodified version more when the source word had a higher predictable context compared to a less predictable one. The results showed that predictability-based adjustments have perceptual consequences and that listeners have predictability-based expectations in perception.

1 Introduction

Speakers tend to produce more reduced forms or shorter duration for probable, and thus predictable, messages or conversely lengthen the phonetic form of less predictable ones [1, 2, 3]. Highly predictable syllables are also shortened and have a smaller intensity range than unpredictable ones [4]. This predictability effect has been extensively studied across various languages and linguistic levels [5, 6]. Predictability modifications have been observed at the word [7], morpheme [8], syllable [6], and phoneme levels [9], and their interactions [10]. Several measures to quantify predictability of a message have been proposed [11]. One of them is surprisal. Surprisal captures the intuition that linguistic expressions that are highly predictable in a given context convey less information than those that are unexpected. Surprisal is defined as the contextual predictability of a unit, given its preceding (or following) units, using Equation (1) where S stands for surprisal and P for probability:

$$S(Unit_i) = -log_2 P(Unit_i|Context) \tag{1}$$

The surprisal-based acoustic adjustments observed in speech production are often considered a speaker-oriented strategy to minimize articulatory effort [12]. However, these adjustments also bear perceptual consequences for listeners, as the speaker-listener dyad functions as a cohesive entity. While some studies report the facilitatory effect of surprisal in perception [13], others find contrasting results, especially in noisy listening conditions [14]. Despite these discrepancies, the role of surprisal in perception remains significant [15, 16].

Given the surprisal-based modifications of acoustic features in production in [4], it follows to raise questions about their perceptual consequences on listeners' phonetic expectations. The present study investigated whether the surprisal-based adjustment in production has listener-oriented consequences in perception. By manipulating the acoustic features of a target consonant–vowel (CV) syllable in two different surprisal contexts in German, we tested listeners' perceptual preference for the manipulation. We hypothesized that listeners expect and prefer acoustic features of a target CV syllable that are congruent with its surprisal context.

 DOI: 10.35096/othr/pub-7098, URN: urn:nbn:de:bvb:898-opus4-70984

2 Method

2.1 Material / stimuli

We used four polysyllabic German words as source words (Table 1), a subset of the stimuli used in [4]. The selected subset has a modal voice quality. A word-medial CV syllable in these words differed in terms of surprisal, which is estimated as the negative log probability of the target syllable, given two preceding syllables. This resulted in source words containing a low-surprisal (LS) or high-surprisal (HS) target syllable. Each source word was used to construct a pair of unmodified vs. modified versions. In the modified version, the target syllable was acoustically manipulated with respect to fundamental frequency, duration, and intensity range, guided by the production patterns reported in [4]. For instance, the low-surprisal target syllable /bo:/ in "Einfuhrver*bo*te" was manipulated to resemble its corresponding high-surprisal counterpart. Similarly, a high-surprisal target syllable was modified to resemble its low-surprisal counterpart. Therefore, acoustic modification resulted in incongruence of expected surprisal. We used Audacity to adjust the intensity range and Praat to modify F0 and duration.

Following the modification, a naturalness rating experiment was conducted to ensure the absence of any audible artifacts that could influence listeners' preferences in the test. Participants were instructed to rate the naturalness of the signal on a scale ranging from 0 (unnatural) to 5 (natural). All stimuli received consistently high naturalness ratings from a group of 20 German listeners. Subsequently, we aimed to determine a signal-to-noise ratio (SNR) that would create a challenging listening condition for participants, preventing ceiling performance while avoiding excessive difficulty. To achieve this, we conducted an experiment to identify the optimal SNR fitting this criterion. Stimuli were presented at three different SNR levels: 5 dB, 0 dB, and -5 dB besides quiet listening condition. The experiment indicated that -5 dB constituted a challenging listening scenario. In summary, four source words (2 LS, 2 HS) underwent acoustic modifications on the target syllable, and we chose a -5 dB SNR for the noise listening condition in the final experiment.

Table 1 – The source words used as a base for the acoustic modification.

Source word	Target syllable	Surprisal type	Modification direction
Einfuhrver*bo*te	[bo:]	low surprisal	enhancement
Sprachmo*du*le	[du:]	low surprisal	enhancement
Berufs*bo*nus	[bo:]	high surprisal	reduction
be*du*delt	[du:]	high surprisal	reduction

2.2 Participants

We recruited 40 German listeners (age range, 20–39 years; mean age 30 years, s.d. = 5.3; 20 females), registered as "participants" on the Prolific website. Participants had not seen the stimuli before. None of the participants reported any hearing problems. They received monetary compensation for taking part in the perception study. Participants were naive with regard to the experimental purpose.

