The combined effects of contextual predictability and noise on the acoustic realisation of German syllables

Omnia Ibrahim, Ivan Yuen, Marjolein van Os, et al.

ARTICLES YOU MAY BE INTERESTED IN

JASA ARTICLE

CrossMark
click for updates

The combined effects of contextual predictability and noise on the acoustic realisation of German syllables

Omnia Ibrahim,[a] (iD) Ivan Yuen, (iD) Marjolein van Os, Bistra Andreeva, (iD) and Bernd Möbius (iD)

Department of Language Science and Technology, Saarland University, Saarbrücken, 66123, Germany

ABSTRACT:

Speakers tend to speak clearly in noisy environments, while they tend to reserve effort by shortening word duration in predictable contexts. It is unclear how these two communicative demands are met. The current study investigates the acoustic realizations of syllables in predictable vs unpredictable contexts across different background noise levels. Thirty-eight German native speakers produced 60 CV syllables in two predictability contexts in three noise conditions (reference = quiet, 0 dB and −10 dB signal-to-noise ratio). Duration, intensity (average and range), $F0$ (median), and vowel formants of the target syllables were analysed. The presence of noise yielded significantly longer duration, higher average intensity, larger intensity range, and higher $F0$. Noise levels affected intensity (average and range) and $F0$. Low predictability syllables exhibited longer duration and larger intensity range. However, no interaction was found between noise and predictability. This suggests that noise-related modifications might be independent of predictability-related changes, with implications for including channel-based and message-based formulations in speech production. © 2022 Author(s). *All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).*
https://doi.org/10.1121/10.0013413

I. INTRODUCTION

Linguistic units, e.g., words, syllables or phonemes, occurring in predictable contexts, tend to undergo more reduction than those in less predictable contexts (Aylett and Turk, 2004; Crocker *et al.*, 2016; Frank and Jaeger, 2008) (also known as source coding; see Pate and Goldwater, 2015) but this tendency might be counteracted in noisy environments when it is difficult for listeners to anticipate a message from the degraded signals. In response, speakers often enhance the signals. However, this need contradicts the tendency for speakers to reduce in predictable contexts. This then raises the question as to whether predictability-based modifications will be attenuated or exaggerated in noisy environments. The current study investigates the acoustic realisations of syllables in predictable vs unpredictable contexts across different background noise levels.

Speakers often adapt their speech to different environments, e.g., in a factory or a stadium (see Cooke *et al.*, 2014, for a review). It is generally accepted that speech will be automatically made louder in noisy environments, with the resultant noise-induced speaking style referred to as Lombard speech (Brumm and Zollinger, 2011; Lombard, 1911). This style is considered an adaptation to protect against potential information loss in a non-ideal channel during signal transmission (also known as channel coding; see Pate and Goldwater, 2015). A number of acoustic features have been identified that characterize this speaking style: increased vocal effort, increased intensity, increased fundamental frequency ($F0$), e.g., Patel and Schell (2008), slow speaking rate, long duration, e.g., Lu (2010), and high first formant ($F1$) of vowels (see Junqua, 1996, for a review). These characteristics have been reported to help listeners identify speech in challenging communicative settings, suggesting acoustic enhancement of signals (Hazan and Simpson, 2000).

Lombard speech modification does not manifest uniformly across different segment types or linguistic units. Rather, Lombard speech shows up in units that are critical to intelligibility. For instance, speakers emphasize vowels more than consonants in a noisy environment, presumably the former is more critical to speech audibility (Garnier and Henrich, 2014). Content words undergo larger modifications than function words in terms of fundamental frequency, syllable duration (Patel and Schell, 2008), and lexical stress contrastivity (Arciuli *et al.*, 2014). Moreover, speakers differentially increase vocal effort with levels of noise intensity (Ngo *et al.*, 2017; Wakao *et al.*, 1996). The study by Zhao and Jurafsky (2009) examined the effects of word frequency and noise on the acoustic realisation of Cantonese tones. Their results showed an overall increase in $F0$ for all tones in noise, but they only observed the effect of word frequency for mid tones, with higher $F0$ in low-frequency than high-frequency words. This is because mid-tones are hard to distinguish perceptually and it will be more difficult to identify an unpredictable (i.e., low frequency) word with mid tones without making it acoustically salient. These previous studies then suggest that speech is modified to serve some communication goals for the benefit of listeners.

[a]Electronic mail: omnia@lst.uni-saarland.de

Word frequency is one measure for predictability (Ernestus, 2014). Yet predictability goes beyond word frequency, e.g., local contextual predictability. Local contextual predictability is the predictability of a unit, given its preceding (or following) units, referred to as surprisal (Shannon, 1948). Surprisal quantifies information in terms of bits as the inverse of the unit's log probability given the local context: $S(unit_i) = -\log_2 P(unit_i|Context)$ (Hale, 2016).

Predictability has been shown to modify speech output at multiple levels. For instance, speakers tend to shorten word duration for predictable, but not for less predictable, messages (Buz and Jaeger, 2016). Such predictability effect is often interpreted under the notion of achieving a speaking intent with minimal effort. Conversely, hard-to-understand units tend to have longer duration, possibly resulting from explicit encoding to improve intelligibility (Gahl *et al.*, 2012; Jaeger, 2010). In addition, American English vowels were found to be more centralized in contextually more predictable syllables (Aylett and Turk, 2006), in line with the idea that vowels are more likely to be weakened or deleted in high-frequency than low-frequency words (Bybee, 2001, 2002). In general, vowels were more dispersed in the vowel space when they were contextually less predictable as in Malisz *et al.* (2018) and this effect was observed in six different languages (American English, Czech, Finnish, French, German, and Polish) across three speaking rates (slow, normal, and fast).

From the information theory perspective (Shannon, 1948), message formulation (source coding) aims at representing information as accurately as possible in as few bits as possible, while channel coding aims at protecting information from transmission loss over a non-ideal channel. To the best of our knowledge, no study has examined the effects of channel coding (e.g., Lombard speech) and source coding (e.g., local contextual predictability) on the acoustic realisation of a syllable. The primary goal of this study was to examine any combined effects of background noise and contextual predictability in German syllables. We extracted duration, average intensity, intensity range, and $F0$ features from the syllable, and $F1$, $F2$, and $F2$–$F1$ difference from the vowel. Throughout this paper, we refer to features extracted from the syllable as syllable-based features and vowel formants as vowel-based features. Those features were chosen as the metrics for speech modification because they are known to correlate with the Lombard effect and some of the metrics have been shown to be sensitive to contextual predictability (e.g., Boril and Pollák, 2005; Brandt, 2019; Lu and Cooke, 2009).

Our research question was how syllables with different contextual predictability are acoustically realised as a function of noise. We expected syllables with high surprisal or in noise to be articulated with care when compared to their counterparts with low surprisal or without noise, respectively. In addition, if speech is modified to serve some communicative goals for the listener, it is more likely for modification to be made on the least predicted element. Therefore, we also expected acoustic modifications to be larger for syllables with high surprisal than those with low surprisal in noisy conditions.

