Speech Communication 136 (2022) 53–62

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

The Lombard intelligibility benefit of native and non-native speech for native and non-native listeners

Katherine Marcoux [a,*], Martin Cooke [b], Benjamin V. Tucker [c], Mirjam Ernestus [a]

[a] Centre for Language Studies, Radboud University, P.O. Box 9103, Nijmegen, HD 6500 the Netherlands
[b] Ikerbasque (Basque Science Foundation), Bilbao, Spain
[c] Department of Linguistics, University of Alberta, 4-32 Assiniboia Hall, Edmonton, Alberta T6G 2E7, Canada

ARTICLE INFO

Keywords:
Lombard speech
Intelligibility
Non-native speakers
Non-native listeners

ABSTRACT

Speech produced in noise (Lombard speech) is more intelligible than speech produced in quiet (plain speech). Previous research on the Lombard intelligibility benefit focused almost entirely on how *native* speakers produce and perceive Lombard speech. In this study, we investigate the size of the Lombard intelligibility benefit of both native (American-English) and non-native (native Dutch) English for native and non-native listeners (Dutch and Spanish). We used a glimpsing metric to measure the energetic masking potential of speech, which predicted that both native and non-native Lombard speech could withstand greater amounts of masking to a similar extent, compared to plain speech. In an intelligibility experiment, native English, Spanish, and Dutch listeners listened to the same words, mixed with noise. While the non-native listeners appeared to benefit more from Lombard speech than the native listeners did, each listener group experienced a similar benefit for native and non-native Lombard speech. Energetic masking, as captured by the glimpsing metric, only accounted for part of the Lombard benefit, indicating that the Lombard intelligibility benefit does not only result from a shift in spectral distribution. Despite subtle native language influences on non-native Lombard speech, both native and non-native speech provides a Lombard benefit.

1. Introduction

Whether at grocery stores, restaurants, or cafes, on a daily basis we hear background noise and speak in it, producing Lombard speech (Lombard, 1911). Lombard speech is acoustically different from plain speech, that is speech produced in quiet. These acoustic modifications allow Lombard speech to be better understood in noise compared to plain speech, providing a Lombard intelligibility benefit (e.g., Dreher and O'Neill, 1957; Pittman and Wiley, 2001). Research to date has heavily focused on native speakers producing Lombard speech as well as native listeners. Considering that non-native speakers are influenced by their native language when speaking, the question arises as to whether non-native Lombard speech is produced differently and in turn how this affects the size of the Lombard benefit for non-native speech. This study is the first to investigate the perception of non-native Lombard speech.

Past research has predominantly examined native speakers' production of Lombard speech in English (e.g., Dreher and O'Neill, 1957; Pisoni et al., 1985; Pittman and Wiley, 2001; Van Summers et al., 1988), as well as several other languages such as French (e.g., Garnier and Henrich, 2014), Spanish (e.g., Castellanos et al., 1996), and Dutch (e.g., Bosker and Cooke, 2020). These studies have established that native Lombard speech involves acoustic modifications that make it distinct from plain speech. These include but are not limited to: an increase in fundamental frequency (F0), a wider F0 range, an increase in intensity, and a shift in energy to higher frequencies (for a review see: e.g., Cooke et al., 2014).

Only a handful of studies have investigated Lombard speech produced by non-native speakers. By investigating non-native Lombard speech in combination with native Lombard speech, we can better understand Lombard speech itself. Analyzing non-native Lombard speech will reveal the potential influence of the native language on the non-native Lombard speech, leading to insights as to whether Lombard speech is language general, or whether there may be some aspects that are more language specific.

Two studies that investigated the acoustics of non-native Lombard speech (Marcoux and Ernestus, 2019a, 2019b) examined native

* Corresponding author.
E-mail addresses: k.marcoux@let.ru.nl (K. Marcoux), m.cooke@ikerbasque.org (M. Cooke), benjamin.tucker@ualberta.ca (B.V. Tucker), m.ernestus@let.ru.nl (M. Ernestus).

https://doi.org/10.1016/j.specom.2021.11.007
Received 12 November 2020; Received in revised form 10 September 2021; Accepted 29 November 2021
Available online 1 December 2021

American-English speakers and native Dutch speakers in addition to native Dutch speakers producing non-native English plain and Lombard speech. Their research used the Dutch English Lombard Native Non-Native corpus (DELNN; Marcoux and Ernestus, 2019b; Marcoux and Ernestus, in preparation). Marcoux and Ernestus (2019b) found that the Dutch non-native speakers of English increased their median F0 for Lombard speech compared to plain speech, as is characteristic of native Lombard speech. They also found subtle native (Dutch) language influences on the non-native Lombard speech production in terms of the amount of F0 increase in the different conditions. The stimuli consisted of question-answer pairs, where the word with contrastive focus was early in the answer (early-focus condition) or late in the answer (late--focus condition). This meant that the material that underwent post-focus compression (a narrowing and lowering of the F0 range for the post-focus stimuli; see e.g., Xu, 2011) differed in the two conditions. While the two groups increased their F0 to a similar extent for Lombard speech in the late-focus condition, the non-native speakers had a larger increase in F0 (average of 12.7 Hz increase) than the native English speakers (average of 6.5 Hz increase) for the early-focus condition, reflecting the larger increase in Dutch (average of 11.6 Hz increase for early-focus).

Marcoux and Ernestus (2019a) also examined the increase in F0 range in Lombard speech in natives and non-natives, finding an overall increase in F0 range in both the native and non-native speakers. Again a difference was found between the native and non-native English speakers, with the native speakers having a larger F0 range increase than the non-natives in the late-focus condition. The native Dutch had the smallest increase in Dutch, indicating that the non-native speakers were being influenced by their native Dutch. These results were not driven by individual speakers. These studies on non-native Lombard speech production, although limited to median F0 and F0 range, suggest that non-native speakers are adapting their speech in noise in ways that are characteristic of native Lombard speech, while showing faint influences of the native language.

Three other studies have investigated non-native Lombard speech in two more languages. Villegas and colleagues (2021) showed that Japanese speakers produced louder speech (as measured by sound pressure level) both in native Japanese and in non-native English Lombard speech relative to plain speech. Cai and colleagues (2020) found the same with intensity for Chinese-English late-bilinguals (first language – L1 – Chinese, second language – L2 – English). Mok and colleagues (2018) investigated mean intensity, mean F0, and durations of vowels in L1 Mandarin and L2 English speech. They found that compared to L1 Mandarin plain speech, Lombard speech had higher mean intensity and longer durations, and for two of the three tones studied, higher mean F0. For the L2 English speech, they also found higher intensity and longer durations for Lombard speech compared to plain speech. However, the mean F0 was lower for L2 English Lombard speech in comparison to plain speech. In combination, these studies indicate that non-native speakers may produce Lombard speech, but may apply different modifications than native speakers do.

Past research with native speech has shown that the acoustic characteristics of Lombard speech provide an intelligibility benefit, resulting in Lombard speech being better understood in noise compared to plain speech presented in noise, that is, a Lombard intelligibility benefit (e.g., Dreher and O'Neill, 1957; Pittman and Wiley, 2001). In examining the Lombard benefit, the speech (plain, Lombard) is mixed with noise, and the intelligibility of the masked stimuli is measured. The signal-to-noise-ratio (SNR) is fixed at the same level for the plain and Lombard speech, allowing for a comparison of the intelligibility between the two speech styles. The Lombard benefit has been shown to be influenced by various factors, including the type and the intensity of the noise used to elicit the Lombard speech, as well as the noise used as a masker and the SNR of the masked stimuli (e.g., Lu and Cooke, 2008; Van Summers et al., 1988).

All studies on the Lombard benefit have investigated native speech. While the vast majority has also focused on native listeners, a couple of studies have examined both native and non-native listeners. One such study was conducted by Junqua (1993). He did not find a Lombard benefit for native nor for non-native listeners.