2.3 Experimental procedures

A listening preference task was carried out online on Prolific. The perception experiment was designed and constructed using Labvanced. It contained two listening blocks representing the

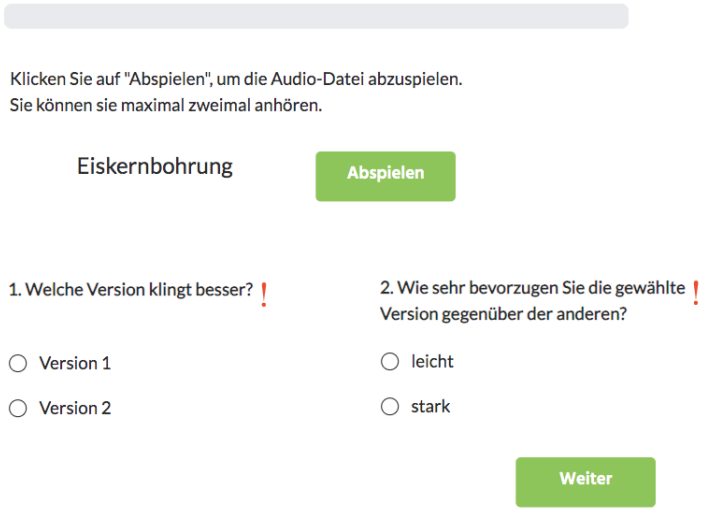

Figure 1 – The main interface for the perceptual experiment displaying a listening preference task. Participants were presented with two questions: 1) 'Which version is better?' and 2) 'To what degree?'

two listening conditions: Quiet and Noise. Each block consisted of 8 trials. Each trial contained two acoustic versions of a polysyllabic word: the unmodified (original) version and its modified counterpart. To eliminate order effects, the presentation order of the original and modified versions was counterbalanced for every word pair. This counterbalancing involved presenting the original version first in some trials and the modified version first in others, ensuring an unbiased participant preference. Participants were instructed to indicate their listening preference for one of the two versions. The task involved responding to two key questions: 1) "Which version is better?" and 2) "To what degree?" (Figure 1). Following the listening preference task, participants completed a questionnaire concerning their language background and musical training. This additional information aimed to gather insights into potential factors influencing participants' preferences.

2.4 Data analysis

The data were analyzed using a mixed-effects logistic regression model to investigate the relationship between the dependent variable, "preference", and several independent variables. The model was specified as follows: glmer(preference ~ Listening condition + Surprisal type+ Order + (1 | Subject ID) + (1 | Word) , family = binomial).

The fixed factors were 2 two-level factors: (1) listening condition (Quiet or Noise), and (2) Surprisal type (LS or HS), representing the primary factors of interest. We added Order as a covariate. Additionally, random intercepts were incorporated for both Subject ID and Word to account for potential variability at the individual subject and word levels. The logistic regression model, implemented using the glmer function from the lme4 package in R [17], assumes a binomial family to accommodate the binary nature of the dependent variable.

3 Results

Figure 2 illustrates how listener preferences vary across different surprisal types in both quiet and noisy listening conditions. The x-axis represents these surprisal types (low surpisal and high surprisal), while the y-axis denotes the corresponding listener preference for the unmodified (original) version. In the analysis of listener preferences, several factors emerged as significant

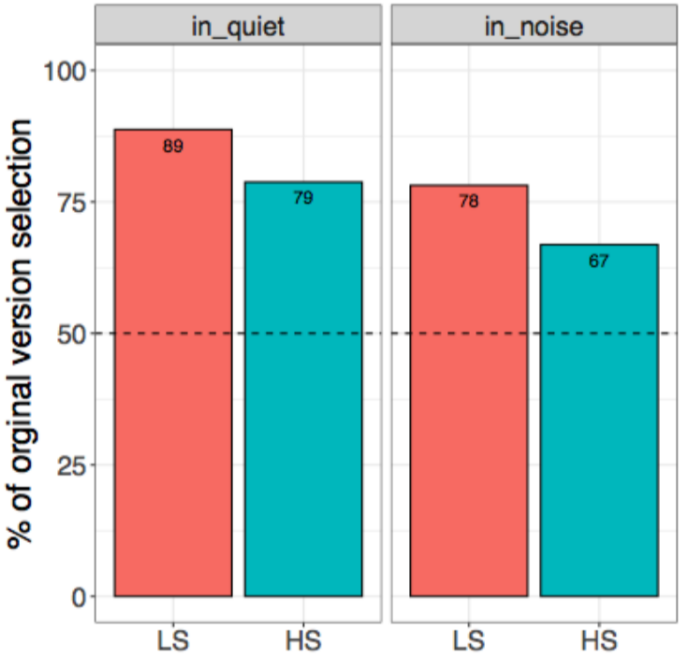

Figure 2 – The percentage of choosing the original version as a function of listening conditions and the surprisal type (low surprisal and high surprisal)