II. METHOD

A. Participants

Thirty-eight native German speakers with no known hearing or speech impairments were recruited (12 M, 26 °F; average age = 27.6, 19–60 years) to take part in a reading aloud task. For recruitment, no upper age limit was imposed, in consideration of the observation that older adults with age-typical mild hearing loss do not differ from young adults in how they modify their speech in background noise (Hazan *et al.*, 2018).

B. Stimuli

Twenty stressed CV syllables were created by combining one of the plosives: /p, b, d, k/ and one of the vowels: /aː, eː, iː, oː, uː/. The CV syllables, consisting of two surprisal groups (high vs low), were crossed with three white-noise conditions (reference = no noise, 0 dB and −10 dB signal-to-noise ratio, SNR), resulting in a total of 60 target syllables. Each target syllable formed part of a polysyllabic word in a sentence context, which was chosen from the DeWaC corpus (Baroni and Kilgarriff, 2006; Brandt *et al.*, 2017). The stimuli were pseudo-randomized for presentation.

A syllable-based language model was used to estimate the probability of a unit given its previous local context, based on large text corpora. The choice to train the language model (LM) with syllable as a unit was motivated by the importance of the syllable as a processing unit in language production and perception (see Krakow, 1999, for a review), and as demonstrated in different phonetic encoding for high frequency vs novel syllables in Bürki *et al.* (2015). Although previous studies, e.g., Bell *et al.* (2009) found an effect of backward predictability on the duration of lexical and function words in American English, a more recent study by Tang and Bennett (2018) found a stronger effect of forward predictability on word duration in Kaqchikel Mayan. Considering the inconsistent findings on backward predictability, we therefore opted to focus on estimating forward predictability, which is consistent with the assumption of linearity in formulating speech (Levelt *et al.*, 1999). Surprisal, defined as: $S(syllable_i) = -\log_2 P(syllable_i|syllable_{i-1}, syllable_{i-2})$, was computed for the target syllables. The surprisal values were derived from a trained language model of the German web corpus (DeWaC) consisting of more than 1.34×10^9 words from written texts in different genres (e.g., newspapers, chats). The choice of using a written corpus was motivated by the limitation of small-sized spoken corpora in representing rarer/infrequent syllable types (Möbius, 2003; Schweitzer and Möbius, 2004). Besides, there is evidence suggesting that written corpora provide comparable estimates of syllable frequency to spoken corpora for German (Samlowski *et al.*, 2011). Syllable boundary assignment was fine-tuned, using the Hidden Markov Model (HMM) syllable tagger (Schmid *et al.*, 2007). A syllable-based forward tri-gram language model was then generated, using the SRI Language Modeling

912   J. Acoust. Soc. Am. **152** (2), August 2022

Ibrahim *et al.*

Toolkit (SRILM) (Stolcke, 2002). According to the language model, target syllables were assigned with high and low surprisal values. A high surprisal syllable is less predictable (S ≈ 5.05), whereas a low surprisal syllable is more predictable (S ≈ 0.58). In total, we had 60 sentences, to be recorded in three noise conditions (reference = no noise, 0 dB and −10 dB SNR).

C. Procedure

Participants were instructed to read aloud a set of stimuli at their habitual pace in a soundproof studio. They wore over-ear headphones (AKGK271 MKII, AKG Harman) and a head-mounted microphone (DPA 4067-F Omni, DPA Microphones), with a display computer in front. Sentence stimuli were orthographically presented in two lines in the centre of the computer screen. Note that some sentence stimuli were shortened to fit within the 2-line limit. Participants were informed about the presence of noise during reading. Noise was played through the headphones. Eleven practice sentences were provided for participants to become familiar with the procedure and for the research assistant to calibrate the equipment in the control room. The experiment consisted of three noise conditions in separate blocks, with the middle block reserved for the reference (no noise) condition. The order of the other two noise conditions (0 dB vs –10 dB) was counterbalanced across participants. This design was adopted in order to minimize the confounding linear correlation between speech modification and the presentation order of noise (from quiet to soft to loud) as observed in our pilot experiment using a subset of 30 DeWaC sentences. The experiment took about 30 min to complete. Productions were recorded and stored as a mono.wav file with a sampling rate of 48 kHz and 24 bits per sample.

D. Data annotation

A total of 1906 sentence stimuli were phonemically and orthographically transcribed, after removing 363 items for mispronounciation ($N = 152$) or disfluency due to the presence of pause/hesitation ($N = 211$). Words, syllables, and segments in each sentence were first automatically annotated using Web-MAUS (Schiel, 1999). Two trained phoneticians subsequently checked all automatic annotations, and manually adjusted the boundaries of the target CV syllables and their constituent segments using Praat (version 6.1.08).

E. Statistical analysis

Four syllable-level acoustic features (duration, average intensity, intensity range, and $F0$), and three vowel-level acoustic features ($F1$, $F2$, and $F2$–$F1$) were chosen as dependent variables, in accordance with previous research (Boril and Pollák, 2005; Brandt, 2019; Brandt et al., 2019; Castellanos et al., 1996; Lu and Cooke, 2009). $F0$ and formants were measured at the mid-point of each vowel because visual inspection of the data revealed minimal trajectory differences across experimental conditions. Formant values

were determined using the default setting of Formant(Burg) in Praat, adjusting the maximum formant value for gender (5 KHz for male and 5.5 KHz for female). All feature values were extracted using in-house Python and Praat scripts. Since the normality assumption was violated in a Shapiro-Wilk test (Shapiro and Wilk, 1965), syllable duration, intensity and F0 were transformed into z-scores per participant (Simpson, 2009; Traunmüller and Eriksson, 1994), and formants were also normalized per participant using the Lobanov method (Adank et al., 2004).

Each normalized acoustic feature was statistically analyzed by fitting linear mixed effects models (LME) using the lmer package (Bates et al., 2015) in R (R Core Team, 2018) and evaluating model fits using the lmerTest package (Kuznetsova et al., 2017). Fixed effects included surprisal group, noise conditions, and their interactions. Surprisal group was coded as a simple contrast, and noise conditions were Helmert contrast coded to create two comparisons: (1) absence vs presence of noise (i.e., noise condition 1), (2) 0 dB vs −10 dB SNR noise levels (i.e., noise condition 2). Covariates included target syllables, part of speech (PoS), and lexical frequency for word stimuli. Sentence stimuli were treated as a random effect. We first constructed a by-sentence random intercept model for each acoustic feature. This model was evaluated against models with intercept + slope for fixed effects to identify the optimal random structure, on the basis of Akaike information criterion (AIC). Each covariate was then evaluated for inclusion. Significance of effects in each model was evaluated by performing maximum likelihood t-tests using Satterthwaite approximations to degrees of freedom. Alpha was Bonferroni-adjusted for multiple pairwise comparisons.

