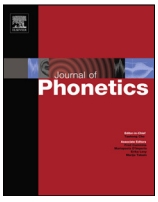
Journal of Phonetics 102 (2024) 101281

Contents lists available at ScienceDirect

Journal of Phonetics

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



Acoustic characteristics of non-native Lombard speech in the DELNN corpus

Katherine Marcoux *, Mirjam Ernestus

Centre for Language Studies, Radboud University, P.O. Box 9103, Nijmegen HD 6500, the Netherlands

ARTICLE INFO

Article history:
Received 22 April 2022
Received in revised form 22 September 2023
Accepted 24 November 2023
Available online 19 December 2023

Keywords:
Lombard speech
Non-native speakers
Native speakers
New speech corpus
Acoustics

ABSTRACT

Lombard speech, speech produced in noise, has been extensively studied in native speakers, while non-native Lombard speech research is limited. This article presents the first corpus of non-native Lombard speech, the Dutch English Lombard Native Non-Native corpus, which includes plain and Lombard read speech from native American-English, non-native English (native Dutch), and native Dutch women. The location of contrastive focus is systematically varied in the sentences. We investigated how intensity, spectral center of gravity, word duration, and VOT varies in the corpus as a function of plain versus Lombard speech and whether it is modulated by the speaker's nativeness and of the language. We did not find differences in how the native and non-native English speakers adapted their English speech in noise, indicating that the Dutch non-native speakers produced Lombard speech similarly to the native English. The comparison of the native Dutch and non-native English sentences produced by the same participants nevertheless suggests that, for all acoustic measurements except word duration, the Dutch speakers adapted their Lombard speech differently in native Dutch than in non-native English. Combined, this would indicate that, when speaking English, Dutch speakers adapt their way of speaking in noise to the way native English speakers do.

1. Introduction

In noisy environments, such as train stations, canteens, and cafes, our way of speaking changes, resulting in what is known as Lombard speech (Lombard, 1911). The specific acoustic modifications that occur when going from a quiet environment, where one produces plain speech, to a noisy environment, where one produces Lombard speech, have been thoroughly studied in native speakers. In contrast, there has been relatively little research dedicated to non-native speakers in noise. This article presents the first corpus of non-native Lombard speech and presents a study based on this corpus that examines non-native Lombard speech acoustics and the potential influence of the native language.

Lombard speech is characterized by changes in acoustics compared to plain speech. These include but are not limited to an increase in fundamental frequency (f0), a widening of the f0 range, an increase in intensity, a shift in energy to higher frequency regions, and changes in duration (for a review see: e.g., Cooke, King, Garnier, & Aubanel, 2014). The extensive research on Lombard speech acoustics has been conducted in several languages including English (e.g., Lu & Cooke, 2008; Pisoni, Bernacki, Nusbaum, & Yuchtman, 1985; Van Summers, Pisoni, Bernacki, Pedlow, & Stokes, 1988), Dutch (e.g., Bosker & Cooke, 2020), Spanish (e.g., Castellanos, Benedí, & Casacuberta, 1996), and French (e.g., Garnier & Henrich, 2014). All this research has focused on native speakers of these languages.

If Lombard speech is an automatic reaction to noisy environments, its properties may be universal to all languages. For instance, increase in intensity may be assumed to depend on how much masking can be expected from the background noise on the speech, rather than depending on exactly which language is spoken. If Lombard speech adjustments are universal, one would expect no differences between native and non-native speakers of the same language. However, Zollinger and Brumm (2011) argue that while Lombard speech does elicit an involuntary response, it is "not a true reflex" (p. R614) since it can be modified by the speaker (e.g., producing a stronger effect when communicating than when in a non-communicative setting such as reading a text; e.g., Junqua, Fincke, & Field, 1999; Villegas, Perkins, & Wilson, 2021) and it can be inhibited with training (e.g., Pick, Siegel, Fox, Garber, & Kearney, 1989). Considering the speaker's ability

* Corresponding author.
E-mail address: katherine.marcoux@ru.nl (K. Marcoux).

https://doi.org/10.1016/j.wocn.2023.101281
0095-4470/

to adapt their Lombard speech to the speaking conditions, we may observe more differences in Lombard speech, for instance, between languages.

If Lombard speech has language specific modifications, we would expect to observe differences when examining native and non-native Lombard speech. The native language influences many characteristics of plain speech by non-native speakers, including the quality of vowels, voice onset time (VOT) of consonants, and intonation, (e.g., Burgos, Cucchiarini, van Hout, & Strik, 2013; Flege & Eefting, 1987; van Maastricht, Krahmer, & Swerts, 2016). Based on this, one may expect that non-native Lombard speech acoustics may also be influenced by the native language.

In addition to the influence of the native language, speech by non-native speakers may be affected by the higher cognitive load non-native speakers experience relative to speaking a native language (e.g., Kormos, 2006; Segalowitz, 2010). Of note, cognitive load may be inversely related to the proficiency in the non-native language, with lower proficiency non-native speakers facing a higher cognitive load than high proficiency non-native speakers. Wester, García Lecumberri, and Cooke (2014) analyzed speech from native English and native Spanish speakers speaking in their native and non-native language (Spanish and English, respectively). They found that speech produced by non-native speakers was characterized by certain acoustic characteristics that are also present in hesitant speech, such as a slower speech rate and a smaller f0 range.

To our knowledge, non-native Lombard speech has only been investigated in a handful of studies in limited acoustic measures. Villegas et al. (2021) investigated native Japanese speakers producing native Japanese and non-native English speech. They found that these speakers showed an increase in sound pressure level in Lombard speech compared to plain speech, although the amount of increase depended on the combination of the language they spoke and the task. Cai, Yin, and Zhang (2020) examined Chinese-English late bilinguals, finding that the speakers increased their intensity in noise in both Chinese (first language, L1) and in English (second language, L2), while the amount of increase depended on the language they spoke (L1/L2) in combination with the noise condition. The same researchers further investigated intensity with similar participants, finding that the L2 speakers increased their intensity more than L1 speakers in the two noise conditions (Cai, Yin, & Zhang, 2021). Mok, Li, Luo, and Li (2018) also examined speakers who had Mandarin as their first language and English as their second, examining mean intensity, f0, and duration of vowels. For the L1 Mandarin speech, they found longer durations, higher intensity and higher f0 (for two of the three tones studied) for vowels in noise. For the L2 English speech, they also found longer durations and higher intensity, but unexpectedly, lower f0 in noise than in quiet.

Marcoux and Ernestus (2019a, 2019b) compared plain and Lombard speech as produced by native American-English speakers, by native speakers of Dutch speaking non-native English, and the same native speakers of Dutch speaking native Dutch. The findings suggest that when going from plain to Lombard speech, the non-native speakers increased their median f0 and f0 range, in accordance with past research on Lombard speech by native speakers (f0; e.g., Pisoni et al., 1985; Van Summers et al., 1988; f0 range: e.g., Garnier & Henrich, 2014; Welby, 2006). Marcoux and Ernestus (2019b) also observed an effect of the native language on the non-native language depending on the position of contrastive focus in the sentence. The Dutch non-native English speakers increased their median f0 when the sentence they read had contrastive focus early in the sentence, as they did in their native Dutch, while the native English speakers did not increase their f0 to a significant extent. Furthermore, for the late-focus sentences, these non-native English speakers had a smaller f0 range increase compared to the native English speakers in their Lombard speech, but not as small as in their native Dutch (Marcoux & Ernestus, 2019a). The authors interpreted these results as an indication of a slight influence of the native language on the non-native Lombard speech.

The purpose of this article is two-fold. First, we present the first corpus that we know of with native and non-native Lombard speech. This corpus is freely available for research purposes and will thus facilitate the research on non-native Lombard speech. Second, on the basis of the corpus, we further explore the acoustic properties of non-native Lombard speech. Like the studies above, we investigated potential differences between native and non-native Lombard speech. Additionally, following Marcoux and Ernestus (2019a, 2019b), we examined how non-native Lombard speech relates to the speaker's native language. Finally, also following Marcoux and Ernestus, we investigated the potential influence of the position of focus in the sentence, firstly because it may affect Lombard speech and secondly because position of focus was manipulated in the corpus.

We chose the corpus to focus on non-native English produced by native speakers of Dutch since Lombard speech has been researched in both languages (English: e.g., Bosker & Cooke, 2018; Lu & Cooke, 2008; Pisoni et al., 1985; Van Summers et al., 1988; Dutch: e.g., Bosker & Cooke, 2020). Additionally, Dutch native speakers tend to have high proficiency in English, and are therefore likely to be able to adapt their speech to the environment, making them a good non-native population to study. Finally, Dutch and English are both Germanic languages, and there are many similarities between them (e.g., post focus compression, non-tonal languages), making them easy to compare, while they still differ in many respects.[1] Because Dutch Lombard speech has not been as extensively studied, studies based on the corpus, including the study reported in this article, also add to our knowledge of native Dutch Lombard speech.