Another study involving non-native listeners was performed by Cooke and García Lecumberri (2012), who found that non-native listeners show a Lombard benefit for native speech. In line with previous research with natives (e.g., Van Summers et al., 1988), Cooke and García Lecumberri (2012) reported that for non-native listeners, the size of the benefit also depends on the SNR level of the stimuli used in the intelligibility experiment, with higher noise levels eliciting a larger Lombard benefit. Also in agreement with past research on native listeners (Lu and Cooke, 2008), the level of noise used to elicit the Lombard speech affected the size of the Lombard benefit. The Lombard benefit for the non-native listeners, however, appears to be smaller than for native listeners who were tested with the same stimuli in another experiment (Lu and Cooke, 2008). With Lombard speech produced in 82 dB SPL of noise and plain and Lombard speech tested at an SNR of -9 dB, the native listeners increased their intelligibility 22% points when going from plain to Lombard speech (Lu and Cooke, 2008). In comparison, under the same conditions, non-native listeners increased their intelligibility by 15% points (Cooke and García Lecumberri, 2012). Together, these findings suggest that the Lombard benefit for non-native listeners is influenced by various factors similarly to native listeners (SNR levels, noise levels in producing Lombard speech, etc.) and that, although non-native listeners experience the Lombard benefit, they may do so to a smaller extent than native listeners.

One possible explanation for the smaller Lombard intelligibility benefit for non-native listeners is that in general non-native listeners can be more adversely affected than natives by noise in word recognition tasks (for a review of the literature see: García Lecumberri et al., 2010; Scharenborg and van Os, 2019). Being more adversely affected by noise could mean that the non-native listeners may need to dedicate more cognitive resources to word recognition, and in turn may not be able to take full advantage of the Lombard cues that contribute to the Lombard benefit. Alternatively, there may be language specific modifications in Lombard speech, which the non-native listeners may not take full advantage of.

The few studies investigating the Lombard benefit for non-native listeners only examined one non-native listener group each (Cooke and García Lecumberri, 2012; Junqua, 1993). The benefit may, however, vary depending on the speaker's and the listener's native language. In examining the intelligibility of plain speech, Bent and Bradlow (2003) found a "matched interlanguage speech intelligibility benefit". That is, a non-native listener understands a non-native speaker with whom they share the same native language as well as a native speaker. Other studies have not found such straightforward results. For instance, Stibbard and Lee (2006) only found weak evidence for the matched interlanguage speech intelligibility benefit, and Major and colleagues (2002) only found the effect for one of several listener groups. Bent and Bradlow (2003) further found a "mismatched interlanguage speech intelligibility benefit," with non-native listeners benefitting from non-native speakers independently of whether they share their native language. Other research has not replicated this mismatched interlanguage speech intelligibility benefit (e.g., Stibbard and Lee, 2006). Therefore, it is still an open question whether and how a speaker's intelligibility is co-determined by the exact combination of the speaker's and listener's native languages.

In our study, we investigated the Lombard benefit of non-native speech, taking into account that this benefit may depend on the combination of the speaker's and listener's native languages. The materials (English target words) were taken from the DELNN corpus (Marcoux and Ernestus, 2019b; Marcoux and Ernestus, in preparation), which, as mentioned above, contains English speech from native (American--English) and non-native (native Dutch) English speakers as well as native Dutch speech. Lombard speech has been documented for both

native English and native Dutch and the Lombard speech in the two languages show similarities (e.g., English: Bosker and Cooke, 2018; Lu and Cooke, 2008; Pisoni et al., 1985; Van Summers et al., 1988; e.g., Dutch: Bosker and Cooke, 2020). Additionally, the acoustics of English Lombard speech produced by Dutch natives have been briefly studied, and it appears that their non-native English Lombard speech is very similar to native English Lombard speech, albeit perhaps with some native language influence (Marcoux and Ernestus, 2019a, 2019b).

We first analyzed the speech signal itself, investigating the masking potential of English target words using the high-energy glimpsing proportion metric (HEGP, Tang and Cooke, 2016). The HEGP metric is an extension of the glimpse proportion metric (GP). GP measures the proportion of spectro-temporal regions (glimpses) in speech tokens where the energy is greater for the speech than for the noise (Cooke, 2006). Lu and Cooke (2008) found that Lombard speech has higher GPs than plain speech and furthermore that the GPs correlate with human intelligibility. The HEGP metric extends GP by examining each frequency band separately and selecting only those glimpses that additionally have an energy that exceeds the average speech-plus-noise energy in that band. The extension of GPs by HEGPs results in an improved correlation with intelligibility scores (Tang and Cooke, 2016). HEGPs range in value from 0 to 1, which ideally maps on to the range from 0 to 100% intelligibility, although in practice the mapping depends on the type of speech material and the masker (see Fig. 3 in Tang and Cooke, 2016). This acoustic measure is independent of the listener's native language and therefore provides language-independent intelligibility information on native versus non-native Lombard speech. As there may be native language influences on non-native plain and Lombard speech, we may expect differences in HEGPs between native and non-native speech, in addition to differences between plain and Lombard speech.

We then tested the same material in an intelligibility experiment with native and non-native listeners. Listeners were asked to identify the English target words, produced by native and non-native English speakers in plain and in Lombard speech, mixed with noise. By having one group of native and two groups of non-native listeners, we can investigate how the speaker's and listener's native languages contribute to a Lombard intelligibility benefit. Canadian listeners served as our native cohort. One non-native listener group consisted of native Dutch individuals, chosen because they shared the native language with the non-native speakers. Our other non-native listener group consisted of native Spanish individuals, chosen as they did not share the native language with any of the speakers. The two non-native listener groups allowed us to examine the matched and mismatched interlanguage speech intelligibility benefit (Bent and Bradlow, 2003).

Listeners' responses were initially analyzed independently from HEGPs. Subsequently, we performed an additional analysis on the listeners' responses in which HEGP predictions were incorporated, in order to clarify whether any seemingly language-dependent differences might be ascribed to the ability to withstand energetic masking.

2. Speech materials

2.1. Speakers

The speech materials for this study were taken from the DELNN corpus (Marcoux and Ernestus, 2019b; Marcoux and Ernestus, in preparation). Of the nine native (American-English) and thirty non-native (native Dutch) female English speakers in the corpus, we selected eight native and eight non-native speakers. One native speaker did not agree to have her recordings used online, leaving us with the needed eight. We selected the eight mid-accented non-native speakers from the 23 who agreed to have their recordings used online based on the results from an overall accentedness rating experiment reported in Marcoux, Süß, and Ernestus (in preparation). In the accentedness experiment, six native American-English listeners rated six sentences per non-native speaker, using a 7-point Likert scale (1 "native-like" to 7 "very strong foreign accent"). These native listeners also rated two native speakers, which resulted in averages of 1.00 and 1.03, respectively, for each of the speakers, confirming their ability to identify native speech. After averaging the ratings per non-native speaker, the eight non-native speakers who were closest to the median rating of 4.7 were selected. Their average accentedness ratings for these eight speakers ranged from 4.4 to 5.0. Further, these selected non-native speakers had an average LexTALE (Lemhöfer and Broersma, 2012) score of 64.0 ($sd = 9.2$), corresponding to a B1 level in the Common European Framework (Council of Europe, 2001). We assume these selected speakers to be normal representations of Dutch natives, who typically start learning English at the age of ten or eleven years. Dutch natives are constantly exposed to English via English movies and series, as most entertainment is not dubbed.

2.2. Stimuli

For the DELNN corpus, speakers read question-answer pairs at their own pace. The 96 target words were taken from the 72 early-focus sentences, where the word with contrastive focus came early in the sentence (for further details see Marcoux and Ernestus, 2019b; Marcoux and Ernestus, in preparation). In some cases, multiple target words were taken from one sentence. The majority of target words were produced as nouns (e.g., *table, gloves, city*), while some were produced as verbs but could function as nouns (e.g., *left, likes, move*). Their frequencies of occurrence ranged from 0.7 to 1958.6 in a million ($M = 145.8$, $sd = 255.3$) as reported in SUBTLEX US (Brysbaert and New, 2009). This large range in frequency of occurrence is a result of the design of the corpus, which limited us in the selection of target words. The target words always came at some point after the word with contrastive focus (anywhere from immediately after the word with contrastive focus to the last word in the sentence); thus they all underwent post-focus compression, a narrowing and lowering of the F0 range (e.g., Xu, 2011). An example question-answer pair is included in (1). Speakers were asked to place emphasis on the words in bold, resulting in contrastive focus in the answers (*Paul* in this example). From the answer in (1), we extracted the word *café*, to be used in the HEGP analysis and intelligibility experiment. See Appendix 1 for all target words used.