III. RESULTS

A. Syllable-level acoustic features

Figure 1 shows the mean z-scores for (a) syllable duration, (b) average intensity, (c) intensity range, and (d) $F0$ in absence vs presence of noise conditions. As expected, syllable duration was longer (Estimate = -8.093×10^{-2}, $t = -3.25$, $p = 0.001$**), average intensity higher (Estimate = -8.043×10^{-1}, $t = -25.96$, $p \leq 0.0001$***), intensity range larger (Estimate = -1.294×10^{-1}, $t = -3.79$, $p = 0.00015$***), and overall $F0$ higher (Estimate = -2.319×10^{-1}, $t = -6.78$, $p \leq 0.0001$***), when noise was present. These noise-induced effects all reached statistical significance, consistent with the predicted Lombard effect. Figure 2 shows the z-scores for the same acoustic features at 0 vs −10 dB SNR noise levels. Higher noise levels induced higher average intensity (Estimate = -3.476×10^{-1}, $t = -9.82$, $p \leq 0.0001$***), larger intensity range (Estimate = -8.107×10^{-2}, $t = -2.08$, $p = 0.037$*) and higher $F0$ (Estimate = -1.547×10^{-1}, $t = -3.96$, $p \leq 0.0001$***), with these effects reaching statistical significance. Unexpectedly, noise levels did not affect syllable duration. In addition, a significant effect of surprisal was observed for syllable duration (Estimate = -2.620×10^{-1},

J. Acoust. Soc. Am. **152** (2), August 2022

Ibrahim et al.   913

FIG. 1. Mean z-scores for syllable-level (A) duration, (B) average intensity, (C) intensity range, and (D) F0 as a function of absence vs presence of noise (with ± SE).

$t = -2.06$, $p = 0.045$*) and intensity range (Estimate $= -3.874 \times 10^{-1}$, $t = -2.32$, $p = 0.025$*), with longer duration and larger intensity range for high surprisal syllables. While the effect of surprisal on syllable duration is replicated, its effect on intensity range has not been previously reported. When the effects of noise (i.e., either in terms of presence or levels) and surprisal were present, we did not observe any significant interaction. This is counter to the predicted interaction that noisy signals will be phonetically made more salient (i.e., phonetically enhanced) for high surprisal syllables than their low surprisal counterparts.
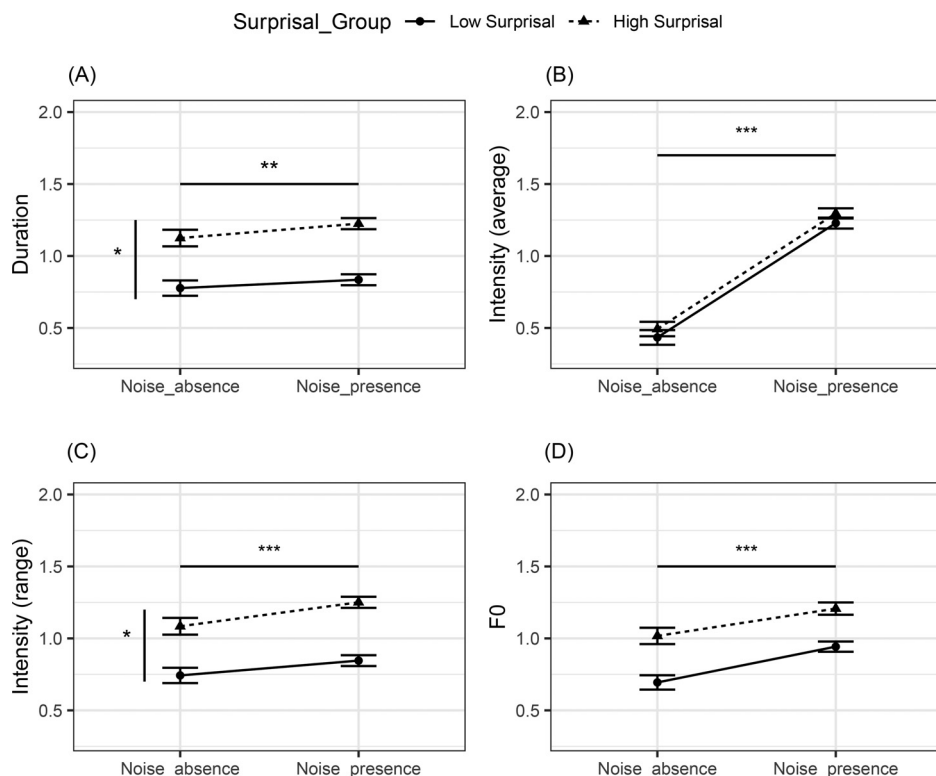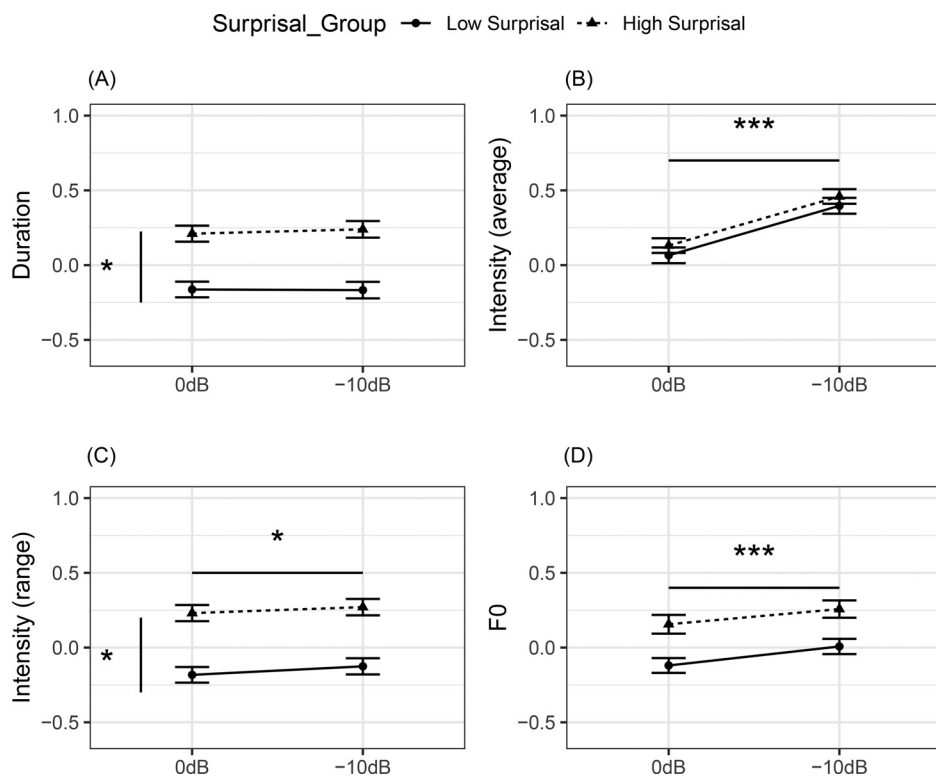