Our corpus, the Dutch English Lombard Native Non-Native (DELNN) corpus, consists of both English plain and Lombard speech from native English speakers as well as from native

[1] While Dutch and (American) English are similar in many phonological aspects, there are also many phonological and phonetic differences. Examples include: differences in the phoneme inventory (e.g., there is no velar fricative in English while there is in Dutch (e.g., Johnson & Babel, 2010; Booij, 1999) and there is no /g/ in Dutch-native words (e.g., Booij, 1999)), in English /æ/ and /ɛ/ differ phonetically (e.g., Hillenbrand, Getty, Clark, & Wheeler, 1995) while in Dutch the vowel /ɛ/ falls between the two English realizations (e.g., Collins & Mees, 1996), VOT differs (e.g., Lisker & Abramson, 1964), lengthening of vowels before voiced obstruents, which affects word duration, is more pronounced in English (e.g., House, 1961) than in Dutch (e.g., Elsendoorn, 1985), spectral CoG differences for certain phones (e.g., Quené et al., 2017), final devoicing in Dutch but not in English (e.g., Booij, 1985), and while there are many similarities in intonation between Dutch and Received Pronunciation (RP) English, there are differences which may lead Dutch speakers producing English to sound monotonous to native RP English speakers (e.g., Collins & Mees, 1996).

Dutch speakers, who also produced native Dutch plain and Lombard speech. The 30 native Dutch women and nine native American-English women read contrastive questions-answer pairs, where the location of contrastive focus in the answers was manipulated. We manipulated the location of contrastive focus as we wanted the DELNN corpus to facilitate research into acoustic reduction, and the degree of reduction is sensitive to whether the word is in focus position or not (e.g., van Bergem, 1993). The corpus is available at https://zenodo.org/record/4267819#.Y4nPpXbMJdg to other researchers and is described in detail in later sections.

In our study based on the corpus, we first compared the difference between plain and Lombard speech in native and non-native English (the *English speech* comparison), to see whether the non-native speakers adapt their speech when going from plain to Lombard speech in the same way as the native English speakers do. We then compared the non-native English and native Dutch speech, produced by the same speakers (the *Dutch speakers* comparison). This comparison shows whether the Dutch adapt their speech, when going from plain to Lombard speech, in the same way in their native Dutch as in their non-native English. This second comparison may shed light on how to interpret the results from the English speech comparison. If no differences are found in the English speech comparison, the Dutch speaker comparison will show whether this is because the Dutch speakers have learnt how to produce Lombard speech in English (we then see a difference between native Dutch and non-native English) or because there are minimal to no differences between Dutch and English (we then see no difference between native Dutch and non-native English). If, in contrast, the English speech comparison shows differences between native English and non-native English, the Dutch speaker comparison may show whether these differences result from a transfer of Lombard properties from native Dutch to non-native English.

In order to investigate whether the acoustic characteristics of the native language may affect Lombard speech in a non-native language, the best acoustic characteristics to study would be those that differ between plain and Lombard speech as well as between the native and the non-native language. Unfortunately, although Lombard speech has been researched in both Dutch and English, the focus of none of these studies has been on acoustic characteristics that substantially differ between Dutch and English (i.e. the characteristics mentioned in footnote 1). It is therefore unknown whether acoustic characteristics that clearly distinguish between Dutch and English play a role in the production of Lombard speech. This complicated the choice of the acoustic characteristics for the present study. In order to ensure that we would find differences between plain and Lombard speech in the native speech, we examined three measures that have been consistently reported to differ between plain and Lombard speech and that can easily be extracted from any dataset of transcribed speech (intensity, spectral center of gravity, word duration). In addition, we examined one measure that we know to differ between Dutch and English but for which it is uncertain whether it is modulated by plain versus Lombard speech.

The first characteristic of Lombard speech we investigated is intensity, which has been shown to increase compared to plain speech (e.g., Dreher & O'Neill, 1957; Junqua, 1993; Lu & Cooke, 2008; Pisoni et al., 1985; Van Summers et al., 1988). Considering that intensity and f0 are correlated (e.g., Gramming, Sundberg, Ternström, Leanderson, & Perkins, 1988), and that non-native Lombard English produced by Dutch native speakers may show f0 characteristics from Dutch (Marcoux & Ernestus, 2019b), we may expect differences in intensity between English Lombard speech produced by American-English and Dutch native speakers. These differences may be a function of the location of contrastive focus in the sentence, as they are for f0.

The second characteristic of Lombard speech we investigated is spectral Center of Gravity (CoG) of the utterance. As mentioned, Lombard speech is characterized by a shift in energy to higher frequency ranges (e.g., Pisoni et al., 1985; Van Summers et al., 1988). One way to measure the shift in energy is to examine spectral CoG, which is the average frequency weighted by the amplitudes of the frequencies. Indeed, the spectral CoG of an utterance has been shown to increase in English Lombard speech (e.g., Lu & Cooke, 2008; Lu & Cooke, 2009).

The spectral CoG of an utterance is determined by its phonemes, and some phonemes (like fricatives) tend to have higher spectral CoG than others (e.g. vowels).[2] In addition, the same phoneme may have different spectral CoG depending on the language. Previous research has shown that Dutch and English plain speech differ in spectral CoG for at least some phonemes. For instance, Dutch /s/ has a lower spectral CoG than English /s/ (e.g., Quené, Orr, & van Leeuwen, 2017). Slight differences in pronunciation for other phonemes between Dutch and English suggest that these phonemes may differ in spectral CoG as well between Dutch and English (see e.g. Gussenhoven & Broeders, 1997). We may therefore expect plain Dutch and English to differ in spectral CoG and that non-native English produced by native Dutch speakers can show the signature of their native language with respect to spectral CoG. However it may be the case that there are no differences in spectral CoG at the utterance level between Dutch and English because some phonemes may have higher spectral CoG in one language and others in the other language, canceling out differences at the utterance level. We decided not to analyze spectral CoG by phoneme since we would have had to have many tokens in similar contexts for each phoneme, which the corpus was not designed for.

The third characteristic we examined is word duration. Previous research has indicated that words and sentences are mostly longer in Lombard than in plain speech (e.g., Dreher & O'Neill, 1957; Junqua, 1993; Van Summers et al., 1988). However, one study has found the opposite, documenting shorter sentence durations in Lombard speech, accompanied by shorter silence durations, which the researchers mention may be due to the speaker's urgency because of the noise (Varadarajan & Hansen, 2006). Because the differences in duration between Lombard and plain speech have not yet been investigated in Dutch to our knowledge, we do not know whether there are differences between English and Dutch in this respect, and, we cannot formulate predictions about

[2] It should be noted that spectral CoG at the phone level describes the place of articulation in bursts or fricatives. In contrast, the spectral CoG at the sentence level reflects the richness of the glottal spectrum.

whether non-native Lombard English produced by native Dutch speakers can show the signature of their native language with respect to duration. We focus on words, rather than on smaller units (e.g. stressed versus unstressed syllables or phonemes).

The fourth acoustic measure we investigated is VOT (Voice Onset Time, the time from the release of the stop consonant as marked by a burst to the start of voicing; Lisker & Abramson, 1964). While the previous three acoustic measures have been extensively researched in Lombard speech, and have been shown to be different in plain and Lombard speech, this is not the case for VOT. We examined VOT as Dutch and English plain speech differ in their VOT lengths. Lisker and Abramson (1964) reported average VOTs for English speakers as 58 ms for word initial /p/ and 80 ms for word-initial /k/ and for Dutch speakers, these values were 10 ms for /p/ and 25 ms for /k/. As a consequence, Lombard speech may affect VOT differently in Dutch than in English, which may surface in non-native English produced by native Dutch speakers. However, in examining native Dutch speakers producing English voiceless plosives, Simon and Leuschner (2010) found that trained and untrained post-secondary school Dutch students were producing longer VOTs in English than in Dutch and that these values were in native English speaker ranges. This suggests that the Dutch are able to adapt their voiceless VOTs to native English, at least in plain speech. Regarding VOT and Lombard speech, we may expect a decrease in VOT length because an increase in air flow, which may be expected in Lombard speech, may hinder vocal fold vibration. Alternatively, we may observe a lengthening of VOT as researchers found for /p/ in clear speech which was explained by the reduced speech rate (Hazan, Grynpas, & Baker, 2012).

As mentioned above, we examined whether the effects would be modulated by the position of focus in the sentence. We may expect that focus will be more relevant for some acoustic measures, than others. For example, words that receive focus are longer in duration (e.g., Cooper, Eady, & Mueller, 1985), and contrastive focus has also been shown to affect VOT (e.g., Choi, 2003).