Table 2 – Glmer model results for the fixed effects. The reference level for Listening-condition is Quiet, for Surprisal type is LS, and for Order it is O1 (original-modified).

	Estimate	**Std. error**	***t***	**Pr (>|t|)**
Intercept	2,4243	0,2921	8,301	<2e-16 ***
Listening-condition in-noise	-0,7437	0,2058	-3,614	**0.000301*** **
Surprisal type HS	-0.7066	0.2578	-2.741	**0.00613** **
Order O2	-0.4141	0.2029	-2.041	**0.041243***

contributors. Firstly, the presence of noise significantly impacted preferences (Table 2, p = .0003***), revealing that listener preference for the unmodified (original) version tended to be lower in the noisy condition compared to the quiet condition. As stated before, the noise listening condition was introduced to prevent ceiling performance. In a noisy setting, listeners may encounter challenges in accurately processing and identifying acoustic features, leading to a reduced preference for the unmodified version. This change in preference indicates a listener adaptation to the acoustic challenges posed by the noisy condition. Turning to the surprisal type, it also significantly influences the listeners' preference (Table 2, p = .006**). Preference scores indicated that listeners generally favored the unmodified version more when the source words had low surprisal compared to high surprisal. Furthermore, we introduced the order of stimuli within each trial as a covariate, revealing a statistically significant effect in our experiment. Listeners' preferences were lower when the original words appeared in the second position (modified - original) (Table 2, p = .041243*). This order effect implies that the first member serves as an auditory referent.

4 Discussion and conclusion

Effective communication often involves finding a balance between speaker-oriented and listener-oriented approaches. The current paper followed up on an observation in Ibrahim et al. [4] that syllable-based surprisal has a significant effect on syllable duration. High-surprisal syllables have longer duration than low-surprisal ones. Such surprisal-based acoustic adjustments are often interpreted as a speaker-oriented strategy to minimize articulatory effort. In the current study, we explored the perceptual consequence of this surprisal effect on listeners' phonetic expectations. Our investigation extended beyond the speaker-oriented perspective to assess whether predictability also holds a listener-oriented dimension. Our study involved the manipulation of acoustic features of a target CV syllable in two surprisal contexts in German.

In general, our findings indicate a preference among listeners for the unmodified (original) version over the modified counterpart. The listener's preference for the unmodified version reflects an expectation or preference for the familiar, original acoustic features. Furthermore, listeners generally favored the unmodified version more when the source words had low surprisal compared to high surprisal. This shift in preference as a function of surprisal type holds in both quiet and noise conditions (Figure 2). However, the tendency to prefer the original version is lower in a high-surprisal than a low-surprisal syllable, suggesting that the listeners prefer the incongruent version of the high-surprisal syllables more than the incongruent version of the low-surprisal syllables. This arose perhaps because the direction of the acoustic modification on a source high-surprisal syllable is to reduce acoustic features, whereas that on a source low-surprisal syllable is to exaggerate acoustic features. This preference pattern could then suggest a listener bias. The listeners tend to be more tolerant of acoustic reduction for the target syllables with high suprisal, but less so of acoustic enhancement for those with low suprisal. The preference for the reduction variant therefore suggests that the effect of surprisal in production might be driven by reduction, emphasizing the reduced articulatory effort view, rather than an enhancement view that could be more listener-oriented.

In conclusion, our results highlight the perceptual consequences of surprisal-based adjustments, indicating that listeners indeed possess surprisal-based expectations in perception. Despite an apparent bias towards reduced acoustic features, these findings contribute to our understanding of the intricate role of surprisal in shaping listener-oriented outcomes in speech perception. Combining the current results with the results from the production experiment in [4], we can argue that the effect of predictability is not solely speaker oriented behavior but rather a balance between both speaker and listener oriented behavior.

5 Acknowledgements

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102. We thank Marjolein van Os for her support.