(1) Did **Simon** meet his professor at the café to talk?
No, **Paul** met his professor at the café to talk.

A total of 662 target word tokens were chosen from the DELNN corpus, which differed from each other in the combination of speech style (plain, Lombard), speaker, and word. Not all speakers and words contributed equally to the stimuli because not all eight native and eight non-native speakers produced all 96 target words both in plain and in Lombard speech in the DELNN corpus. Each speaker contributed between 35 and 48 target word tokens. Each target word was produced the same number of times as plain and as Lombard speech. Therefore, 331 of the target words were produced as plain speech and the other 331 as Lombard speech (this ranged from two to four productions of plain / Lombard speech per target word with an average of 3.4).

The corpus was phonemically transcribed and segmented at the word level using the Montreal Forced Aligner (McAuliffe et al., 2017). The first author examined all oscillograms and spectrograms and improved the segmentation if needed, before the target word tokens were extracted at the zero-crossing boundaries.

2.3. Procedure

The speakers in the DELNN corpus completed the self-paced reading task wearing a pair of Sennheiser HD 215 MKII DJ over ear headphones. Nothing was played via the headphones for the plain speech. In contrast, to elicit Lombard speech, participants heard speech shaped noise (SSN) at 83 dB SPL (calibrated using a Brüel & Kjær Type 4153 artificial ear) through the headphones. Wearing headphones during the plain speech

condition may have caused speakers to produce some amount of Lombard speech because their own voice was attenuated. We wanted the Lombard condition to purely reflect the effect of noise rather than noise in combination with headphone noise attenuation.

The SSN was generated by passing random noise through a filter whose spectrum matched the average spectrum of male and female voices. The average spectrum was created from the recordings of 10 male and 10 female adults reading a phonetically balanced text that was approximately two minutes long. The resulting SSN file had a sampling frequency of 44.1k Hz and was a single-channel WAV file, as is the case with all speech and noise materials used.

3. High energy glimpse proportion analysis

3.1. Procedure

HEGPs were calculated at a global SNR of -1 dB, as this was the SNR designated for the intelligibility experiment (Section 4). Each token was constructed by mixing the speech signal with a randomly-chosen fragment of the SSN masker used to elicit Lombard speech in the DELNN corpus. The speech and masker signals making up each noisy token were independently passed through a gammatone filterbank consisting of 55 filters with center frequencies in the range of 50–8000 Hz. The instantaneous Hilbert envelope at the output of each filter was subsequently smoothed with a first-order leaky integrator (8 ms time constant) and downsampled to 100 Hz (i.e. 10 ms frames) for glimpse calculation. Candidate glimpses were then defined as those 10 ms time-frequency regions where the energy in the resulting spectro-temporal representation of the speech exceeded that of the masker (i.e., a local SNR of 0 dB was employed). Subsequently, in each frequency band, only those candidates occupying time regions where the speech-plus-noise mixture energy exceeded the mean energy of the mixture in that band were retained. Details of the HEGP calculation are provided in Tang and Cooke (2016).

3.2. Analyses

HEGPs were transformed into logits using the equation: ln (proportion/1-porportion) (Jaeger, 2008) and analyzed using linear mixed effects models (lmers) from the *lme4* package (version 1.1.21) (Bates et al., 2015) in R (version 3.5.1) (R Core Team, 2016). Visualizations were made using the *ggplot2* package (version 3.2.1) (Wickham, 2016). The predictors of interest were Speech Style (plain, Lombard) and Speaker Nativeness (native, non-native). Speaker and Target Word were crossed-random intercepts and the significant predictors of interest were also tested as random slopes.

In the analysis, the Nelder-Mead optimizer was used as it provided the most robust convergence results. Outliers were defined as data points more than 2.5 standard deviations away from the grand mean, and we removed 18 such outliers prior to modeling. We started with all the predictors and interactions of interest and did a backwards fitting procedure for the fixed effects structure, and then a forward fitting procedure for the random slopes. This meant that we started by including an interaction between Speech Style and Speaker Nativeness (our predictors of interest) and then removed the interaction since it was not significant ($t < 1.96$) (backwards fitting procedure). The simple effects of the predictors of interest were confirmed as significant via the summary() function ($t > 1.96$). Once the fixed effects structure was established, the significant fixed predictors were tested as random slopes as well as the interaction and anova() was used to determine if the addition improved the model (forward fitting procedure). If the model resulted in a convergence warning, we did not proceed with that model as it indicated that model was too complex given the dataset. As a final step, we took the previous model and removed the data points resulting in absolute standardized residuals exceeding 2.5 and refitted the model, which is reported below. The significance of the fixed effects of this final model were confirmed via the function summary() and via Anova(), from the *car* package (version 3.0.6) (Fox and Weisberg, 2019), which computes Type II Wald chi-square tests. For the model reported below, the levels plain speech (predictor Speech Style) and non-native speakers (predictor Speaker Nativeness) are on the intercept.

3.3. Results

The statistical model revealed a significant simple effect of Speech Style as well as of Speaker Nativeness (Table 1). Lombard speech had higher HEGPs than plain speech and the native speakers had higher HEGPs than the non-natives (Fig. 1). The lack of a significant interaction between Speaker Nativeness and Speech Style ($\beta = 0.0$. $t = -1.5$) suggests that the difference in resistance to masking between plain and Lombard speech was similar for the native and non-native speech. The random effects structure revealed that HEGPs differed per Target Word, per Speaker, and that the effect of Speech Style varied per Speaker and per Target Word.

4. Intelligibility experiment

4.1. Methods

4.1.1. Participants

Our native participants were 42 Canadians (32 females) with an average age of 21.5 years (*M*) and a standard deviation (*sd*) of 3.8 years. These participants were born and raised only in English by native English-speaking parents in Canada. They had no knowledge of Dutch or German as well as no contact with Dutch or German speakers. No participant had spent more than three months in non-English speaking countries. Additionally, participants reported no hearing loss or reading problems. These participants were recruited at the University of Alberta, in Edmonton, Canada.

Our non-native participants were 47 native speakers of Spanish or Spanish/Basque bilinguals, hereafter referred to as Spanish (39 females; $M = 20.2$ years, $sd = 2.4$ years) and 46 native Dutch individuals (36 females; $M = 21.7$ years, $sd = 2.7$ years). The Spanish participants were students at the Universidad del País Vasco/Euskal Herriko Unibertsitatea (UPV/EHU) in Vitoria, Spain, while the Dutch participants were students at Radboud University, Nijmegen, The Netherlands. The Spanish participants were enrolled in English Philology at the Faculty of Arts, while the Dutch participants came from different majors, with the additional requirements that they did not study linguistics and that more than half of their classes were not in English. These different study requirements for the two non-native groups were in place so that they had similar levels of English proficiency. On average, both the Spanish and Dutch participants had a B2 level of English in the Common European Framework (Council of Europe, 2001) as indicated by their LexTALE scores; $M = 69.6$, $sd = 8.5$ and $M = 66.5$, $sd = 14.6$, respectively (Lemhöfer and Broersma, 2012). We did not find a difference between

Table 1
lmer model of the logit HEGPs. Plain speech and non-native speaker are on the intercept. After each predictor, we indicate which level is being contrasted with the intercept. The β indicates the size of the difference between the intercept level and the contrasted level.