FIG. 2. Mean z-scores for syllable-level (A) duration, (B) average intensity, (C) intensity range, and (D) F0 as a function of noise levels: 0 vs −10 dB (with ± SE).
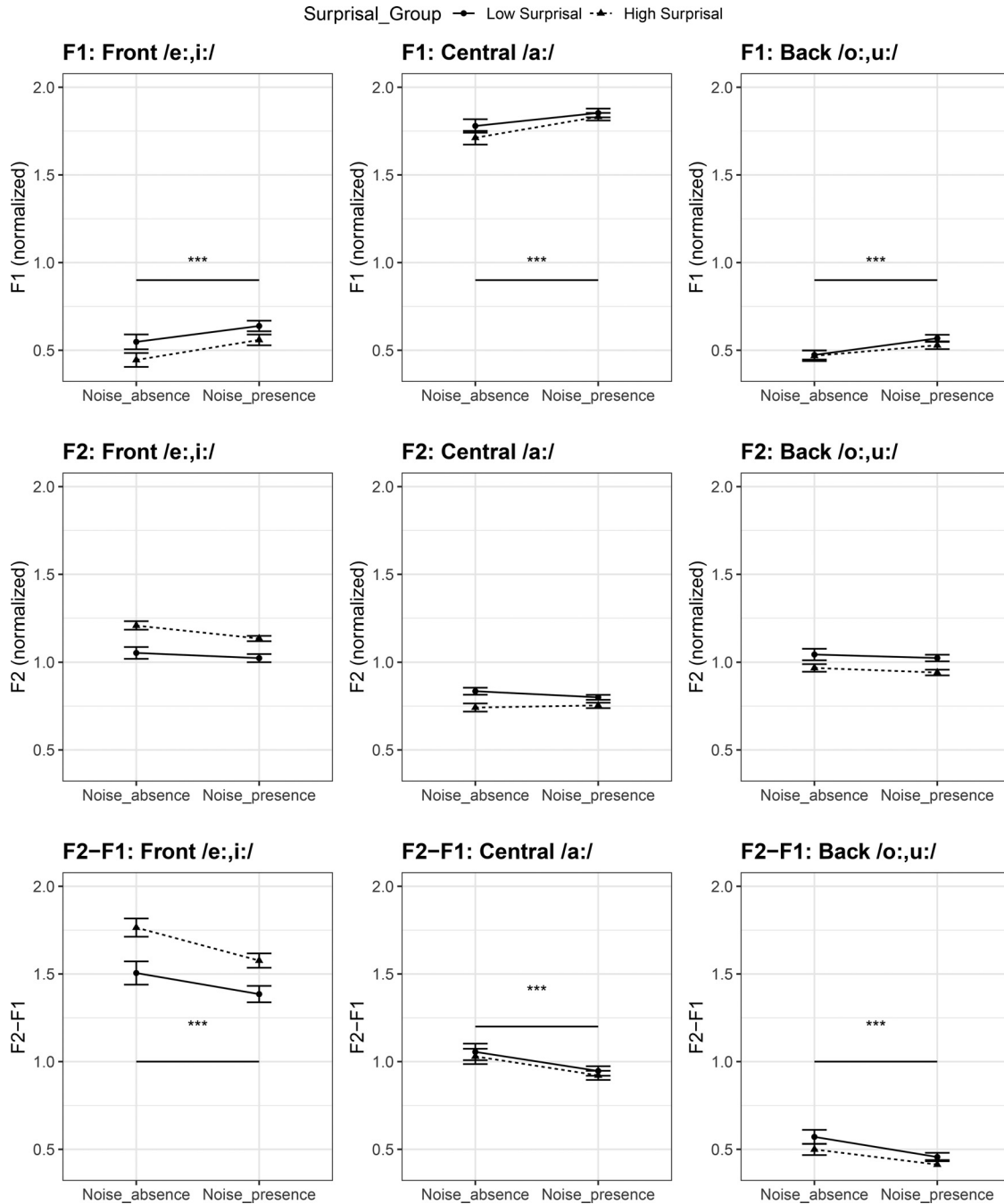
Ibrahim *et al.*

FIG. 3. Mean z-scores of F1, F2 measured at the mid-point of the vowel, and F2 -F1 difference for front /i:, e:/, central /a:/ and back /o:, u:/ vowels in the absence vs presence of noise (with ± SE).

B. Vowel-level acoustic features

We predicted that vowels will be produced with effort in high surprisal syllables or noisy conditions (i.e., with less reduction). Under those scenarios, front vowels will be more peripheral by lowering $F1$ and raising $F2$ with a larger $F2–F1$ difference. The central vowel will be more open by raising $F1$ with a smaller $F2–F1$ difference. Back vowels will be more peripheral by lowering $F1$ and $F2$. In our analysis, vowels were separated into "front" (/i:/ and /e:/), "central" (/a:/), and "back" (/o:/ and /u:/). Figure 3 shows the z-scores of (a) $F1$, (b) $F2$, (c) $F2–F1$ as a function of the absence vs presence of noise in three different vowel sub-groups. Counter to our

predictions, front vowels exhibited higher F1 (Estimate $= -0.124$, $t = -5.92$, $p \leq 0.0001$***) and lower $F2$ (Estimate $= 0.059$, $t = 2.81$, $p = 0.005$**) with a smaller $F2–F1$ (Estimate $= 0.184$, $t = 5.48$, $p \leq 0.0001$***) in the presence of noise than in its absence, with these effects reaching statistical significance. As for the central vowel, $F1$ was significantly higher (Estimate $= -0.103$, $t = -3.45$, $p = 0.0006$***) with a significantly smaller $F2–F1$ difference (Estimate $= 0.117$, $t = 3.48$, $p = 0.0005$***) in the presence of noise than in its absence as predicted. Back vowels significantly raised $F1$ (Estimate $= -0.07$, $t = -3.68$, $p = 0.0002$***) with no difference in $F2$, resulting in a significantly smaller $F2–F1$

J. Acoust. Soc. Am. **152** (2), August 2022

Ibrahim *et al.* 915

difference (Estimate $= 0.093$, $t = 3.81$, $p = 0.0001***$) in the presence of noise, contrary to our prediction. No additional effects of noise levels, surprisal or any interactions were observed.

These results suggest that formant modification is primarily attributable to the presence of noise and that formants are not modified to be peripheral. To check if the formant change is a side-effect of higher $F0$ or intensity as an adjustment to noise, we conducted a series of correlations. The results revealed significant positive correlations between average intensity and $F1$ for front ($r = 0.34**$), central ($r = 0.13*$), and back vowels ($r = 0.40***$), but only a significant positive correlation between F0 and F1 for back vowels ($r = 0.10**$).[1] These patterns suggest that the overall $F1$ raising for the three vowel types might be due to increasing intensity.