The article serves to introduce the DELNN corpus in detail as well as to analyze several acoustic features of said corpus. Therefore we first describe the corpus in detail. This is followed by a section on the data we extracted from the corpus for the present acoustic study.

2. DELNN corpus

2.1. Speakers

The DELNN corpus includes the recordings from 39 women, allowing us to have a homogenous sample. Men and women differ in some acoustic measures such as f0 and some research has reported slight differing pitch and energy changes in Lombard speech for women and men (e.g., Junqua, 1993). All speakers reported no vision or hearing issues nor dyslexia or stuttering.

Of the 39 speakers in the DELNN corpus, 30 were native speakers of Dutch, with an average (*M*) age of 21.3 years, and nine were native speakers of American-English (*M* = 22.1 years). The Dutch native speakers were students at Radboud University, Nijmegen, The Netherlands (RU) and were completing their studies in Dutch. They all had native Dutch speaking parents and according to their LexTALE scores (*M* = 69.4, standard deviation (*SD*) = 15.8) (Lemhöfer & Broersma, 2012) on average they had a B2 English level proficiency as per the Common European Framework (Council of Europe, 2001), meaning that they are independent users. The American-English speakers were studying at RU at the time of recording and all had been residing in The Netherlands for less than a year and a half (ranging from three to 18 months). They were raised in the United States by at least one native English speaking parent.

2.2. Speech materials

We designed the speech materials such that the corpus would be a valuable database to address a range of questions on the acoustic characteristics of non-native English produced by native speakers of Dutch.

The corpus consists of question–answer pairs in which there is a target word embedded. For the English speech materials, there are three target word categories, chosen because of their difficulty for native Dutch individuals speaking English. Each of the English target word categories consists of 12 target words, resulting in a total of 36 English target words (see Appendix A), with an average of 2.5 syllables per target word (*SD* = 1.0). The corpus thus provides data for research on how these target words are produced in the different conditions incorporated in the corpus. We will not specifically examine these three target word categories in this article. We nevertheless discuss the details of these target word categories since this article serves as an introduction to the DELNN corpus.

The first category of target words consists of /θ/-initial words (e.g., *theater*). The /θ/ is problematic for native Dutch speakers as /θ/ does not exist in their phoneme inventory. They tend to produce other phones in its place, most often /t/ (Hanulikova & Weber, 2010). The second category of target words is English-Dutch cognates with a schwa in prestress position in American-English, which is represented by a <a> or <o> in the orthography, and which corresponds to a full vowel in Dutch (e.g., *parade*). These schwa target words may pose difficulty as the Dutch may tend to produce the schwa as the full vowel that is present in the orthography and in their native Dutch. This may be especially so when the word receives contrastive focus, while the full vowel may be more likely to be correctly produced as schwa in non-accent position, due to vowel reduction in Dutch in this position (e.g., Booij, 1999). The final category consists of target words ending in voiced obstruents (e.g., *club*). These words are difficult for Dutch speakers since Dutch has final devoicing (e.g., Berendsen, 1986; Booij, 1985; Simon, 2010), meaning that voiced obstruents are produced as voiceless in syllable-final positions. Dutch speakers may also apply final devoicing to English words. Hereafter, these three categories are referred to as: /θ/, schwa, and voiced obstruent target words, respectively, as indicative of the problematic phoneme for the native Dutch speakers.

For each target word, four question–answer pairs were created, resulting in a total of 144 English question–answer pairs. Each question had an average of 9.2 words (*SD* = 1.1) and

each answer an average of 9.3 (*SD* = 1.3). As an illustration, the four question–answer pairs for the target word *parade*, belonging to the schwa target word category, are presented below, where the speakers were instructed to emphasize the words in bold.

1. *Did the family go to the **beach** in Barcelona? No, they went to the **parade** in Barcelona.*
2. *Did the **friends** go to the parade in Barcelona? No, the **family** went to the parade in Barcelona.*
3. *Did Lily enjoy the flower **garden** in the spring? No, she enjoyed the flower **parade** in the spring.*
4. *Did **Ellen** enjoy the flower parade in the spring? No, **Lily** enjoyed the flower parade in the spring.*

As can be seen in the examples above, a target word appears in all answers of the question–answer pairs and in half of the questions. The location of contrastive focus was manipulated: half of the question-answers pairs had late-focus (examples 1 and 3), where the contrastive focus was on the target word, and the other half had early-focus (examples 2 and 4), where the target word was in a similar position as in the late-focus sentences, but the contrastive focus was earlier in the sentence, on a different word. This manipulation of the position of focus in the sentence allows for the investigation of how focus may affect the acoustic realization of words, including how focus may manipulate the difference between plain and Lombard speech. The target words were never sentence final.

The Dutch sentences were very similar in structure to the English sentences except that instead of having 36 target words for a total of 144 question–answer pairs, the Dutch had 24 target words resulting in 96 question–answer pairs. Twelve of these target words were the Dutch translations of the English schwa words and for the other 12 target words, nouns were chosen (see Appendix B). The Dutch target words were on average 3.0 syllables in length (*SD* = 0.7). Each question had an average of 7.5 words (*SD* = 0.7), and each answer an average of 8.2 (*SD* = 0.8). Below are the four Dutch question–answer pairs for the Dutch target word *parade*. Early-focus can be found in 5 and 7, and late-focus in 6 and 8, respectively. As mentioned, the late-focus condition indicates that the contrastive focus occurs later in the sentence, on the target word.

5. *Zagen de **jongens** de parade gisteren? Nee, de **studenten** zagen de parade gisteren.*
   *'Did the **guys** see the parade yesterday? No, the **students** saw the parade yesterday.'*
6. Zagen de studenten de **voorstelling** gisteren? Nee, ze zagen de **parade** gisteren.
   *'Did the students see the **performance** yesterday? No, they saw the **parade** yesterday.'*
7. Bezochten de **buren** de parade vanmiddag? Nee, de **kinderen** bezochten de parade vanmiddag.
   *'Did the **neighbors** visit the parade this afternoon? No, the **children** visited the parade this afternoon.'*
8. Bezochten de kinderen de **speeltuin** vanmiddag? Nee, ze bezochten de **parade** vanmiddag.
   *'Did the children visit the **playground** this afternoon? No, they visited the **parade** this afternoon.'*

2.3. Lists

From these question–answer pairs, three main lists were created for each language. Every list contained all question–answer pairs of the given language. The first half of each list was produced as plain speech and the second half as Lombard speech. This led to four blocks, early-focus plain, late-focus plain, early-focus Lombard, and late-focus Lombard. As described above, four question–answer pairs were created per target word, one for each of these blocks. The four question–answer pairs consisted of two matched question–answer pairs, of which one of each pair is early- and one is late-focus, for example (1)–(2) and (3)–(4). In the lists, both members of a pair occurred in either the plain or the Lombard blocks. Apart from this matching criterion, the order of question–answer pairs was randomly permuted.

For each language, for counterbalance purposes, an additional three lists were created from the three main lists. These additional lists had the question–answer pairs that were in Lombard speech as plain speech, and vice versa. Furthermore, the order of the early- and late-focus blocks were flipped within the plain and Lombard conditions. These additional lists adhered to the same criteria as the main lists: plain precedes Lombard and the matching question–answer pairs occurred in the same half. A filler was added as the first item of each block.

This resulted in six lists of 144 English question–answer pairs and four fillers and six lists of 96 Dutch question–answer pairs with four fillers. This formation of lists means that each speaker produced different question–answer pairs in the plain and Lombard conditions.

2.4. Procedure

The question–answer pairs were presented on a desktop computer in Microsoft PowerPoint, in which each question–answer pair was presented on its own slide. The participants were instructed to read both the question and the answer and place emphasis on the words in bold. The task was self-paced and participants could proceed to the next slide using the spacebar. Participants had a break after each block.

All recordings were made at Radboud University, in a sound attenuated room. Participants sat 15 cm from the microphone. The distance from the microphone to the speakers was estimated by the experiment leader. The wheels on the participant's chair were blocked so that the chair would not move during the session.

Participants wore Sennheiser HD 215 MKII DJ headphones throughout the entire experiment. Nothing was played via the headphones apart from the noise in the noise condition to elicit Lombard speech. That is, the participants' own speech was not fed back through their headphones, in either condition, which may have affected their ability to self-monitor their speech. We chose not to feed back the participants' own speech in order to be able to investigate the pure effect of hearing noise on speech planning and articulation, without the potential effect of noise on how well one can hear one's own voice.

The noise played through the headphones in the noise condition was speech shaped noise (SSN), played at 83 dB SPL

(77 dBA). The speech shaped random noise's spectrum was the average spectrum of two minutes of speech recorded from ten women and ten men reading a phonetically balanced text. The SSN file was a single-channel WAV file with a sampling frequency of 44.1 k Hz. The SSN level output was calibrated using the Brüel & Kjær Type 4153 artificial ear.