Fixed effects:	β	*t-value*
Intercept	0.3	9.4
Speech Style: Lombard	0.1	6.3
Speaker Nativeness: Native	0.1	2.1
Random effects		*SD*
Target word (Intercept)		0.1
Speech Style by Target Word		0.1
Speaker (Intercept)		0.1
Speech Style by Speaker		0.1
Residual		0.1

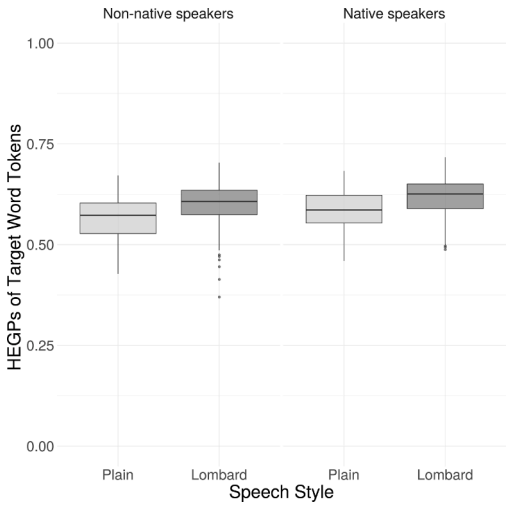

Fig. 1. HEGPs of target word tokens produced in plain and Lombard speech split by speaker nativeness. The data in the graph are the raw data (pre-logit transformation), including outliers, which were excluded from the analyses. The box represents the lower and upper quartiles, with the horizontal line within the box indicating the median. The dots indicate potential outliers while the lines extend to the minimum and maximum, excluding potential outliers.

the Dutch and Spanish participants' English proficiency as indicated by the LexTALE scores (independent non-parametric two samples Wilcoxon rank test was conducted as the Dutch data was not normally distributed, $W = 862$, $p = 0.09$). None of the non-native participants reported any hearing loss or reading problems and had not been in English-speaking countries for more than two months. Moreover, the Dutch participants were not speakers from the DELNN corpus (Marcoux and Ernestus, 2019b; Marcoux and Ernestus, in preparation). All participants gave informed consent and were compensated financially or with course credit for their participation in the experiment.

4.1.2. Speech materials

The same speech materials as in the HEGP analysis were used for the intelligibility study. The 662 target word tokens were mixed with a random section of the noise from the same SSN file that was used to calculate the HEGPs. For the intelligibility experiment, a 30 millisecond ramp of noise preceded and followed the tokens. An SNR of -1 dB was chosen on the basis of a pilot which used Dutch non-native listeners of English. These pilot participants did not partake in the intelligibility experiment but had a similar background as the participants that did. Pilot participants were tested on a subset of the tokens using SNRs ranging from -2 to +4 dB. An SNR of -1 dB ensured that the performance on the easiest condition (native speakers, Lombard speech) was not at ceiling while the most difficult condition (non-native speaker, plain speech) was not at floor.

The experiment also included one filler word per speaker (see below), which were not included in the analyses. These filler words were included so the listener could adapt to each speaker. These fillers also came from the DELNN corpus and were also nouns.

4.1.3. Experimental lists

Each of the 12 experimental lists contained 192 trials of interest, made up of 96 distinct target words, produced twice, once by a native and once by a non-native speaker. Within each list, eight speakers (four native, four non-native) produced 24 target words each. Of the target word tokens, half were plain and half were Lombard speech. The lists were blocked by speaker and the first item of each block was a filler. The fillers were included at the beginning of each speaker block so the listeners could adapt to the individual speakers. The lists were pseudorandomized so that no more than two native or two non-native speakers followed each other and that the second instance of the target word was at least 10 trials after the first.

To create the 12 experimental lists, we started with one list containing eight speakers – four native and four non-native. Speakers were grouped into pairs, each speaker had a corresponding speaker that produced the same target words in the other speech style (plain and Lombard). A second list was created by taking four of the speakers from the first list and adding four different speakers to it. From the two lists, we created two mirror lists, which contained the other speakers from each pair (eight speakers), and what was plain speech was now Lombard speech and vice versa. The order of speakers in these four lists and the words in each of these speaker blocks was randomized with the restrictions described above, resulting in the final 12 experimental lists. Each participant completed three practice trials (not target words) by one native and two non-native speakers (selected from our 16 speakers) followed by an experimental list.

4.1.4. Procedure

In the intelligibility experiment, listeners heard isolated words in plain and Lombard speech styles produced by native and non-native speakers mixed with SSN. For each trial, participants heard the stimulus while seeing a blank screen and, after 0.1 second, they were instructed to "Write down the word you heard:". If participants attempted to continue the experiment without a response, a message appeared on the screen asking them to "Please fill in the blank". The computers in the three countries had autocorrect activated, but there were misspellings that the auto-correct did not catch, such as "fundation" for "foundation". The experiment included three self-paced breaks, each after two speaker blocks.

In addition to the intelligibility experiment itself, the Dutch and Spanish participants completed the aforementioned LexTALE task (Lemhöfer and Broersma, 2012). For the task, participants needed to indicate whether the 60 test items presented orthographically were English words or non-words. Of the 60 test items, 40 were words and 20 were non-words. This task is used to estimate general English proficiency as per the levels in the Common European Framework (Council of Europe, 2001).

The Spanish participants completed the intelligibility experiment on Mac Mini computers using MacOS Sierra and Sennheiser HD 380 pro headphones. Up to four participants completed the experiment simultaneously in separate alcoves in a sound attenuated room. The Dutch participants completed the same experiment on Dell Latitude 5590 laptops using Windows 10 and Sennheiser HD 215 MKII DJ headphones. The Canadian participants completed the experiment on Dell Optiplex 3020 computers running Windows 7 and wore MB Quart QP 805 DEMO headphones. For both the Dutch and Canadian participants, up to two participants were run at once, each in their own sound attenuated booth.

The participants heard the stimuli at an average fixed volume of 71 dBA. In Spain and The Netherlands, the volume coming from the headphones was calibrated on a subset of concatenated stimuli using the Brüel & Kjaer artificial ear type 4153 and in Canada it was calibrated using the EXTECH Instruments 407750 Digital Sound Level Meter with RS232 and Sound Level Calibrator 407744.

4.2. Analyses

To analyze intelligibility, participants' responses were cleaned of spurious characters such as "." and "\". Further, when multiple worded or multiple answers were given, only the first word was considered (e.g., only "gang" was considered in "gang or game"). Answers were coded as correct (1) or incorrect (0). An answer was only considered correct if there were no misspellings. Correctly spelled homonyms of the target word, such as "weak" for "week" were also considered correct.

Intelligibility was analyzed with generalized linear mixed effects models (glmers) with the binomial link function. The number of maximal iterations was increased to 100,000 in "bobyqa". The

predictors of interest were Speech Style (plain, Lombard), Speaker Nativeness (native, non-native), and Listener Group (Canadian, Dutch, Spanish). The control predictors were Final, Trial Number, Occurrence, and Focus. Final (final, non-final) was included to indicate whether the target word token was produced as the final word in the sentence or not, since final words are typically lengthened and may therefore be easier to understand. Scaled and centered Trial Number was included since listeners' accuracy may improve during the experiment due to learning or drop during the experiment because of fatigue. Occurrence (first, second) indicated whether it was the listeners' first or second occurrence of hearing the target word and was included because priming could make the second occurrence easier to comprehend. Finally, Focus indicated whether the target word token came immediately after the word with contrastive focus in the original sentence in the corpus or at some later point in the sentence. Focus was included since we were unsure whether the effect of post-focus compression on intelligibility varies based on the closeness to the word with contrastive focus. Speaker, Listener, and Target Word were the crossed-random effects.

The same statistical procedure was followed as when analyzing HEGPs in determining the best model, using a backwards fitting procedure for the fixed effects structure and then a forward fitting procedure for the random slopes. We began with a theory-based approach with simple effects and interactions among our predictors of interest (Speech Style, Speaker Nativeness, and Listener Group) and simple effects for our control predictors (Final, Trial Number, Occurrence, Focus). In the model reported below, plain speech (Speech Style), non-native speakers (Speaker Nativeness), and Canadian listeners (Listener Group) are on the intercept.

Because the Canadian listeners are on the intercept, the model does not provide detailed information about the potential differences between the Dutch and Spanish listeners. We therefore releveled the final model with the Dutch listeners on the intercept and also report the relevant results of this releveled model.

4.3. Results

Due to technical issues, 193 trials for the Spanish participants were lost (one Spanish participant lost 52 trials while the rest randomly lost between 0 and 9 trials; the total loss is approximately 2% of the Spanish data). This resulted in 8832 and 8829 data points for the Dutch and Spanish listener cohorts, respectively.