IV. DISCUSSION

The present study has extended previous research by examining the combined effects of noise and local contextual predictability (i.e., surprisal) on the acoustic realisations of syllables in German. We hypothesized that syllables with high surprisal or in noise will be hyperarticulated when compared to their counterparts with low surprisal or without noise, and that this effect will be stronger when the noise level increases. In addition, we expected the magnitude of hyperarticulation to be larger for syllables with high surprisal than those with low surprisal in noisy conditions, if speakers choose to make the least predicted element intelligible.

Our results provided evidence for the expected effect of noise on hyperarticulation: longer syllable duration, higher average intensity, larger intensity range, and higher $F0$, higher $F1$, and smaller $F2$–$F1$ difference in the presence of noise than in its absence. In addition, noise levels increased average intensity, intensity range, and $F0$. Our results also revealed an effect of surprisal, resulting in longer syllable duration and larger intensity range for syllables with high surprisal. Contrary to our expectation, the effect of surprisal was not more pronounced in the presence of noise or at variable noise levels (as reflected in the lack of any interaction effects).

A. Effects of noise

The presence of noise affected both syllable-level features (syllable duration, average intensity, intensity range, and $F0$) and vowel-level features ($F1$, $F2$, and $F2$–$F1$ difference), in similar directions as observed in previous studies (Bapineedu *et al.*, 2009; Davis *et al.*, 2006; Dreher and O'Neill, 1957; Fricke, 1970; Garnier *et al.*, 2010; Godoy *et al.*, 2014; Junqua, 1996; Lu and Cooke, 2008; Meekings *et al.*, 2016; Ngo *et al.*, 2017; Patel and Schell, 2008; Pisoni *et al.*, 1985; Pittman and Wiley, 2001; Summers *et al.*, 1988). This pervasive effect suggests speakers' attempt to make their voice (i.e., the carrier) and consequently the message (i.e., the content) more salient and detectable in a noisy

environment along multiple acoustic dimensions (at the expense of being redundant).

However, not all the measured acoustic features were manifested for the sake of increasing detectability or enhancing phonological contrasts. Although the syllable-level features were acoustically enhanced to become more detectable in noise, vowel formants did not follow the same pattern. Previous studies have shown that vowel peripherality benefits intelligibility (Ferguson and Kewley-Port, 2007; Smiljanić and Bradlow, 2005). Therefore, we would expect lower $F1$ and higher $F2$ for closed front vowels, or higher $F1$ for open central vowels, or lower $F1$ and $F2$ for closed back vowels in a noisy condition.

Our results did not yield those expected patterns for the three vowel groups. On the contrary, the $F1$ of the front, central, and back vowels under investigation increased uniformly in a noisy condition. The $F1$ increase might be related to the corresponding increase in average $F0$, which could arise from tense vocal folds because previous studies have reported a moderate correlation between $F0$ and formants that stems from co-variation in the size of laryngeal and supra-laryngeal structures (Fant, 1970; Fitch and Giedd, 1999). Another possible explanation might be related to the observed higher average intensity and larger intensity range in noise, as previous studies have shown that the intensity contour co-varies with the mouth-opening area (Chandrasekaran *et al.*, 2009; He *et al.*, 2019). The higher $F1$ might then be a consequence of having a large open-mouth area (i.e., analogous to jaw-lowering) to increase intensity in noise. To determine one of the postulated explanations, we examined correlations among $F1$, $F0$, and average intensity. The results yielded a positive correlation between $F1$ and average intensity, suggesting that $F1$ is an ancillary effect of increase in intensity.

Despite the consistent increase in $F1$ for front, central and back vowels, noise exerts differential effects on $F2$ of these vowels. While $F2$ lowers for front vowels (see Lu and Cooke, 2008) for similar observation), it does not statistically change for central or back vowels. The lower $F2$ for front vowels is not consistent with the idea that vowels are modified to be more peripheral in the vowel space for better perceptual distinction. This interpretation is further supported by the smaller $F2$–$F1$ difference across front, central and back vowels in a noisy condition. Note that $F2$–$F1$ difference is a derived measure by subtracting $F2$ from $F1$. Given the weak and selective effect of noise on $F2$, the overall $F1$ increase across all vowels could drive the resultant $F2$–$F1$ differences. It seems that noise does not enhance the peripherality of vowel distinctions.

The lack of vowel peripherallity in the presence of noise or with noise levels could also be due to the nature of our selected vowels. Three out of fvei vowels (/i:, a:, u:/) are point vowels. They are characterized by extreme $F1$ and $F2$ frequencies. These vowels are likely subjected to anatomical constraints within the oral tract as to the extent and direction that they can be made more peripheral. Such observation is in line with Wedel *et al.* (2018) showing that the manner

vowels are hyperarticulated is vowel-specific in American English, rather than the result of an expanded vowel space *per se*. On top, German has a crowded vowel space, which might constrain possible acoustic modifications, in consideration of the need to preserve phonological vowel distinctions. From a perceptual perspective, it may be more appropriate to enhance the signal *via* intensity, duration, and *F*0 than increasing vowel peripherallity in response to noise.

As expected, noise levels (0 vs −10 dB SNR) significantly increased average intensity and *F*0 and expanded intensity range. Previous work investigating different noise levels shows similar results. For instance, Ngo *et al.* (2017) found an increase in intensity and *F*0 with the increase in noise levels. In contrast to Ngo *et al.* (2017), we did not find such effect on syllable duration or vowel formants. It seems that speakers are less likely to modify syllable duration or vowel-level features such as formants in response to noise levels. One potential explanation could be related to the temporal order of the manipulated noise levels in our experimental design. Speakers always recorded no-noise sentences in between the two noise conditions, which could serve as a "reset" of the expected incremental adaptation of noise levels as observed in Ngo *et al.* (2017).

Broadly, our findings of enhancing multiple acoustic features in response to noise suggest that speakers strategically choose to encode redundant acoustic signal to minimize errors in an unreliable channel or adverse acoustic conditions. Furthermore, the increase in noise levels leading to higher intensity and *F*0 suggests that speakers make a dynamic assessment of their environment. This finding is consistent with the "hyper"- and "hypo"-articulation (H&H) model of speech production (Lindblom, 1990), where speech is viewed as an adaptive system that is sensitive to the real-time contexts for speakers.

B. Effects of contextual predictability

Our study revealed that syllables with high surprisal not only exhibit longer duration but also larger intensity range than those with low surprisal. Conversely, our findings could also be interpreted as syllables with low surprisal to exhibit shorter duration and smaller intensity range. Irrespective of the direction of change, surprisal affects not only syllable duration but also intensity range, extending the findings of Aylett and Turk (2006). This is in general agreement with previous work on the effect of predictability for duration at the level of word (Buz and Jaeger, 2016), morpheme (Tang and Bennett, 2018), syllable (Aylett and Turk, 2006), phoneme (Bybee, 2002), or interactions between levels (e.g., Hasegawa-Johnson *et al.*, 2009). Duration is shortened in more predictable contexts as evidenced in 319 different languages (Pimentel *et al.*, 2021).