We used three different microphones rather than one, due to technical issues. Two were the same model, Sennheiser 65, and the other one was a Sennheiser ME 64. Of the 39 participants, all but four participants completed their recordings using one microphone for both sessions. Of the four participants who recorded using two different microphones, three of them used one microphone for one language and another microphone for the other language, which were of the same model for two participants. Of note, all the native English participants recorded using a Sennheiser 65. The Sennheiser ME 64 has a sensitivity of 31 mV/Pa, corresponding to 70.2 dB, while the Sennheiser ME 65′s sensitivity is 10 mV/Pa, corresponding to 80 dB.

The microphone was connected to an AudiTon amplifier (see Appendix C), and in turn to a Roland R-05 WAVE/MP3 solid-state recorder, where the recording was saved as a wav file with a 44.1 kHz sampling rate with 16-bit resolution. The AudiTon was located outside the recording booth so that the researcher could adjust it, without disturbing the participant. The AudiTon amplifier provides a calibration tone, which, in combination with the information about the microphone's sensitivity (the number of millivolts per Pascal), allows researchers to calculate the intensity level of the recording (see Appendix C).

Since the intensity of the participants' voices was expected to vary between plain and Lombard speech, we adjusted the AudiTon before the plain speech block and the Lombard speech block. Therefore, before the first (plain) and third (Lombard) blocks, participants read a short passage so that the AudiTon amplifier could be calibrated to the highest loudness without peaking. Furthermore, each block started with a filler question–answer pair, in case further calibration was needed. If the calibration was not ideal and the speaker was too close to peaking, we calibrated as needed during the session, unbeknownst to the participants.

After recording the English stimuli, the Dutch participants completed the LexTALE task (Lemhöfer & Broersma, 2012), in which they were presented with 60 English words and non-words on the screen and had to indicate for each whether it was a real English word. This task indicates the individual's overall English proficiency level as per the European Common Framework (Council of Europe, 2001). At the end of the first session all participants completed a language background questionnaire. The Dutch participants returned within a week to record the Dutch stimuli, which followed the same recording procedure. We decided to always present the English stimuli in the first session and the Dutch in the second session, as we did not want participants to drop out when they learned that the following session would be in a foreign language. This meant that the session of the language was confounded with the order of the session.

All participants gave informed consent and were compensated upon completion of the recording session(s) with course credit or gift vouchers. The larger project within which this experiment is embedded received ethics approval from the Ethics Assessment Committee Humanities at Radboud University on July 13, 2016, with reference number Let/MvB16U.019446. In total, the English recording session lasted approximately an hour and the Dutch session, about 45 minutes.

2.5. Word and phone level transcriptions

The speech recordings were aligned at the word and phone level using the Montreal Forced Aligner (MFA: McAuliffe, Socolof, Mihuc, Wagner, & Sonderegger, 2017) for the English data and Kaldi (Povey, Ghoshal, Boulianne, Burget, Glembek, & Goel, 2011) for the Dutch data. While two different forced alignment systems were used for the two languages, we believe that this does not pose an issue. The two systems are highly related since MFA (McAuliffe, Socolof, Mihuc, Wagner, & Sonderegger, 2017) uses Kaldi (Povey et al., 2011) as its basis, which is the forced alignment system used for Dutch. Moreover, comparison of the performance of each forced alignment system with human annotators, yielded similar results. This information and further details on the alignment process can be found in Appendix D. The word and phone level transcriptions were used for the acoustic measures; the silences were removed from the sentences before measuring intensity and spectral CoG, word durations were measured on the word level transcriptions, and the VOT measures were annotated by humans who used the transcriptions to orient themselves.

The DELNN corpus includes the speech recordings as well as the Praat TextGrids with orthographic transcriptions and with phone level transcriptions.

3. Methods for the acoustic measures

3.1. Materials

The materials for the acoustic measures only consisted of the answers from the question–answer pairs, not including the filler trials. While listening to the corpus we noticed that the non-native participants especially had difficulties with question–answer pairs that contained the target words *massage*, *thermodynamics*, and *thermometer*. We eliminated answers with these target words for all four acoustic measures. This meant that all acoustic measures were calculated for maximally 132 English stimuli per participant (144 total sentences minus the three difficult target words produced in the four blocks) and an additional 96 Dutch stimuli for the native Dutch speakers. Further, in order to ensure that there was no clipping in the sound files, utterances with series of 1.00 or −1.00, as indicated by Praat (version 6.0.37; Boersma & Weenink, 2018), respectively, were excluded from the intensity analysis. This resulted in the exclusion of another 125 answers for the analyses of intensity. Additionally, for the intensity analysis we excluded the 228 answers for which the calibration tones were missing.

We analyzed word duration and VOT measures of the target words, because only these words informed us about the

role of plain and Lombard speech in combination with the effect of contrastive focus. Because speakers produced another real word instead of the target word in three instances and because of a technical error in the alignment process, we lost a total of four target word tokens for the analysis of word duration.

For the analysis of VOT, we only focused on word-initial /p/ or /k/ target words followed by either a vowel or /r/ or /l/ in English, since these phones allowed for a clear onset of voicing. For Dutch stimuli, the /p/ and /k/ had to be followed by a vowel, and not by an /r/. This was due to the fact that there is great variation in Dutch /r/ pronunciations, which does not provide a consistently clear onset of voicing and is therefore problematic for VOT measurements. This left us with *cab*, *cadaver*, *computer*, *club*, *crib*, *parade*, *police*, *pub*, and *professor* in English. For the Dutch stimuli, we extracted VOT values from *computer*, *kadaver*, *kostuum*, *parade*, and *politie* (in English: *computer*, *cadaver*, *costume*, *parade*, and *police*). We excluded 12 tokens where the voiceless plosive was followed by a voiceless vowel, the vowel was absent, or where there was prevoicing or frication as we were unable to accurately measure the VOT in these tokens.

3.2. Procedure

To calculate mean intensity values (in dB) of the answers, we used Praat's (version 6.0.37; Boersma & Weenink, 2018) command "To Intensity..." with a pitch floor of 100 Hz and an auto time step. In calculating intensity, the "subtract mean" option was not selected, as Direct Current offset was minimal (i.e. silence already corresponded to 0). The dB averaging method was used. We normalized the intensities as recorded by our combination of equipment as described in Appendix C.

We obtained spectral CoG values over the entire utterance by converting the sound file to spectrum using the slow setting in Praat and then computing the center of gravity in the power spectrum, in Hertz (Hz).

The durations (word durations and VOT) were calculated in Praat (version 6.0.37; Boersma & Weenink, 2018) by subtracting the start time from the end time of each token and the values were converted to milliseconds. If the target word was produced multiple times in one utterance, due to repetition by the speaker, we took the first instance, even if this occurred just before a restart.

For the VOT measurements, three annotators, trained by the authors, marked the VOT from the start of the burst to the start of the periodicity (at the zero-crossing boundaries), as is illustrated in Appendix E. In order to see whether VOT duration only varies because of an overall lengthening of the speech materials, corresponding to slower speech rate, we included a durational measure as a control predictor in the VOT analyses. The annotators therefore also marked the end of the vowel (which included an intervening liquid in the case of *club*, *crib* and *professor*) so we could calculate the vowel duration. The end of the vowel was chosen rather than the whole target word as vowels and consonants are lengthened differently in Lombard speech, with vowels being elongated more than consonants (e.g., Castellanos et al., 1996; Garnier & Henrich, 2014; Junqua, 1993). Inter-rater agreement among the annotators is described in Appendix F.

3.3. Analysis

For each acoustic measure, two separate analyses were conducted. One investigated *English speech*, comparing the English data produced by native and the non-native English speakers. The other examined *Dutch speakers*, analyzing the same speakers, producing native Dutch and non-native English data. The two analyses are presented separately in the results sections. Although it is not the focus of our study we also ran another analysis, comparing the native Dutch and the native English speech, just because these data are available. The results can be found in Appendix G.

Our predictors of interest for the intensity, spectral CoG, duration, and VOT analyses for the *English speech* and the *Dutch speakers* analyses were Speech Style (plain, Lombard), Speaker Nativeness (native, non-native) and Focus (at the sentence level, focus indicates early- or late-focus, at the word level focus indicates whether the word received contrastive focus or not). We added Focus because it is an important manipulation in the corpus and because earlier results suggest that Focus may modulate differences between plain and Lombard speech. The crossed-random intercepts were Speaker and Answer for intensity and spectral CoG, and Speaker and Target Word for duration and VOT.

We included scaled and centered Trial Number as well as scaled and centered Occurrence (block number) as control variables. In some cases, there were technical issues and the experimenter asked the participant to redo the affected stimuli, resulting in Trial Numbers higher than 144 and Occurrences higher than four.