These data are visualized in Fig. 2 and the final statistical model is shown in Table 2. This final model lacks some of the predictors and interactions that we tested because they were not statistically significant. While Fig. 2 may suggest there is a three-way interaction among our predictors of interest (Speech Style * Speaker Nativeness * Listener Group), this was not borne out in the statistical analysis, and therefore this interaction was removed from the model (Speech Style: Lombard * Speaker Nativeness: Native * Listener Group: Dutch $\beta = 0.2$, $z = 1.5$, $p = 0.1$ and Speech Style: Lombard * Speaker Nativeness: Native * Listener Group: Spanish $\beta = 0.1$, $z = 0.6$, $p = 0.6$). The interaction of Speech Style with Speaker Nativeness was no longer significant ($\beta = -0.2$, $z = -0.9$, $p = 0.4$) after Speech Style was added as a random slope to the Speaker random intercept and therefore also removed. The control predictor Focus was also not significant and removed from the model ($\beta = 0.1$, $z = 0.1$, $p = 0.9$).

With regard to non-native speech, which is at the intercept, we found no difference between the Canadian (at the intercept) and Dutch listeners, while the Spanish listeners performed worse compared to the Canadians. A releveled model with the Dutch listeners, instead of the Canadian listeners, on the intercept, showed that the Spanish listeners also performed worse than the Dutch listeners (Listener Group: Spanish $\beta = -0.2$, $z = -2.0$, $p < 0.05$).

Overall the native speakers were better understood than the non-native speakers, and this was more so for the native (Canadian) listeners (interaction of Speaker Nativeness and Listener Group) than the non-native listeners. The Dutch and Spanish listeners showed a similarly sized effect of the nativeness of the speech (in the releveled model with the Dutch listeners on the intercept, the interaction of Speaker Nativeness with the Spanish Listener Group was not significant: $\beta = 0.1$, $z = 0.9$, $p = 0.4$).

Lombard Speech was better understood than plain speech. This Lombard benefit was larger for non-native listeners (interaction of Listener Group and Speech Style), with the Dutch and the Spanish listeners showing a similarly sized effect (in line with this, the releveled model with the Dutch listeners on the intercept showed no significant interaction of Speech Style with the Spanish Listener Group: Lombard $\beta = 0.0$, $z = 0.5$, $p = 0.6$). As there is no three-way interaction of Speaker

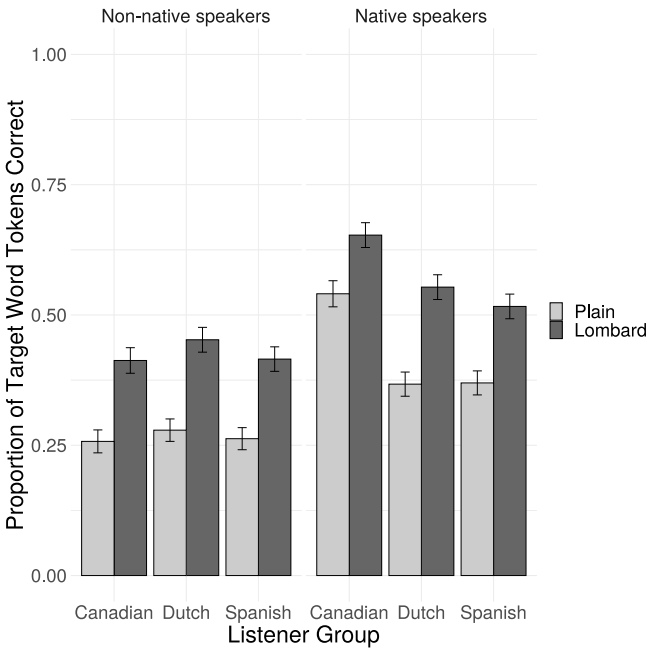

Fig. 2. The proportion of target word tokens correct for plain and Lombard speech split by Speaker Nativeness and by Listener Group. The error bars indicate 95% confidence intervals.

Table 2
The glmer model of intelligibility scores for native and non-native plain and Lombard speech by native and non-native listeners. Plain speech, non-native speaker, and native listener are on the intercept. After each predictor, we indicate which level is being contrasted with the intercept. The β indicates the size of the difference between the intercept level and the contrasted level.

Fixed effects	β	z	P
(Intercept)	-1.6	-4.9	<0.001
Speaker Nativeness: Native	1.7	4.8	<0.001
Listener Group: Dutch	0.0	-0.4	0.7
Listener Group: Spanish	-0.3	-2.2	<0.05
Speech Style: Lombard	0.9	5.3	<0.001
Trial Number	0.1	4.6	<0.001
Occurrence: Second	0.4	9.2	<0.001
Final: Non-final	-0.5	-2.2	<0.05
Speaker Nativeness: Native* Listener Group: Dutch	-0.9	-11.4	<0.001
Speaker Nativeness: Native* Listener Group: Spanish	-0.8	-10.3	<0.001
Listener Group: Dutch * Speech Style: Lombard	0.2	3.1	<0.01
Listener Group: Spanish * Speech Style: Lombard	0.3	3.5	<0.001
Random effects			*SD*
Listener (Intercept)			0.3
Target Word (Intercept)			1.5
Speech Style: Lombard by Target Word			1.3
Listener Group: Dutch by Target Word			0.4
Listener Group: Spanish by Target Word			0.7
Speaker (Intercept)			0.7
Speech Style: Lombard by Speaker			0.4

Nativeness, Listener Group, and Speech Style, the effect of Lombard speech appears similar for native and non-native speech.

Regarding the control variables, when the stimulus was produced as the last word in the sentence in the corpus, participants did better, suggesting that final lengthening improved intelligibility. As trial number increased, performance improved, indicating that learning occurred over the course of the experiment. Additionally, the second occurrence of the word was better understood than the first, suggesting that priming aided the participants.

From the model's random effect structure, we learned that the intelligibility scores varied for Listeners, Target Words, and Speakers. Additionally, Speech Style affected different Target Words differently as well as affecting Speakers differently. We also observed that the different Listener Groups were affected differently by the different Target Words.

4.4. Intelligibility analysis including HEGPs as predictor

4.4.1. Analyses

We extended the analysis of participants' intelligibility scores provided in Section 4.2 by including the additional predictor HEGPs. As explained above, the HEGPs form an acoustic measure indicating the ability of a word token to resist the noise masking used in the intelligibility experiment. This additional analysis can be seen as a control analysis to see whether the effects of Speech Style and its interactions are still statistically significant after inclusion of an acoustic predictor that may partly account for the Speech Style effect.

The HEGPs from Section 3 were scaled and centered, because not doing so led to convergence issues. In determining the model, the same fitting procedure as in the previous analysis (Section 4.2) was followed. Compared to the previous analysis, we began by including the additional interaction of Listener Group with HEGPs. We did not include all interactions between the HEGPs and the variables of interest because the results from the statistical analysis of HEGPs in Section 3.3 showed that HEGPs are highly correlated with both Speaker Nativeness and Speech Style. We therefore excluded the interactions between HEGPs and these variables of interest and all higher order interactions containing these variables.

We established whether the model including the HEGPs explained more variance than the model without this predictor by comparing their Akaike Information Criterion (AICs). Because models can only be compared on the basis of their AIC if they are based on the same dataset, we made the comparison on the models before the removal of the residual outliers.

4.4.2. Results

When fitting this model, as with the previous model (Section 4.2), the three-way interaction (Speech Style * Speaker Nativeness * Listener Group) as well as the interaction of Speech Style with Speaker Nativeness were not significant and therefore removed from the model (Speech Style: Lombard * Speaker Nativeness: Native * Listener Group: Dutch $\beta = 0.3$, $z = 1.8$, $p = 0.1$, Speech Style: Lombard * Speaker Nativeness: Native * Listener Group: Spanish $\beta = 0.2$, $z = 1.0$, $p = 0.3$, and Speech Style: Lombard * Speaker Nativeness: Native $\beta = -0.1$, $z = -0.8$, $p = 0.4$). Additionally, the two-way interaction between Listener Group and HEGPs was not significant and therefore removed from the model (Listener Group: Dutch * HEGPs $\beta = 0.0$, $z = -0.2$, $p = 0.9$ and Listener Group: Spanish * HEGPs $\beta = 0.0$, $z = -0.9$, $p = 0.4$). The control predictor Focus was again also not significant and removed from the model ($\beta = 0.1$, $z = 0.3$, $p = 0.7$).