The observed syllable duration effect is in line with the idea that longer syllable duration reflects explicit encoding to improve intelligibility of hard to understand units see Jaeger, 2010 and Gahl *et al.*, 2012 at the word level. As previously mentioned, the syllable duration effect could come from shortening of predictable syllable or lengthening of unpredictable syllable. This seems to argue in favor of accounts that treat frequent syllables as a holistic phonetic motor plan/unit for ease of retrieval, relative to less frequent syllables which are computed on-line (Bürki *et al.*, 2015; Laganaro, 2019; Whiteside and Varley, 1998).

At the segmental level, previous studies have found that American English vowels were more centralized in contextually more predictable syllables (Aylett and Turk, 2006), vowel space was more dispersed when they were contextually less predictable (Malisz *et al.*, 2018), and formant trajectory were generally affected by surprisal (Brandt *et al.*, 2021). Contrary to the work of Aylett and Turk (2006); Brandt *et al.* (2021), and Malisz *et al.* (2018), surprisal does not alter the first two formants of the five vowels /i:, e:, a:, o:, u:/ in the current study. Our results would seem counter to previous findings, but such divergence could arise from the following notable differences between studies: namely, the way surprisal is estimated. For instance, Aylett and Turk measured syllable probabilities by taking "uni-gram," "bi-gram," and "tri-gram" into consideration to arrive at two new factors: wide context and narrow context redundancy. They then estimated "language redundancy" according to the distribution of these new factors. In Brandt *et al.* (2021) and Malisz *et al.* (2018), a phone-based LM was used to calculate surprisal. Counter to these studies, the current study estimated surprisal in terms of a syllable-based trigram model. Because of our choice to train the LM with syllable as a unit, such mixed findings raise further questions as to predictability being estimated through these different measures and how syllable-based predictability might interact with phone-based predictability. Moreover, the current study estimates forward predictability. When a listening condition is noisy, it is possible that a listener might not immediately commit to a linguistic unit, e.g., the syllable, in perception. If a speaker considers such a listener perspective in formulating speech, backward predictability might also affect the acoustic realisation of syllables, a topic that deserves further investigation.

There are also other possible explanations for the lack of surprisal effect on vowel formants. Our study is different from previous stated research in the nature of analyzed vowels. While Brandt *et al.* (2021) and Malisz *et al.* (2018) studied all the vowels in the vowel spaces of the investigated languages, our study focused mainly on point vowels which could have a dispersion ceiling effect as discussed previously. Another important constraint for vowel formants variability in our study is that all our syllables are stressed syllables, while Aylett and Turk (2006) analyzed both stressed and unstressed syllables. Stressed syllables will constrain the degree of vowel reduction on highly predictable syllables.

C. The combined effect of noise and contextual predictability

The present study was designed to investigate the combined effect of predictability and background noise on the

J. Acoust. Soc. Am. **152** (2), August 2022

Ibrahim *et al.*   917

acoustic realization of German syllables. Initially, we hypothesized that speakers will phonetically enhance high surprisal syllables more than low surprisal syllables in noisy conditions if speakers choose to make the least predicted element intelligible and detectable in noise.

In our study, only duration and intensity range could allow us to answer this question as they were subject to both effects of surprisal and noise (Fig. 1). Contrary to our expectation, this study did not find a significant interaction between surprisal and noise effects. Our results are consistent with Zhao and Jurafsky (2009), who investigated the effects of word frequency (another measure of predictability) and noise on the acoustic realization of Cantonese tones. Although they found a main effect of noise on all tone types and word frequency for Cantonese mid-range tones, they did not find any interaction between noise and word frequency.

This parallel/additive effects of both noise and surprisal suggest separate, independent processing. The information theory framework (Shannon, 1948) could help us interpret those results. In Pate and Goldwater (2015), two types of signal encoding are distinguished: source code and channel code. While source coding focuses on message, channel coding is concerned with finding nearly optimal codes to transmit the message over a noisy channel with a low error rate (i.e., near the channel capacity). In our study, the predictability effect can be considered as message modulation (source coding) while the noise effect communication channel modulation (channel coding). The lack of an interaction finding suggests that the transmission channel is expanded (i.e., modified) without compromising source coding. Alternative explanations are also possible for the lack of interaction between surprisal and noise. One is related to the type of noise being experimentally manipulated. White noise might not have been adequate to degrade high-suprisal syllables to a larger extent than low-surprisal syllables. It is possible that noise manipulation using babble noise could have induced stronger degradation effects on the intelligibility of high- vs low-surprisal syllables to increase the chance of observing any interaction. The other explanation is related to the relatively small effect size of the interaction, compared to that of the main effect, suggesting that the test of interaction between noise and surprisal may be statistically underpowered in our study (see supplementary material).[1] Upon closer inspection of patterns from individual participants, only 8% of all participants (3 out of 38) showed significant interactions, suggesting that the interaction effect is not a preponderant group pattern. On this basis, we are more inclined to interpret our results as suggestive of the additive effects of predictability and noise. However, it is important for future studies to increase the sample size by collecting more data per speakers to further test the hypothesised interaction.

These results have implications for the need to go beyond message coding to include channel coding in formulating speech production models. Although channel coding is not part of linguistic representation (message formulation)

during planning, it shapes the phonetic output. Our study has only explored one type of channel. Different types of channel abound, e.g., talking to L2 learners or robots, etc. Channel characteristics imply the need to consider contexts as part of an enriched formulation of phonetic output during planning.

V. CONCLUSIONS

Despite the variety of studies on speech enhancement strategies, these studies have focused on the Lombard effect or predictability effect separately while our work extends previous literature by examining both factors in tandem. The present study investigated whether surprisal, or contextual unpredictability, is modulated by different levels of background white noise. The presence of noise affected all syllable-based and most of the vowel-based metrics. Noise levels affected only syllable-based measures [intensity (average, range) and F0 but not duration]. Contextual predictability only influenced duration and intensity range. Contrary to our expectation, the effect of surprisal was not more pronounced in the presence of noise or for noisier levels, as reflected by the lack of any interaction effects. These findings suggest that speakers might aim at being maximally informative to avoid any potential mishearing, even when it entails articulatory effort to produce predictable syllables.

[1]See supplementary materials at https://www.scitation.org/doi/suppl/10.1121/10.0013413 for the full tables of correlations and linear mixed effects model results.

Adank, P., Smits, R., and van Hout, R. (2004). "A comparison of vowel normalization procedures for language variation research," J. Acoust. Soc. Am. 116(5), 3099–3107.

Arciuli, J., Simpson, B. S., Vogel, A. P., and Ballard, K. J. (2014). "Acoustic changes in the production of lexical stress during lombard speech," Lang. Speech 57(2), 149–162.

Aylett, M., and Turk, A. (2004). "The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech," Lang. Speech 47(1), 31–56.

Aylett, M., and Turk, A. (2006). "Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei," J. Acoust. Soc. Am. 119(5), 3048–3058.

Bapineedu, G., Avinash, B., Gangashetty, S. V., and Yegnanarayana, B. (2009). "Analysis of lombard speech using excitation source information," in Proceedings of the Tenth Annual Conference of the International Speech Communication Association, September 6–10, Brighton, UK, pp. 1091–1094.