The VOT analysis had the following additional control variables: Plosive (p, k), Previous Phone (voiced, voiceless, or silence), Syllable Stress, and Duration (vowel duration). The control predictor Plosive was included since /p/ and /k/ have been shown to have different VOT lengths (e.g., Lisker & Abramson, 1964). Since context may influence VOT (e.g., Yao, 2009), we also considered Previous Phone as a control predictor. Syllable Stress, indicating whether the syllable was stressed, was included as it has been reported to also affect VOT (e.g., Cho & McQueen, 2005; Lisker & Abramson, 1967; Simonet, Casillas, & Díaz, 2014). The control predictor Duration, was included based on the results from Hazan, and colleagues (2012), who found that increased word duration, corresponding to slower speech rate, explained the longer VOT for /p/ they found in clear speech.

We used R (version 3.5.1; R Core Team, 2016) to perform linear mixed effects models from the *lme4* package (version 1.1.21; Bates, Mächler, Bolker, & Walker, 2015), using the Nelder-Mead optimizer as it led to the best convergence. Before beginning the analysis, we first removed outliers, defined as 2.5 standard deviation above or below the grand mean. We then began with a hypothesis-based model, which included interactions among the predictors of interest (Speech Style, Nativeness and Focus) and simple effects for the control variables. If a predictor was not significant and not in a significant interaction ($t < 1.96$), then it was removed from the model. We added *p*-values to the results tables for the benefit of the readers using *lmerTest* (version 3.1.3; Kuznetsova, Brockhoff, & Christensen, 2017). Once everything in the fixed structure was set, we proceeded to the random structure. We

checked whether the addition of random slopes improved the models, using anova(). If an addition did not lead to an improvement, it was not included. If the fitting produced a warning, we did not proceed with that model. Further, if the addition of the random structure resulted in a correlation of 0.75 or higher between the slope and intercept, the slope was removed.[3] We checked the final model to ensure that all fixed predictors included were significant, using the summary() function as well as the Anova() function from the *car* package (version 3.0.6; Fox & Weisberg, 2019).

In the instance that there was a three-way interaction in the final model among the predictors of interest (Speech Style * Nativeness * Focus), as was the case with the intensity data in the *Dutch speaker* analysis, we split by Focus to better understand the data. We produced the split model by starting from the model with the three-way interaction, excluding Focus as predictor, and removed from the resulting model non-significant interactions and simple effects until only significant interactions and simple effects remained (without re-entering simple effects or interactions that were not significant in the model with the three-way interaction).

Plain speech (Speech Style), early-focus or not contrastive focus (Focus), and non-native speaker (Speaker Nativeness) were on the intercept. We chose plain speech and early-focus as we consider them our baseline. In order to compare the effects of predictors between the *English speech* data and the *Dutch speakers*' data, we had non-native speakers on the intercept as the non-native speakers appear in both comparisons. The plots were created using the *ggplot2* (version 3.2.1; Wickham, 2016) package.

4. Results

4.1. Intensity

The intensity data are shown in Fig. 1. The corresponding models comparing native and non-native English (*English speech*) and comparing non-native English and native Dutch (*Dutch speakers*) intensity values are presented in Table 1.

4.1.1. English speech

For fitting the model comparing the native with the non-native English intensity data, we removed 17 outliers, which resulted in 4901 data points being analyzed. The model (see the two columns for *English speech* in Table 1) revealed a significant simple effect of Speech Style which was modulated by Focus. Together these effects indicate that intensity increased significantly for Lombard speech compared to plain speech and that the increase was even larger for sentences with late-focus. Additionally, there was a significant effect of Trial Number, with intensity increasing as the recording session progressed. The random structure showed that the intensity differed per Answer as well as per Speaker (as indicated by the significant random intercepts of Answer and Speaker). Moreover, the effect of Speech Style differed by Answer and by Speaker (as shown by the significant random slopes of Speech Style per Answer and Speech Style per Speaker).

4.1.2. Dutch speakers

The fitting procedure for the model comparing the non-native English and native Dutch speech produced by the same participants implied the removal of 61 outliers resulting in a total of 6411 data points being analyzed. There was a three-way interaction in the model (Speech Style * Nativeness * Focus: $\beta = -0.3$, $t = -3.0$) and we split the data by Focus to better understand the effects of Speech Style and Nativeness (see Table 1). For the early-focus sentences, the model revealed significant simple effects of Speech Style and Trial Number (see the Early-focus columns under *Dutch speakers* in Table 1). Lombard speech had higher intensity than plain speech and, as the recording session progressed, intensity increased as well. For the late-focus sentences, there was a significant effect of Speech Style that was modulated by Nativeness. Lombard speech had higher intensity than plain speech but slightly less so in native Dutch. Trial number was also significant for the late-focus sentences; intensity increased over the recording session.

The random structures for both the early- and late-focus sentences revealed that intensity differed per Answer and per Speaker (as shown by the significant random intercepts of Answer and Speaker). Additionally both the effect of Speech Style and the effect of Nativeness varied per Speaker (as shown by the significant random slopes of Speech Style per Speaker and Nativeness per Speaker).

4.1.3. Intensity interim discussion

Lombard speech had higher intensity than plain speech in all speech, native English, native Dutch, and non-native English. Further, as the experiment progressed, intensity increased. The increase in intensity was observed when participants were producing plain speech as well as when producing Lombard speech. One possible explanation could be that the participants became more confident over the course of the experiment. Alternatively, participants may have become more relaxed over the course of the experiment and may have started leaning towards the microphone. Note, however, that the participants could not move their chairs.

For the *English speech* (native and non-native English data), in addition to the effect of Lombard speech having higher intensity, we observed that the Lombard sentences with late-focus were produced with even higher average intensity than the Lombard early-focus sentences. In English and Dutch, material after the focus undergoes post-focus compression (PFC, English: e.g., Cooper et al., 1985; Xu & Xu, 2005, Dutch e.g., Hanssen, Peters, & Gussenhoven, 2008; Rump & Collier, 1996), with a lowering and narrowing of the f0 range (e.g., Xu, 2011) as well as a lowering of intensity (e.g., Chen, 2015). Considering that intensity decreases after the focus, it is of no surprise that the Lombard late-focus sentences had a larger increase in intensity than the Lombard early focus sentences, as less material underwent PFC, allowing for a larger increase.

When examining *Dutch speakers* producing native Dutch and non-native English speech, we found a different pattern from the *English speech* data. For the late-focus sentences, the speakers produced lower intensity in their native Dutch

[3] This applies to only two models (including the additional models in the appendix). If we allow for the higher correlation in the random structure we get convergence issues for one model and for the other we get the same results for the fixed structure but a more complex random structure.

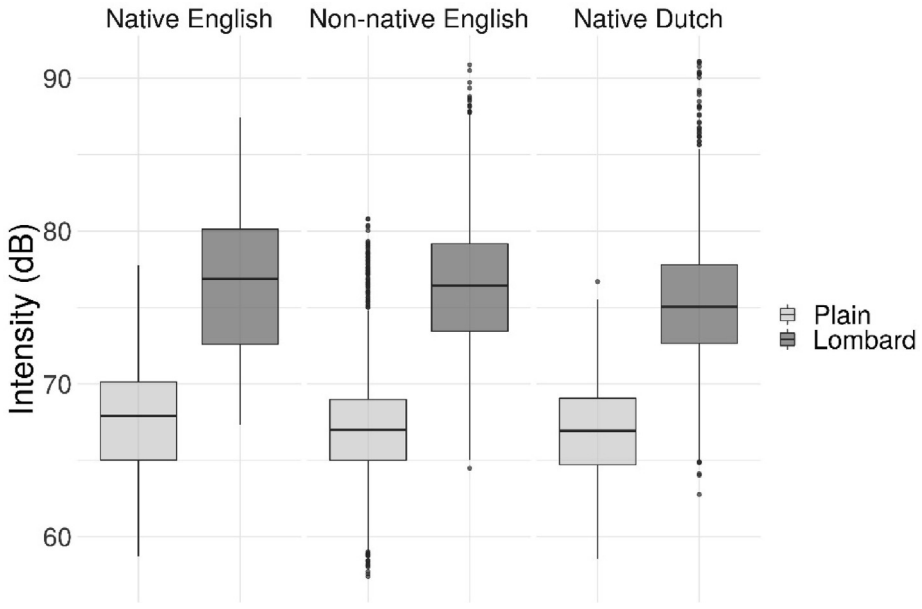


Fig. 1. Average intensity data for native English, non-native English, and native Dutch split by speech style. In this figure and all figures below, a box indicates the upper quartile, median, and lower quartile from top to bottom respectively. The ends of the whiskers indicate the minimum and maximum, respectively, excluding the potential outliers, which are indicated by the dots. The data are visualized before outliers were removed for the statistical analysis. For visualization of the data of all figures with the outliers removed see Appendix H.