The final statistical model is shown in Table 3. The AIC score of the model with HEGPs (25,683.6, $df = 27$) was substantially lower than that of the model without the HEGPs as predictor (25,925.0, $df = 26$), which shows that the model with HEGPs better explains the data despite the extra degree of freedom.

The HEGPs contribute to explaining the variance in the data. The higher the HEGPs, the better intelligibility (main effect of HEGPs). The effect size of HEGPs is similar to the effect size of Speech Style.

Table 3
The glmer model of intelligibility of native and non-native plain and Lombard speech by native and non-native listeners including HEGPs as predictor. Plain speech, non-native speaker, and native listener are on the intercept. After each predictor, we indicate which level is being contrasted with the intercept. The β indicates the size of the difference between the intercept level and the contrasted level.

Fixed effects	β	z	p
(Intercept)	-1.2	-4.2	<0.001
Speaker Nativeness: Native	1.5	4.8	<0.001
Listener Group: Dutch	0.0	-0.2	0.8
Listener Group: Spanish	-0.2	-1.9	0.1
Speech Style: Lombard	0.6	3.9	<0.001
HEGPs	0.5	15.7	<0.001
Trial Number	0.1	4.8	<0.001
Occurrence: Second	0.4	9.1	<0.001
Final: Non-final	-0.6	-2.7	<0.01
Speaker Nativeness: Native* Listener Group: Dutch	-0.9	-11.6	<0.001
Speaker Nativeness: Native* Listener Group: Spanish	-0.9	-10.6	<0.001
Listener Group: Dutch * Speech Style: Lombard	0.2	2.8	<0.01
Listener Group: Spanish * Speech Style: Lombard	0.2	3.0	<0.01
Random effects			*SD*
Listener (Intercept)			0.3
Target Word (Intercept)			1.4
Speech Style: Lombard by Target Word			1.2
Listener Group: Dutch by Target Word			0.4
Listener Group: Spanish by Target Word			0.7
Speaker (Intercept)			0.6
Speech Style: Lombard by Speaker			0.3

The inclusion of the HEGPs changed the results for the non-native speech. In contrast to the previous model, the model with the HEGPs does not show a difference between either the Dutch or the Spanish listeners with the Canadian listeners in comprehending non-native plain speech. The Dutch and the Spanish listeners also did not differ from each other, as evidenced by a model with the Dutch listeners on the intercept (Listener Group: Spanish $\beta = -0.2$, $z = -1.9$, $p = 0.1$).

The HEGPs model also showed different results for the native speech. Native speech was only better understood than non-native speech (simple effect of Speaker Nativeness) by the Canadian listeners. The interactions of Speaker Nativeness and both Listener Groups showed that this was less for the Dutch and Spanish listeners, and when we releveled the model, the simple effect of Speaker Nativeness was not significant (Speaker Nativeness: Native $\beta = -0.6$, $z = -1.8$, $p = 0.1$) and we did not find a difference of the effect of the nativeness of the speech between the Dutch and Spanish listeners (Speaker Nativeness: Native * Listener Group: Spanish $\beta = 0.1$, $z = 0.8$, $p = 0.4$). This indicates that if the listener is Dutch or Spanish, the benefit of the native speech can be explained by the difference in HEGPs, while the Canadian listeners benefit from the native speech even when HEGP differences are accounted for.

The inclusion of the HEGPs did not affect the pattern of results for the difference between plain and Lombard speech. Lombard speech was better understood than plain speech (simple effect of Speech Style), and this effect was larger for the Dutch and Spanish listeners (interaction of Listener Group and Speech Style) compared to Canadian listeners. The Dutch and the Spanish benefitted similarly from the Lombard speech, as indicated by the releveled model, in which the interaction of Speech Style and Spanish Listener Group was not significant (Listener Group: Spanish * Speech Style: Lombard $\beta = 0.0$, $z = 0.3$, $p = 0.8$).

The control predictors remained the same as in the model in Section 4.3 (Table 2), with significant effects of Final, Trial Number, and Occurrence. Furthermore, the random structure also remained the same as in the previous model.

5. Discussion

This article compares the size of the Lombard intelligibility benefit for words in noise between native (American-English) and non-native (native Dutch) English speakers, when heard by native and non-native listeners. Perception of non-native Lombard speech has not previously been investigated. This study sheds light on the nature of non-native Lombard speech while also touching upon non-native speech perception.

By computing high-energy glimpsing proportions (HEGPs), we first gained information about the speech signal itself and the stimuli's capacity to withstand noise masking. HEGPs only consider the acoustics of the speech signal and are therefore language independent and can serve as an objective measure. Previous studies have shown that a large part of the capacity to withstand masking can be explained by the shifts in the spectral energy distribution of speech (e.g., Lu and Cooke, 2009). HEGPs only consider the quantity of audible information and not the quality. Therefore they do not take other factors that may be relevant for speech intelligibility into account, such as coarticulation and changes in vowel formants which may occur with reduction.

The native and the non-native Lombard speech showed similar increases in HEGP scores when going from plain to Lombard speech. This suggests that both native and non-native Lombard speech are more resistant to noise than plain speech. The finding for native speech is in line with previous research using glimpsing proportions (GPs), which is the basis of the HEGPs that we used (e.g., Lu and Cooke, 2009). Importantly, the current study extends this finding to Lombard speech produced by non-native speakers. This suggests that beneficial alterations to the spectral energy distribution are also present in non-natively produced Lombard speech.

The intelligibility of the same speech materials was tested with human listeners. In addition to a native listener group (Canadians), we tested two non-native listener groups; one that shared the native language with the non-native English speakers (native Dutch) and one that did not (native Spanish). All listener groups showed a clear Lombard benefit for both the native and the non-native speech. Together, the HEGP analysis and the intelligibility experiment therefore show that, like native speakers, non-native speakers produce Lombard speech and that their Lombard speech is more intelligible in noise than their plain speech.

While our finding that native Lombard speech is more intelligible than plain speech when present in noise is in line with previous research investigating the Lombard benefit with native speakers (e.g., Dreher and O'Neill, 1957; Pittman and Wiley, 2001; Van Summers et al., 1988), to our knowledge, the Lombard intelligibility benefit with non-native speech is a novel finding. It could indicate that the production of Lombard speech is easily acquired by language learners or that it results from mechanisms that learners transfer from their native language to non-native languages. As learners are typically not explicitly taught about Lombard speech and previous research showed small differences in the median F0 and F0 range between the native and non-native English Lombard speech presented in our experiment (Marcoux and Ernestus, 2019a, 2019b), we believe that the latter explanation is more likely. This indicates that learners may implement their native Lombard speech alternations in non-native languages, but that this does not greatly hinder the Lombard benefit for the listeners.

The native listeners exhibited a smaller Lombard benefit than the non-native listeners. The increased intelligibility of Lombard speech compared to plain speech across native and non-native speech was 13.4% points for the native listeners, while, for the non-native listeners, this was greater at 16.5% points. It should be noted that although the non-native listeners may have had more problems than the natives with correctly spelling the target words (we counted every misspelling as an incorrect answer), this should not have modulated the size of the Lombard benefit, as we may expect the same spelling problems for both the Lombard and the plain speech.