Baroni, M., and Kilgarriff, A. (2006). "Large linguistically-processed web corpora for multiple languages," in Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006), April 6, Trento, Italy, pp. 87–90.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). "Fitting linear mixed-effects models using lme4," J. Stat. Softw. 67(1), 1–48.

Bell, A., Brenier, J. M., Gregory, M., Girand, C., and Jurafsky, D. (**2009**). "Predictability effects on durations of content and function words in conversational English," J. Mem. Lang. **60**(1), 92–111.

Boril, H., and Pollák, P. (**2005**). "Design and collection of czech lombard speech database," in *Proceedings of the 9th European Conference on Speech Communication and Technology*, September 4–8, Lisbon, Portugal, pp. 1577–1580.

Brandt, E. (**2019**). "Information density and phonetic structure: Explaining segmental variability," Ph.D. thesis, UdS Saarbrücken, Saarbrücken, Germany.

Brandt, E., Andreeva, B., and Möbius, B. (**2019**). "Information density and vowel dispersion in the productions of bulgarian l2 speakers of german," in *Proceedings of the 19th International Congress of Phonetic Sciences*, August 5–9, Melbourne, Australia, pp. 3165–3169.

Brandt, E., Möbius, B., and Andreeva, B. (**2021**). "Dynamic formant trajectories in german read speech: Impact of predictability and prominence," Front. Commun. **6**, 643528.

Brandt, E., Zimmerer, F., Andreeva, B., and Möbius, B. (**2017**). "Mel-cepstral distortion of German vowels in different information density contexts," in *Proceedings of the Annual Conference of the International Speech Communication Association*, August 20–24, Stockholm, Sweden, pp. 2993–2997.

Brumm, H., and Zollinger, S. A. (**2011**). "The evolution of the lombard effect: 100 years of psychoacoustic research," Behaviour **148**(11/13), 1173–1198.

Bürki, A., Cheneval, P. P., and Laganaro, M. (**2015**). "Do speakers have access to a mental syllabary? ERP comparison of high frequency and novel syllable production," Brain Lang. **150**, 90–102.

Buz, E., and Jaeger, T. F. (**2016**). "The (in)dependence of articulation and lexical planning during isolated word production," Lang. Cogn. **31**(3), 404–424.

Bybee, J. (**2001**). *Language Use as Part of Linguistic Theory* (Cambridge University Press, Cambridge, UK), pp. 1–18.

Bybee, J. (**2002**). "Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change," Lang. Var. Change **14**(3), 261–290.

Castellanos, A., Benedí, J.-M., and Casacuberta, F. (**1996**). "An analysis of general acoustic-phonetic features for spanish speech produced with the lombard effect," Speech Commun. **20**(1), 23–35.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., and Ghazanfar, A. A. (**2009**). "The natural statistics of audiovisual speech," PLoS Comput. Biol **5**(7), 1–18.

Cooke, M., King, S., Garnier, M., and Aubanel, V. (**2014**). "The listening talker: A review of human and algorithmic context-induced modifications of speech," Comput. Speech Lang. **28**(2), 543–571.

Crocker, M. W., Demberg, V., and Teich, E. (**2016**). "Information Density and Linguistic Encoding (IDeaL)," Künstl. Intell. **30**(1), 77–81.

Davis, C., Kim, J., Grauwinkel, K., and Mixdorff, H. (**2006**). "Lombard speech: Auditory (a), visual (v) and av effects," in *Proceeding of 3rd International Conference on Speech Prosody*, May 2–5, Dresden, Germany, pp. 248–252.

Dreher, J. J., and O'Neill, J. (**1957**). "Effects of ambient noise on speaker intelligibility for words and phrases," J. Acoust. Soc. Am. **29**(12), 1320–1323.

Ernestus, M. (**2014**). "Acoustic reduction and the roles of abstractions and exemplars in speech processing," Lingua **142**, 27–41.

Fant, G. (**1970**). *2 Acoustic Theory of Speech Production* (Walter de Gruyter, Berlin, Germany).

Ferguson, S. H., and Kewley-Port, D. (**2007**). "Talker differences in clear and conversational speech: Acoustic characteristics of vowels," J. Speech. Lang. Hear. Res. **50**(5), 1241–1255.

Fitch, W. T., and Giedd, J. (**1999**). "Morphology and development of the human vocal tract: A study using magnetic resonance imaging," J. Acoust. Soc. Am. **106**(3), 1511–1522.

Frank, A. F., and Jaeger, T. (**2008**). "Speaking rationally: Uniform information density as an optimal strategy for language production," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, July 23–26, Washington, DC, pp. 939–944.

Fricke, J. E. (**1970**). "Syllabic duration and the Lombard effect," Int. J. Audiol. **9**(1), 53–57.

Gahl, S., Yao, Y., and Johnson, K. (**2012**). "Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech," Mem. Lang. **66**(4), 789–806.

Garnier, M., and Henrich, N. (**2014**). "Speaking in noise: How does the Lombard effect improve acoustic contrasts between speech and ambient noise?," Comput. Speech Lang. **28**(2), 580–597.

Garnier, M., Henrich, N., and Dubois, D. (**2010**). "Influence of sound immersion and communicative interaction on the lombard effect," J. Speech. Lang. Hear. Res. **53**(3), 588–608.

Godoy, E., Koutsogiannaki, M., and Stylianou, Y. (**2014**). "Approaching speech intelligibility enhancement with inspiration from lombard and clear speaking styles," Comput. Speech Lang. **28**(2), 629–647.

Hale, J. (**2016**). "Information-theoretical complexity metrics," Linguistics Lang. Compass **10**(9), 397–412.

Hasegawa-Johnson, M., Cole, J., Chen, K., Partha, L., Juneja, A., Yoon, T., Borys, S., and Zhuang, X. (**2009**). "Prosodic hierarchy as an organizing framework for the sources of context in phone-based and articulatory-feature-based speech recognition," in *Linguistic Patterns of Spontaneous Speech*, edited by S. Tseng (Academica Sinica, New York), pp. 101–128.

Hazan, V., and Simpson, A. (**2000**). "The effect of cue-enhancement on consonant intelligibility in noise: Speaker and listener effects," Lang. Speech **43**, 273–294.

Hazan, V., Tuomainen, O., Kim, J., Davis, C., Sheffield, B., and Brungart, D. (**2018**). "Clear speech adaptations in spontaneous speech produced by young and older adults," J. Acoust. Soc. Am. **144**(3), 1331–1346.

He, L., Zhang, Y., and Dellwo, V. (**2019**). "Between-speaker variability and temporal organization of the first formant," J. Acoust. Soc. Am. **145**(3), EL209–EL214.

Jaeger, T. F. (**2010**). "Redundancy and reduction: Speakers manage syntactic information density," Cogn. Psychol. **61**(1), 23–62.