Table 1
Lmer models of native and non-native English (English speech) and non-native English and native Dutch (Dutch speakers) intensity split by focus position. In this table and all tables below, β and t-values are rounded to the second decimal point, while *p*-values are rounded to the third decimal point unless they are <0.001.

	English speech			Dutch speakers					
				Early Focus			Late Focus		
Fixed Effects	β	*t*	*p*	β	*t*	*p*	β	*t*	*p*
Intercept	67.42	128.61	<0.001	67.56	128.40	<0.001	67.56	111.23	<0.001
Speech Style: Lombard	8.22	14.87	<0.001	7.13	11.12	<0.001	8.18	13.04	<0.001
Focus: Late	0.27	1.45	0.149	NA	NA	NA	NA	NA	NA
Nativeness: Native	–	–	–	–	–	–	−0.23	−0.49	0.625
Trial Number	0.34	9.04	<0.001	0.83	11.51	<0.001	0.42	4.09	<0.001
Speech Style: Lombard * Focus: Late	0.37	3.79	<0.001	NA	NA	NA	NA	NA	NA
Speech Style: Lombard * Nativeness: Native	–	–	–	–	–	–	−0.32	−2.94	0.003
Random Effects			*SD*			*SD*			*SD*
Answer (Intercept)			1.04			1.00			1.18
Speech Style by Answer			0.40			-			-
Speaker (Intercept)			3.14			3.24			3.18
Speech Style by Speaker			3.34			3.44			3.24
Nativeness by Speaker			-			1.94			2.17
Residual			1.22			1.07			1.15

than in their non-native English in Lombard speech. This is not likely to be due to stimuli differences in the two languages because we do not observe a general effect of nativeness, both in plain and Lombard speech, but rather only in Lombard speech.

Hence, while all speakers increased their intensity when speaking in noise, when speaking Dutch, the Dutch native speakers did less so for late focus-sentences. As the Dutch speakers did not do the same in non-native English, apparently, the Dutch speakers adapted their way of producing Lombard speech when speaking non-native English, similarly to the way native English speakers produce Lombard speech.

4.2. Spectral CoG

The spectral CoG data for all speakers – native English, non-native English, and native Dutch – are visualized in Fig. 2. Table 2 presents the corresponding models comparing native and non-native English (*English speech*) and comparing non-native English and native Dutch (*Dutch speakers*) in spectral CoG.

4.2.1. English speech

The fitting procedure for the model comparing the native with the non-native English data implied the removal of 127 outliers for the analysis of 5020 data points. The columns labeled *English speech* in Table 2 above list the results of the model. The model revealed that the only significant predictor of interest was Speech Style, indicating that spectral CoG values increased for Lombard speech compared to plain speech for both native and non-native English to a similar extent. The random structure revealed that spectral CoG varied per Answer and per Speaker (shown by the significant random intercepts of Answer and of Speaker). Additionally, the

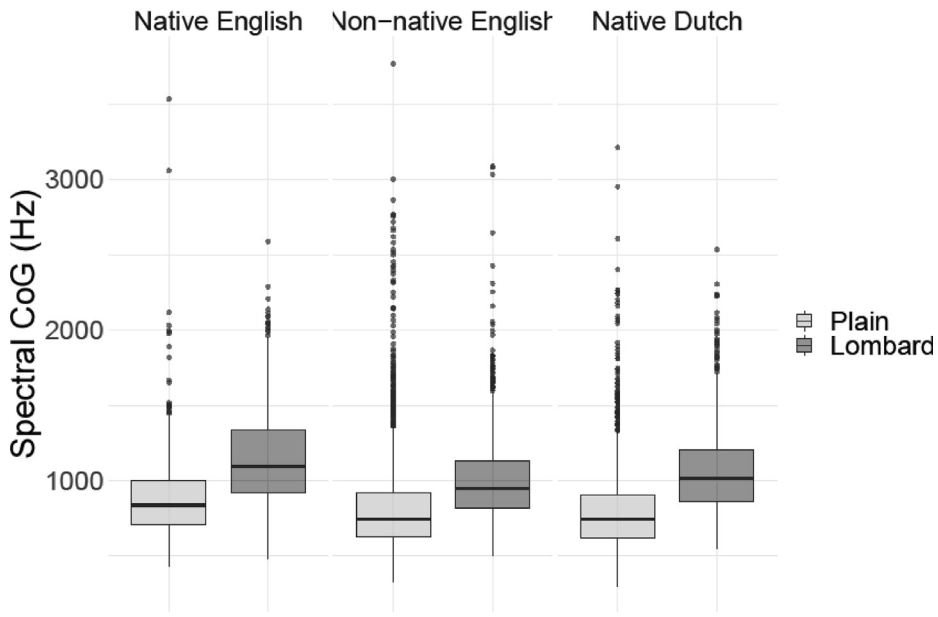


Fig. 2. Spectral CoG data for native English, non-native English, and native Dutch split by speech style.

Table 2
Lmer models of native and non-native English (English speech) and non-native English and native Dutch (Dutch speakers) spectral CoG.

	English speech			Dutch speakers		
Fixed Effects	*β*	*t*	*p*	*β*	*t*	*p*
Intercept	819.93	30.41	<0.001	800.92	29.67	<0.001
Speech Style: Lombard	208.36	8.49	<0.001	190.98	8.31	<0.001
Nativeness: Native	–	–	–	−7.61	−0.36	0.720
Speech Style: Lombard * Nativeness: Native	–	–	–	58.43	5.68	<0.001
Random Effects			*SD*			*SD*
Answer (Intercept)			145.93			153.91
Speech Style by Answer			67.96			56.48
Speaker (Intercept)			147.60			127.14
Speech Style by Speaker			146.92			120.48
Residual			132.55			140.46

effect of Speech Style differed per Answer and per Speaker (significant random slope of Speech Style by Answer and of Speech Style by Speaker).

4.2.2. Dutch speakers

For fitting the model comparing the non-native English and native Dutch speech produced by the same participants, we removed 140 outliers, which resulted in a total of 6700 data points being analyzed. The columns labeled *Dutch speakers* in Table 2 above list the results of the model. The statistical model revealed simple effects of Speech Style as well as an interaction of Speech Style with Nativeness. Together these effects indicate that spectral CoG increased for Lombard speech, and that this increase was larger in Dutch than in English. The random structure is similar to the *English speech* spectral CoG model, with spectral CoG varying per Answer and per Speaker as well as differing for Speech Style per Answer and per Speaker.

4.2.3. Spectral CoG interim discussion

Our analyses showed that spectral CoG was higher in Lombard than in plain speech. The size of the Lombard effect seems similar for the native and non-native speakers of English but to be larger for native Dutch. This suggests that in regards to spectral CoG, the Dutch speakers were doing something slightly different in Lombard speech in their native Dutch than in their non-native English. This suggests that the Dutch native speakers were adapting their spectral CoG to the native English speakers when producing non-native English Lombard speech. Note that, as for intensity, it is unlikely that the difference between Lombard speech in native Dutch and in non-native English can be ascribed to the differences in stimuli or to differences in phoneme inventory. If that were the case, we would have found a difference for plain speech as well.

4.3. Duration

The data for the durations of the target words in milliseconds are visualized in Fig. 3 below. The corresponding models comparing native and non-native English (*English speech*) and comparing non-native English and native Dutch (*Dutch speakers*) target word duration are shown in Table 3. For the former model, we removed 63 outliers leaving 5081 data points, and, for the latter model, we removed 87 outliers leaving 6749 data points to be analyzed.

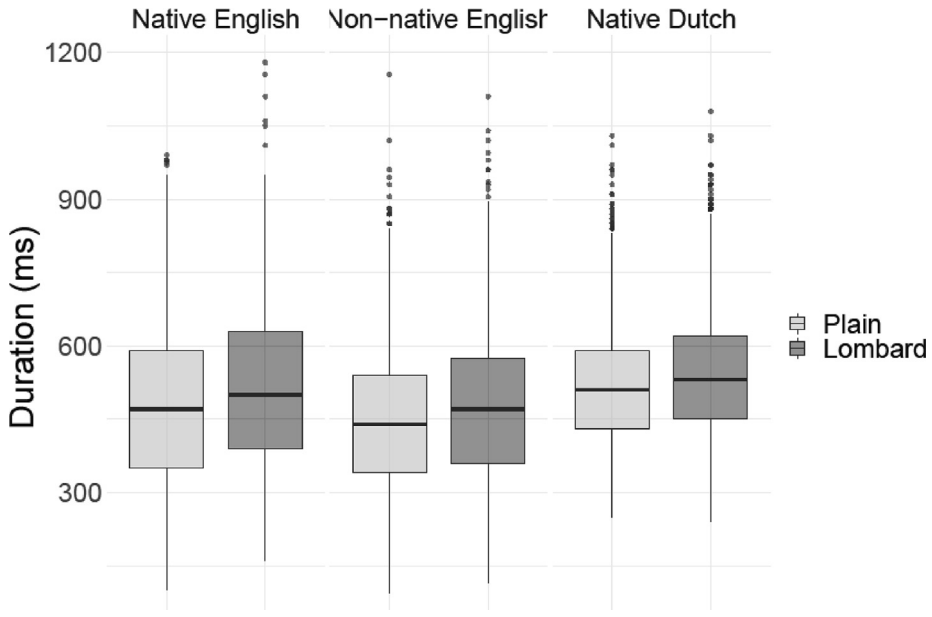


Fig. 3. The durations of target words produced by native English, non-native English, and native Dutch split by speech style.