Our finding that non-native listeners showed a larger Lombard benefit than native listeners contrasts with the findings reported by Cooke and García Lecumberri (2012), who documented a smaller Lombard benefit for non-native listeners compared to the same materials with native listeners (Lu and Cooke, 2008). Lu and Cooke (2008) reported native listeners identifying plain speech embedded in noise at 42% accuracy, with a 22% point increase in intelligibility for identifying Lombard speech produced in 82 dB SPL embedded in noise. For non-native listeners, this was a baseline of 36% accuracy and a 15% point increase in intelligibility for Lombard speech (Cooke and García Lecumberri, 2012). This difference with our findings could be explained in several ways, including differences in stimuli and masker properties. The listeners tested in Cooke and García Lecumberri (2012) and in Lu and Cooke (2008) identified letter and number combinations at SNR -9 dB. In the current study, listeners were asked to identify English words at -1 SNR. For our stimuli, we used 96 different target words, which did not belong to any one category or topic. Our stimuli were produced in sentences where many words, including the target words, were reduced (e.g. *police* was often pronounced as /pli:s/). We spliced the target words out of their sentences and presented them in isolation, which makes especially reduced words hard to understand (e.g., Ernestus et al., 2002). The difference in stimuli and masker between our study and Cooke and García Lecumberri (2012) and Lu and Cooke's (2008) could in part explain the difference in the size of the Lombard benefit.

We also analyzed the intelligibility scores including HEGPs as a predictor. The predictor HEGP was statistically significant, suggesting that part of the Lombard benefit can be explained by a shift in energy to higher frequency regions. Since Speech Style was still significant as well, these analyses show that the HEGPs predict part of the Lombard benefit while other acoustic characteristics of Lombard speech not captured by HEGPs also contribute to the Lombard benefit. Acoustic characteristics of Lombard speech that are not considered in HEGPs include increase in duration (e.g., Castellanos et al., 1996; Dreher and O'Neill, 1957; Garnier and Henrich, 2014; Junqua, 1993; Van Summers et al., 1988) and shifts in the vowel space (e.g., Garnier, 2008). The inclusion of the HEGPs as predictor did not affect any of the interactions involving Speech Style.

In addition to observing the Lombard benefit in both native and non-native speech for all listener groups, we obtained results further elucidating the differences between native and non-native speech and between native and non-native listening. The Spanish listeners performed worse than the Canadian native and Dutch non-native listeners when listening to non-native plain speech. Dutch listeners may outperform Spanish listeners because of their greater exposure to English in daily life. Perhaps more interestingly, when we compared the model without the predictor HEGPs with a model with the predictor HEGPs, the latter model showed a better fit with the data and showed that the Spanish performed as well as the Dutch and native listeners for non-native plain speech. This difference between the two models in whether the Spanish listeners performed as well as the other two groups suggests that the Spanish listeners are differently affected than the native and Dutch listeners by the energetic masking as indicated by the HEGPs. As this experiment was not designed to test for possible differences among groups in sensitivity to HEGPS, future research should further investigate this possible difference.

All listeners benefitted from listening to native rather than non-native speech. This is not in line with the matched and mismatched interlanguage speech intelligibility benefit, as the Dutch and Spanish listeners did not find the non-native English (native Dutch) speech as intelligible as the native speech (Bent and Bradlow, 2003). The higher intelligibility for native speech mimics the difference in HEGPs between the native and the non-native speech indicating that native English speech better withstood masking. Unfortunately, we cannot establish whether this difference in HEGPs is due to acoustic properties inherent to native versus non-native speech or whether it is due to differences between English and Dutch (such as the different realizations of

fricatives) with the Dutch characteristics surfacing in non-native English produced by the native Dutch speakers. This finding requires further investigation.

The Canadian native listeners benefited more from listening to native speech compared to the Spanish and the Dutch. Stated differently, the Canadian listeners suffered more from listening to non-native speech than the Dutch and the Spanish listeners. When the HEGP metric was incorporated in the analysis, the difference between the native listeners on the one hand and the non-native listeners on the other hand was larger, with the Dutch and Spanish listeners no longer showing a benefit for the native speech. Therefore, when the HEGPs are taken into consideration, the Dutch and Spanish listeners do show a matched and mismatched interlanguage speech intelligibility benefit respectively (Bent and Bradlow, 2003). This suggests again that different listener groups may benefit differently from the energetic masking as indicated by the HEGPs.

The materials in this study were restricted to native English and non-native English produced by native speakers of Dutch and we only tested native listeners of English, Dutch, and Spanish. The choice of these languages and these listeners may have affected the results: different results may have been obtained had we chosen to study speakers of native languages that are more dissimilar than English and Dutch, and listeners of native languages that are more dissimilar than English, Dutch, and Spanish. Considering there are various factors that influence the matched and mismatched interlanguage speech intelligibility benefit (Bent and Bradlow, 2003), including the language choices of the native and non-native speakers, the proficiency of the non-native speakers (e.g., Stibbard and Lee, 2006), and the proficiency of the non-native listeners (e.g., Imai et al., 2005; Pinet et al., 2011), we leave it to future research to investigate to what extent the results obtained in our study generalize to other languages and listener groups. Especially, the acoustic characteristics not captured by HEGPs may differ more among languages that are less similar to each other than Dutch and English are. If so, the Lombard benefit may depend on the combination of the speaker's and the listener's native languages.

In this study, we set out to examine the size of the Lombard intelligibility benefit for native and non-native speech. We approached this by analyzing HEGPs to understand the speech signal itself and by conducting an intelligibility experiment with native and non-native listener groups to also understand the role of the listener's native language. We found that, like native speakers, non-native speakers can produce Lombard speech with higher HEGPs than plain speech and that is clearly beneficial for the listener. Although there may be influences from the speaker's native language on non-native Lombard speech (Marcoux and Ernestus, 2019a, 2019b), non-native speech can still show a Lombard intelligibility benefit.

Funding

This project has received funding from the European Union's Horizon 2020 research innovation programme under the Marie Skłodowska-Curie grant agreement No. 675324.

CRediT authorship contribution statement

Katherine Marcoux: Conceptualization, Methodology, Formal analysis, Investigation, Writing – original draft. **Martin Cooke:** Software, Writing – review & editing. **Benjamin V. Tucker:** Writing – review & editing. **Mirjam Ernestus:** Conceptualization, Methodology, Writing – review & editing, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

We would like to thank Dr. Louis ten Bosch for his help in consultation as well as in the creation of the speech shaped noise file and the script to mix the noise with the speech to create the stimuli used in the intelligibility experiment. Additionally, we would like to thank Dr. Esther Janse for her comments and feedback in the creation of this project and for her input on the article.

Appendix 1. List of the target words

Baby	Flower	Neighborhood	Sundays
Balloon	Food	Night	Table
Banana	Force	Notebook	Teacher
Beach	Foundation	Pants	Theater
Birthday	French	Parade	Theme
Board	Fridays	Party	Theology
Bonfire	Game	Pizza	Theory
Boy	Garden	Police	Therapist
Cadaver	Gloves	Professor	Thriller
Café	Gorilla	Rain	Throne
City	Guests	Road	Time
Classes	House	Room	Today
Club	Jazz	Salami	Tomato
Computer	July	Sample	Tomorrow
Conference	Left	Sandwiches	Town
Counselor	Lemonade	Saturday	Tubes
Dance	Letter	Snack	Walk
Day	Likes	Spanish	Watch
Desk	Literature	Spring	Week
Detail	Meeting	Square	Wild
Dinner	Minutes	Store	Woman
Drink	Month	Street	Wood
Education	Morning	Summer	Year
Fall	Move	Sun	Zoo

References

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. J. Stat. Softw. 67, 1–48. https://doi.org/10.18637/jss.v067.i01.

Bent, T., Bradlow, A.R., 2003. The interlanguage speech intelligibility benefit. J. Acoust. Soc. Am. 114, 1600–1610. https://doi.org/10.1121/1.1603234.

Bosker, H.R., Cooke, M., 2018. Talkers produce more pronounced amplitude modulations when speaking in noise. J. Acoust. Soc. Am. 143, EL121–EL126. https://doi.org/10.1121/1.5024404.

Bosker, H.R., Cooke, M., 2020. Enhanced amplitude modulations contribute to the Lombard intelligibility benefit: evidence from the Nijmegen Corpus of Lombard Speech. J. Acoust. Soc. Am. 147, 721–730. https://doi.org/10.1121/10.0000646.

Brysbaert, M., New, B., 2009. Moving beyond Kučera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. Behav. Res. Methods 41, 977–990. https://doi.org/10.3758/BRM.41.4.977.