References

[1] AYLETT, M. and A. TURK: *The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech*. *Language and Speech*, 47(1), pp. 31–56, 2004. doi:10.1177/00238309040470010201.

[2] FRANK, A. and T. F. JAEGER: *Speaking rationally: Uniform information density as an optimal strategy for language production*. In *Proceedings of the Annual Meeting*

of the Cognitive Science Society. 2008. URL https://escholarship.org/uc/item/7d08h6j4.

[3] CROCKER, M. W., V. DEMBERG, and E. TEICH: *Information Density and Linguistic Encoding (IDeaL). KI - Künstliche Intelligenz*, 30(1), pp. 77–81, 2016.

[4] IBRAHIM, O., I. YUEN, M. VAN OS, B. ANDREEVA, and B. MÖBIUS: *The combined effects of contextual predictability and noise on the acoustic realisation of german syllables. The Journal of the Acoustical Society of America*, 152(2), pp. 911–920, 2022. doi:10.1121/10.0013413. URL https://pubs.aip.org/jasa/article/152/2/911/2839439/The-combined-effects-of-contextual-predictability.

[5] PIMENTEL, T., C. MEISTER, E. SALESKY, S. TEUFEL, D. BLASI, and R. COTTERELL: *A surprisal–duration trade-off across and within the world's languages*. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 949–962. Association for Computational Linguistics, 2021.

[6] AYLETT, M. and A. TURK: *Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. Journal of the Acoustical Society of America*, 119(5), pp. 3048–3058, 2006. doi:10.1121/1.2188331. URL https://doi.org/10.1121/1.2188331. https://doi.org/10.1121/1.2188331.

[7] BUZ, E. and T. F. JAEGER: *The (in)dependence of articulation and lexical planning during isolated word production. Language, Cognition and Neuroscience*, 31(3), pp. 404–424, 2016. doi:10.1080/23273798.2015.1105984.

[8] TANG, K. and R. BENNETT: *Contextual predictability influences word and morpheme duration in a morphologically complex language (kaqchikel mayan). Journal of the Acoustical Society of America*, 144(2), pp. 997–1017, 2018. doi:10.1121/1.5046095.

[9] BYBEE, J.: *Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. Language Variation and Change*, 14(3), p. 261–290, 2002. doi:10.1017/S0954394502143018.

[10] HASEGAWA-JOHNSON, M., J. COLE, K. CHEN, L. PARTHA, A. JUNEJA, T. YOON, S. BORYS, and X. ZHUANG: *Prosodic hierarchy as an organizing framework for the sources of context in phone-based and articulatory-feature-based speech recognition*. In S. TSENG (ed.), *Linguistic Patterns of Spontaneous Speech*, pp. 101–128. Academica Sinica, 2009.

[11] HALE, J.: *Information-theoretical complexity metrics. Language and Linguistics Compass*, pp. 1–16, 2016. doi:10.1111/lnc4.12196.

[12] LINDBLOM, B.: *Explaining phonetic variation: A sketch of the h&h theory*. In *Speech production and speech modelling*, pp. 403–439. Springer, 1990.

[13] VAN OS, M., J. KRAY, and V. DEMBERG: *Rational speech comprehension: Interaction between predictability, acoustic signal, and noise. Frontiers in Psychology*, 13, p. 914239, 2022. doi:10.3389/fpsyg.2022.914239. URL https://www.frontiersin.org/articles/10.3389/fpsyg.2022.914239/full.

[14] MARRUFO-PÉREZ, M. I., A. EUSTAQUIO-MARTÍN, and E. A. LOPEZ-POVEDA: *Speech predictability can hinder communication in difficult listening conditions. Cognition*, 192,

p. 103992, 2019. doi:10.1016/j.cognition.2019.06.004. URL https://linkinghub.elsevier.com/retrieve/pii/S0010027719301659.

[15] DUBNO, J. R., J. B. AHLSTROM, and A. R. HORWITZ: *Use of context by young and aged adults with normal hearing*. *The Journal of the Acoustical Society of America*, 107(1), pp. 538–546, 2000.

[16] STAUB, A., M. GRANT, L. ASTHEIMER, and A. COHEN: *The influence of cloze probability and item constraint on cloze task response time*. *Journal of Memory and Language*, 82, pp. 1–17, 2015.

[17] BATES, D., M. MÄCHLER, B. BOLKER, and S. WALKER: *Fitting linear mixed-effects models using lme4*. *Journal of Statistical Software*, 67(1), pp. 1–48, 2015. doi:10.18637/jss.v067.i01.