Table 3
Lmer models of native and non-native English (English speech) and non-native English and native Dutch (Dutch speakers) target word durations.

	English speech			Dutch speakers		
Fixed Effects	*β*	*t*	*p*	*β*	*t*	*p*
(Intercept)	385.61	17.39	<0.001	401.16	20.44	<0.001
Speech Style: Lombard	67.06	13.77	<0.001	34.76	5.40	<0.001
Nativeness: Native	2.95	0.13	0.900	69.69	2.59	0.012
Focus: Contrastive	80.70	39.26	<0.001	77.81	13.55	<0.001
Occurrence	−25.35	−13.41	<0.001	−9.55	−2.92	0.004
Nativeness: Native * Focus: Contrastive	63.96	14.81	<0.001	–	–	–
Speech Style: Lombard * Focus: Contrastive	–	–	–	7.46	2.68	0.007
Random Effects			*SD*			*SD*
Speaker (Intercept)			60.58			46.54
Speech Style by Speaker			19.62			13.04
Focus by Speaker			–			29.55
Target word (Intercept)			109.70			100.05
Nativeness by Target word			26.68			–
Residual			64.29			57.18

4.3.1. English speech

The statistical model for native and non-native English revealed a significant simple effect of Speech Style, indicating that the words were longer in Lombard speech. Additionally, we observe a significant effect of Focus (contrastive focus) and an interaction between Focus and Nativeness. Together this suggests that when a word receives contrastive focus, the word is longer, and that when native English speakers produce it, it is even differentially longer. Finally, we observed a significant effect of Occurrence, indicating that the subsequent occurrences of a target word were shorter. The random structure revealed that duration differed per Speaker and per Target Word (as shown by the significant random intercepts of Speaker and Target Word) and that the effect of Speech Style differed by Speaker while the effect of Nativeness differed by Target Word (as shown by the significant random slopes of Speech Style by Speaker and Nativeness by Target Word).

4.3.2. Dutch speakers

From the model comparing non-native English and native Dutch we observed significant simple effects of Speech Style and Focus, which also interact with each other. The target words were longer in Lombard speech compared to plain speech and longer when carrying contrastive focus. When the target word was produced in Lombard speech with contrastive focus, the duration was even longer. Additionally, we observed significant simple effects of Nativeness and Occurrence, with the target words begin shorter when the speakers produced non-native English, and subsequent occurrences of a target word being shorter. The random structure indicates that word duration varied per Target Word and per Speaker (significant random intercepts of Speaker and Target Word) and that the effects of both Speech Style and Focus varied per Speaker (significant random slopes of Speech Style by Speaker and Focus by Speaker).

4.3.3. Duration interim discussion

The sets of duration data showed several similar patterns. Most importantly for our research questions, the target words produced as Lombard speech were longer than those produced as plain speech. Additionally, subsequent productions of target words were shorter in duration, in line with past research (e.g., Bell, Brenier, Gregory, Girand, & Jurafsky, 2009; Fowler & Housum, 1987). Also in line with previous research, target words with contrastive focus had increased durations (e.g., Cooper et al., 1985).

There were also differences among the word duration datasets. In examining the target word durations in *English speech*, the native English speakers showed a larger effect of focus than the non-native English speakers. The *Dutch speakers* (native Dutch and non-native English speech) showed a larger effect of focus in Lombard than in plain speech, but due to the absence of a three way interaction of Nativeness with Speech Style and Focus for the *English speech* dataset, it is unclear whether the Dutch speakers differ in this respect from the native English speakers. Finally, the dataset of *Dutch speakers* show an effect of nativeness, with Dutch target words having longer durations. This is most likely due to differences between the Dutch and English stimuli, with the Dutch stimuli having more syllables on average per target word (Dutch: M = 3.0, SD = 0.7, English: M = 2.2, SD = 1.0).

4.4. VOT

The VOT data for native English, non-native English, and native Dutch are presented in Fig. 4. The statistical models for native and non-native English (*English speech*) and for the non-native English and native Dutch (*Dutch speakers*) are shown in Table 4 below. For the models, we removed 41 and 50 outliers leaving 1358 and 1618 data points to be analyzed, respectively.

4.4.1. English speech

In analyzing native and non-native English VOT data, we found significant simple effects of Speech Style, Occurrence, and Duration. The VOTs were shorter when produced as Lombard speech compared to plain speech, following occurrences of the target word had shorter VOTs, and as the following segment duration increased, the VOT was longer. Additionally, there were simple effects of Nativeness and Focus, which interacted with each other. The VOT was longer if the speaker was a native English speaker and if the target word received contrastive focus, and it was lengthened further if both were the case. The random structure showed that VOT differed per Speaker and per Target Word (significant intercepts of Speaker and Target Word) and that the effect of Speech Style varied by Speaker and that of Speech Style by Target Word (significant slopes of Speech Style by Speaker and Speech Style by Target Word).

4.4.2. Dutch speakers

From the model on non-native English and native Dutch speech, we see significant simple effects of Speech Style, and Nativeness, as well as an interaction of the two. The simple effects show that, in plain speech (the condition at the intercept), the native Dutch speakers produced shorter VOTs in Dutch than in non-native English and that in non-native English (again the condition at the intercept) they produced shorter VOTs in Lombard speech than in plain speech. The interaction of Speech Style and Nativeness shows that the effect of Speech Style on VOT is smaller in Dutch than in non-native English. In order to observe whether Speech Style affected VOT for native Dutch speech, we releveled the model with native Dutch speakers on the intercept. This releveled model indicated that this is the case (β_{Noise} = −4.29, t = −3.89). The Dutch thus also shortened their VOT when going from plain to Lombard speech in their native Dutch, although to a lesser extent than in non-native English, as indicated by the interaction in Table 4.

We also found an effect of Focus and an interaction of Nativeness with Focus. Table 4, with non-native English on the intercept, shows that contrastive focus in non-native English lengthened VOT. A releveled model, with native Dutch on the intercept, was used to determine whether the effect of Focus was significant for native Dutch speech as well. The releveled model did not reveal a significant effect of Focus for native Dutch speech (β_{Focus} = −1.49, t = −1.60), indicating that the VOTs in Dutch were not longer in contrastive focus than in non-focus position.

Additionally, we see two effects that we did not observe in the *English speech* dataset. The model reveals a significant effect of Plosive, in which the /p/ had a shorter VOT than the /k/. Furthermore, if the plosive was preceded by silence, the VOT was shorter. To investigate whether the absence of these effects in the *English speech* dataset may be a power issue, we took the final model and reduced the *Dutch speakers* dataset to the same size as the *English speech* dataset to see if the effects of Plosive and Previous Phone remained in the smaller sample. We found that the fixed effect for Plosive (β_p = −11.80, t = −2.07) remained significant while Previous Phone ($\beta_{silence}$ = −4.05, t = −1.72, β_{voiced} = −1.44, t = −0.67) was no longer significant.

The random structure revealed that VOT differed per Speaker and per Target Word (significant random intercepts of Speaker and Target Word) and that the effect of Speech Style varied per Speaker (significant random slope of Speech Style by Speaker).

4.4.3. VOT interim discussion

Our VOT data revealed that VOTs were longer in native English than in non-native English and they were even shorter in Dutch. This is in line with past research that /p/ and /k/ have larger VOTs in native English than in native Dutch (e.g., Lisker & Abramson, 1964). Furthermore, the difference between native Dutch and non-native English VOT length indicates that when speaking non-native English, the Dutch are adapting to a certain extent and lengthening their VOT.