Cai, X., Yin, Y., Zhang, Q., 2020. A cross-language study on feedforward and feedback control of voice intensity in Chinese–English bilinguals. Appl. Psycholinguist. 41, 771–795. https://doi.org/10.1017/S0142716420000223.

Castellanos, A., Benedí, J.M., Casacuberta, F., 1996. An analysis of general acoustic-phonetic features for Spanish speech produced with the Lombard effect. Speech Commun. 20, 23–35. https://doi.org/10.1016/S0167-6393(96)00042-8.

Cooke, M., 2006. A glimpsing model of speech perception in noise. J. Acoust. Soc. Am. 119, 1562–1573. https://doi.org/10.1121/1.2166600.

Cooke, M., García Lecumberri, M.L., 2012. The intelligibility of Lombard speech for non-native listeners. J. Acoust. Soc. Am. 132, 1120–1129. https://doi.org/10.1121/1.4732062.

Cooke, M., King, S., Garnier, M., Aubanel, V., 2014. The listening talker: A review of human and algorithmic context-induced modifications of speech. Comput. Speech Lang. 28, 543–571. https://doi.org/10.1016/j.csl.2013.08.003.

Council of Europe, 2001. Common European Framework of Reference for Languages: Learning, teaching, assessment. Press Syndicate of the University of Cambridge.

Dreher, J.J., O'Neill, J., 1957. Effects of ambient noise on speaker intelligibility for words and phrases. J. Acoust. Soc. Am. 29, 1320–1323. https://doi.org/10.1121/1.1908780.

Ernestus, M., Baayen, H., Schreuder, R., 2002. The recognition of reduced word forms. Brain Lang. 81, 162–173. https://doi.org/10.1006/brln.2001.2514.

Fox, J., Weisberg, S., 2019. An R Companion to Applied Regression, third ed. Sage Publications, Thousand Oaks CA. ed.

García Lecumberri, M.L., Cooke, M., Cutler, A., 2010. Non-native speech perception in adverse conditions: a review. Speech Commun. 52, 864–886. https://doi.org/10.1016/J.SPECOM.2010.08.014.

Garnier, M., 2008. May speech modifications in noise contribute to enhance audio-visible cues to segment perception? In: Göcke, R., Lucey, P., Lucey, S. (Eds.), International Conference on Auditory-Visual Speech Processing 2008. International Speech Communication Association Archive. Moretown Island, pp. 95–100.

Garnier, M., Henrich, N., 2014. Speaking in noise: how does the Lombard effect improve acoustic contrasts between speech and ambient noise? Comput. Speech Lang. 28, 580–597. https://doi.org/10.1016/J.CSL.2013.07.005.

Imai, S., Walley, A.C., Flege, J.E., 2005. Lexical frequency and neighborhood density effects on the recognition of native and Spanish-accented words by native English and Spanish listeners. J. Acoust. Soc. Am. 117, 896–907. https://doi.org/10.1121/1.1823291.

Jaeger, T.F., 2008. Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. J. Mem. Lang. 59, 438–446. https://doi.org/10.1016/j.jml.2007.11.007.

Junqua, J., 1993. The Lombard reflex and its role on human listeners and automatic speech recognizers. J. Acoust. Soc. Am. 93, 510–524. https://doi.org/10.1121/1.405631.

Lemhöfer, K., Broersma, M., 2012. Introducing LexTALE: a quick and valid lexical test for advanced learners of English. Behav. Res. Methods 44, 325–343. https://doi.org/10.3758/s13428-011-0146-0.

Lombard, E., 1911. Le signe de l'elevation de la voix (the sign of the elevation of the voice). Ann. Mal. L'Oreille du Larynx 37, 101–119.

Lu, Y., Cooke, M., 2008. Speech production modifications produced by competing talkers, babble, and stationary noise. J. Acoust. Soc. Am. 124, 3261–3275. https://doi.org/10.1121/1.2990705.

Lu, Y., Cooke, M., 2009. The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise. Speech Commun. 51, 1253–1262. https://doi.org/10.1016/j.specom.2009.07.002.

Major, R.C., Fitzmaurice, S.F., Bunta, F., Balasubramanian, C., 2002. The effects of nonnative accents on listening comprehension: implications for ESL assessment. TESOL Q. 36, 173–190. https://doi.org/10.2307/3588329.

Marcoux, K., Ernestus, M., 2019a. Differences between native and non-native Lombard speech in terms of pitch range. In: Ochmann, M., Vorländer, M., Fels, J. (Eds.), Proceedings of the ICA 2019 and EAA Euroregio. 23rd International Congress on Acoustics, Integrating 4th EAA Euroregio 2019. Deutsche Gesellschaft für Akustik. Berlin, Germany, pp. 5713–5720. https://doi.org/10.18154/RWTH-CONV-239240.

Marcoux, K., Ernestus, M., 2019b. Pitch in native and non-native Lombard speech. In: Calhoun, S., Escudero, P., Tabain, M., Warren, P. (Eds.), Proceedings of the 19th International Congress of Phonetic Sciences. Canberra, Australia. Australasian Speech Science and Technology Association Inc., Melbourne, Australia, pp. 2605–2609.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., Sonderegger, M., 2017. Montreal forced aligner: trainable text-speech alignment using Kaldi. In: Lacerda, F., House, D., Heldner, M., Gustafson, J., Strömbergsson, S., Włodarczak, M. (Eds.), Interspeech 2017. International Speech Communication Association. Stockholm, pp. 498–502. https://doi.org/10.21437/interspeech.2017-1386.

Mok, P., Li, X., Luo, J., Li, G., 2018. L1 and L2 phonetic reduction in quiet and noisy environments. In: 9th International Conference on Speech Prosody 2018, pp. 848–852. https://doi.org/10.21437/SpeechProsody.2018-171.

Pinet, M., Iverson, P., Huckvale, M., 2011. Second-language experience and speech-in-noise recognition: effects of talker–listener accent similarity. J. Acoust. Soc. Am. 130, 1653–1662. https://doi.org/10.1121/1.3613698.

Pisoni, D., Bernacki, R., Nusbaum, H., Yuchtman, M., 1985. Some acoustic-phonetic correlates of speech produced in noise. In: ICASSP '85. IEEE International Conference on Acoustics, Speech, and Signal Processing, 10. Institute of Electrical and Electronics Engineers, pp. 1581–1584. https://doi.org/10.1109/icassp.1985.1168217.

Pittman, A.L., Wiley, T.L., 2001. Recognition of speech produced in noise. J. Speech Lang. Hear. Res. 44, 487–496. https://doi.org/10.1044/1092-4388(2001/038.

R Core Team, 2016. R: a language and environment for statistical computing.

Scharenborg, O., van Os, M., 2019. Why listening in background noise is harder in a non-native language than in a native language: a review. Speech Commun. 108, 53–64. https://doi.org/10.1016/j.specom.2019.03.001.

Stibbard, R.M., Lee, J.I., 2006. Evidence against the mismatched interlanguage speech intelligibility benefit hypothesis. J. Acoust. Soc. Am. 120, 433–442. https://doi.org/10.1121/1.2203595.

Tang, Y., Cooke, M., 2016. Glimpse-based metrics for predicting speech intelligibility in additive noise conditions. In: Morgan, N., Georgiou, P., Narayanan, S., Metze, F. (Eds.), Interspeech 2016. International Speech Communication Association. San Francisco, pp. 2488–2492. https://doi.org/10.21437/Interspeech.2016-14.

Van Summers, W., Pisoni, D.B., Bernacki, R.H., Pedlow, R.I., Stokes, M.A., 1988. Effects of noise on speech production: acoustic and perceptual analyses. J. Acoust. Soc. Am. 84, 917–928. https://doi.org/10.1121/1.396660.

Villegas, J., Perkins, J., Wilson, I., 2021. Effects of task and language nativeness on the Lombard effect and on its onset and offset timing. J. Acoust. Soc. Am. 149, 1855–1865. https://doi.org/10.1121/10.0003772.

Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York.

Xu, Y., 2011. Post-focus compression: cross-linguistic distribution and historical origin. In: Lee, W.S., Zee, E. (Eds.), Proceedings of the 17th International Congress of Phonetic Sciences. Hong Kong. City University of Hong Kong, pp. 152–155.