Junqua, J.-C. (**1996**). "The influence of acoustics on speech production: A noise-induced stress phenomenon known as the lombard reflex," Speech Commun. **20**(1), 13–22.

Krakow, R. A. (**1999**). "Physiological organization of syllables: A review," J. Phon. **27**(1), 23–54.

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (**2017**). "lmertest package: Tests in linear mixed effects models," J. Stat. Softw. **82**(13), 1–26.

Laganaro, M. (**2019**). "Phonetic encoding in utterance production: A review of open issues from 1989 to 2018," Lang. Cogn. **34**(9), 1193–1201.

Levelt, W. J. M., Roelofs, A., and Meyer, A. S. (**1999**). "A theory of lexical access in speech production," Behav. Brain Sci. **22**, 1–38.

Lindblom, B. (**1990**). "Explaining phonetic variation: A sketch of the h&h theory," in *Speech Production and Speech Modelling*, edited by W. Hardcastle and A. Marchal (Springer, Dordrecht, the Netherlands), pp. 403–439.

Lombard, E. (**1911**). "Le signe de l'elevation de la voix" ("The sign of the rise in the voice"), Ann. Diseases Ear, Larynx, Nose Pharynx **37**, 101–119.

Lu, Y. (**2010**). "Production and perceptual analysis of speech produced in noise," Ph.D. thesis, University of Sheffield, Sheffield, UK.

Lu, Y., and Cooke, M. (**2008**). "Speech production modifications produced by competing talkers, babble, and stationary noise," J. Acoust. Soc. Am. **124**(5), 3261–3275.

Lu, Y., and Cooke, M. (**2009**). "Speech production modifications produced in the presence of low-pass and high-pass filtered noise," J. Acoust. Soc. Am. **126**(3), 1495–1499.

Malisz, Z., Brandt, E., Möbius, B., Oh, Y. M., and Andreeva, B. (**2018**). "Dimensions of segmental variability: Interaction of prosody and surprisal in six languages," Front. Commun. **3**, 25.

Meekings, S., Evans, S., Lavan, N., Boebinger, D., Krieger-Redwood, K., Cooke, M., and Scott, S. K. (**2016**). "Distinct neural systems recruited when speech production is modulated by different masking sounds," J. Acoust. Soc. Am. **140**(1), 8–19.

Möbius, B. (**2003**). "Rare events and closed domains: Two delicate concepts," Int. J. Speech Technol. **6**, 57–71.

Ngo, T. V., Kubo, R., Morikawa, D., and Akagi, M. (**2017**). "Acoustical analyses of tendencies of intelligibility in lombard speech with different background noise levels," J. Signal Process. **21**(4), 171–174.

Pate, J. K., and Goldwater, S. (**2015**). "Talkers account for listener and channel characteristics to communicate efficiently," Mem. Lang. **78**, 1–17.

Patel, R., and Schell, K. W. (**2008**). "The influence of linguistic content on the lombard effect," J. Speech. Lang. Hear. Res. **51**(1), 209–220.

Pimentel, T., Meister, C., Salesky, E., Teufel, S., Blasi, D., and Cotterell, R. (**2021**). "A surprisal–duration trade-off across and within the world's

J. Acoust. Soc. Am. **152** (2), August 2022

Ibrahim *et al.*   919

languages," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, November 7–11, Online and Punta Cana, Dominican Republic, pp. 949–962.

Pisoni, D., Bernacki, R., Nusbaum, H., and Yuchtman, M. (**1985**). "Some acoustic-phonetic correlates of speech produced in noise," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, March 26–29, Tampa, FL, pp. 1581–1584.

Pittman, A. L., and Wiley, T. L. (**2001**). "Recognition of speech produced in noise," J. Speech. Lang. Hear. Res. **44**(3), 487–496.

R Core Team (**2018**). *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria).

Samlowski, B., Möbius, B., and Wagner, P. (**2011**). "Comparing syllable frequencies in corpora of written and spoken language," in *Proceedings of the 12th Annual Conference of the International Speech Communication Association*, August 27–31, Florence, Italy, pp. 637–640.

Schiel, F. (**1999**). "Automatic phonetic transcription of non-prompted speech," in *Proceedings of the International Congress of Phonetic Sciences*, August 1–7, San Francisco, CA, pp. 607–610.

Schmid, H., Möbius, B., and Weidenkaff, J. (**2007**). "Tagging syllable boundaries with joint n-gram models," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech 2007)*, August 27–31, Antwerp, Belgium, pp. 2857–2860.

Schweitzer, A., and Möbius, B. (**2004**). "Exemplar-based production of prosody: Evidence from segment and syllable durations," in *Proceedings of Speech Prosody 2004*, March 23–26, Nara, Japan, pp. 459–462.

Shannon, C. E. (**1948**). "A mathematical theory of communication," Bell Syst. Tech. J. **27**(3), 379–423.

Shapiro, S. S., and Wilk, M. B. (**1965**). "An analysis of variance test for normality (complete samples)," Biometrika **52**(3–4), 591–611.

Simpson, A. P. (**2009**). "Phonetic differences between male and female speech," Linguistics Lang. Compass **3**(2), 621–640.

Smiljanić, R., and Bradlow, A. R. (**2005**). "Production and perception of clear speech in croatian and english," J. Acoust. Soc. Am. **118**(3), 1677–1688.

Stolcke, A. (**2002**). "Srilm - an extensible language modeling toolkit," in *Proceedings of the International Conference of Spoken Language Processing*, September 16–20, Denver, CO, pp. 901–904.

Summers, W. V., Pisoni, D. B., Bernacki, R. H., Pedlow, R. I., and Stokes, M. A. (**1988**). "Effects of noise on speech production: Acoustic and perceptual analyses," J. Acoust. Soc. Am. **84**(3), 917–928.

Tang, K., and Bennett, R. (**2018**). "Contextual predictability influences word and morpheme duration in a morphologically complex language (Kaqchikel Mayan)," J. Acoust. Soc. Am. **144**(2), 997–1017.

Traunmüller, H., and Eriksson, A. (**1994**). "The frequency range of the voice fundamental in the speech of male and female adults," Technical Report, available at https://www2.ling.su.se/staff/hartmut/f0_m&f.pdf.

Wakao, A., Takeda, K., and Itakura, F. (**1996**). "Variability of Lombard effects under different noise conditions," in *Proceeding of Fourth International Conference on Spoken Language Processing*, October 3–6, Philadelphia, PA, pp. 2009–2012.

Wedel, A., Nelson, N., and Sharp, R. (**2018**). "The phonetic specificity of contrastive hyperarticulation in natural speech," J. Mem. Lang. **100**, 61–88.

Whiteside, S., and Varley, R. (**1998**). "A reconceptualisation of apraxia of speech: A synthesis of evidence," Cortex **34**(2), 221–231.

Zhao, Y., and Jurafsky, D. (**2009**). "The effect of lexical frequency and Lombard reflex on tone hyperarticulation," J. Phon. **37**(2), 231–247.