In general, VOTs were shortened in Lombard speech. More importantly for our research question, in the *Dutch speakers* dataset, the Dutch speakers were affected differently by Lombard speech in their native Dutch and non-native English, shortening their VOT less in Dutch Lombard speech than in non-native English Lombard speech. This was not the case for the *English speech* dataset, where the effect of Lombard speech was similar for native and non-native English. Combined, this indicates that the non-native English speakers were

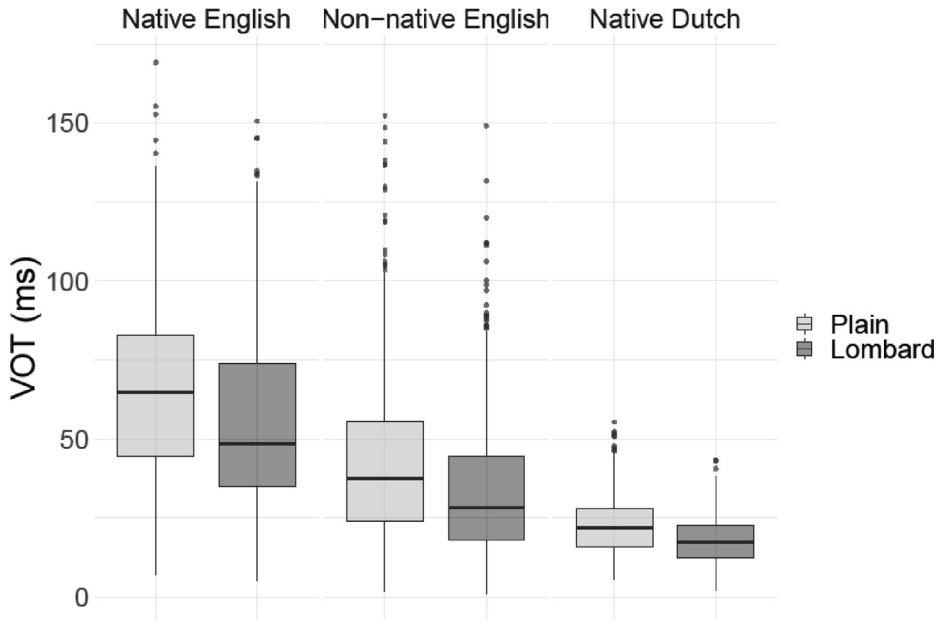


Fig. 4. The VOT of /p/ and /k/ produced by native English, non-native English, and native Dutch split by speech style.

Table 4
Lmer models of native and non-native English (English speech) and non-native English and native Dutch (Dutch speakers) VOT.

	English speech			Dutch speakers		
Fixed Effects	*β*	*t*	*p*	*β*	*t*	*p*
Intercept	35.56	6.62	<0.001	46.62	9.67	<0.001
Speech Style: Lombard	−4.53	−2.06	0.046	−7.75	−8.34	<0.001
Nativeness: Native	12.37	3.66	<0.001	−16.70	−2.88	0.014
Focus: Contrastive	2.91	3.20	0.001	3.35	4.74	<0.001
Plosive: p	–	–	–	−11.98	−2.16	0.054
Previous phone: Silence	–	–	–	−5.09	−2.34	0.020
Previous phone: Voiced	–	–	–	−1.94	−0.97	0.330
Duration	38.80	3.61	<0.001	–	–	–
Occurrence	−2.70	−3.30	<0.001	–	–	–
Speech Style: Lombard * Nativeness: Native	–	–	–	3.46	2.95	0.003
Nativeness: Native * Focus: Contrastive	13.67	7.08	<0.001	−4.83	−4.14	<0.001
Random Effects			*SD*			*SD*
Speaker (Intercept)			9.75			6.35
Speech Style by Speaker			5.09			3.25
Target Word (Intercept)			14.82			10.22
Speech Style by Target Word			3.70			–
Residual			14.56			11.29

making changes to VOT similar to native speakers in their non-native Lombard speech, but unlike what they do in their native Dutch speech.

Additionally, we observed that focus had an effect on the VOT data in English; words with contrastive focus having lengthened VOTs, and this was more so the case for native English speech. In contrast, for native Dutch, the VOT was not affected by focus. In terms of focus, we thus find differences between native Dutch and non-native English on the one hand, while also finding differences between native and non-native English on the other hand.

As for the control variables, for the *English speech* data, duration of the following segment was influential in lengthening VOT, in line with Hazan and colleagues' (2012) research on clear speech. Additionally, in English, VOT was shorter in following occurrences of the word. Surprisingly, we did not find a difference between the /p/ and /k/ plosives in the *English speech* data. For the control variables in the *Dutch speakers* dataset, the VOTs of /p/ were shorter than of /k/, in line with previous research (Lisker & Abramson, 1964) and there was an effect of the previous sound.

5. Discussion

The present article focused on non-native Lombard speech. Non-native Lombard speech may differ from native Lombard speech because non-native speakers have a higher cognitive load when speaking in a non-native language (e.g., Kormos, 2006; Segalowitz, 2010) and this may be even more the case in noise. Moreover, non-native Lombard speech adaptations may show the signature of the speakers' native language. So far, only a few studies have examined non-native Lombard speech, restricting themselves to a small number of acoustic cues that are known to be modified in Lombard speech (inten-

sity, vowel duration, f0 and f0 range; Cai et al., 2020, 2021; Marcoux & Ernestus, 2019a, 2019b; Mok, Li, Luo, & Li, 2018; Villegas et al., 2021).

In order to facilitate the study of non-native Lombard speech, we compiled the Dutch English Lombard Native Non-Native (DELNN) corpus, which is the first corpus to our knowledge to include non-native Lombard speech, in combination with native speech of the two languages involved. The corpus is freely available for research purposes and was designed such that it can be used to address various questions about non-native Lombard speech in future research. The DELNN corpus includes nine native American-English women producing English plain and Lombard speech and 30 native Dutch women producing native Dutch and non-native English plain and Lombard speech. These speakers read 144 English question–answer pairs, half of which were produced as plain speech and the other half as Lombard speech. Furthermore, a target word — words selected for their difficulty for Dutch speakers in English — was embedded in each answer. These target words constituted three categories: words starting with /θ/, a phoneme which does not occur in Dutch, English-Dutch cognates with schwa in prestress position in English and a full vowel in their orthography and in the Dutch counterpart, and words ending in voiced obstruents, which do not occur in Dutch because of final devoicing. Of note, each answer contained contrastive focus. For half of the answers, the target word received contrastive focus (late-focus condition), while for the other half, contrastive focus was earlier in the sentence (early-focus). The native Dutch speakers additionally read 96 Dutch question–answer pairs, also of which half were produced as plain speech and half as Lombard speech, and half with early-focus and half with late-focus.

Using the DELNN corpus, we examined four acoustic measures from the answers in the question–answer pairs: intensity and spectral CoG of the complete answers, durations of the target words, and VOTs of a subset of these target words. We chose these measures because the first three are known to differ substantially between plain and Lombard speech in both languages. The fourth measure we chose in order to also include a measure that differs between Dutch and English, although we did not know whether it would also differ between plain and Lombard speech. For each acoustic measure, there were two comparisons, one examining native and non-native English speakers (*English speech*) and the other examining the same non-native speakers in their non-native English and native Dutch (*Dutch speakers*).

Our analyses revealed that the non-native speakers were producing Lombard speech, adapting all four acoustic measures in noise in the same direction as the native English speakers and as in their native language, increasing intensity, spectral CoG, and word duration, and decreasing VOT compared to plain speech. These adaptations have also been shown in previous studies dedicated to speech by native speakers, for intensity, spectral CoG, and word duration (e.g., Dreher & O'Neill, 1957; Lu & Cooke, 2008; Van Summers et al., 1988). For non-native Lombard speech, research has been done on intensity, vowel duration, f0, and f0 range, but to our knowledge this study is the first to document changes in spectral CoG, word duration, and VOT for non-native Lombard speech. Further, the current article adds to the limited research on native Dutch Lombard speech.

Regarding VOT, our data showed shorter VOTs for the voiceless plosives /p/ and /k/ as compared to plain speech. Further research with more controlled stimuli is needed to better understand the decrease we found in VOT length in Lombard speech.

Our analyses not only showed that the non-native speakers adapted the four acoustic measures in noise in the same direction as the native English speakers, but also that they did so to a similar extent. This shows that non-native speakers need not adapt their speech in noise less or differently than native speakers do just because they are non-native. At least the non-native speakers we investigated in this study (native speakers of Dutch speaking English at a B2 level, as per the European Common Framework; Council of Europe, 2001) adapted their speech as much as native speakers.

Since we did not observe differences between how the native and non-native English speakers adapted their speech in noise for the four measures, the comparison between native Dutch and non-native English Lombard speech may show why. It may show whether a difference is absent because the two languages adapt their speech in a similar manner in noise, or because the Dutch speakers have learned how to successfully adapt their Lombard speech in their non-native language.

When comparing native Dutch and non-native English (from the same speakers), we found differences in their Lombard speech adaptations for three of the four measures. While the speakers increased their intensity in Lombard speech in the late-focus sentences both in their non-native English and their native Dutch, they did less so in their native Dutch. With respect to the increase in spectral CoG in Lombard speech, this increase was larger in native Dutch than in these speakers' non-native English. Finally, the shortening in VOT in Lombard speech was smaller in native Dutch compared to non-native English. These differences in Lombard speech adaptations for these three measures between native Dutch and non-native English speech by the same speakers is not likely to result from differences in stimuli in the two languages. If the stimuli were influential (considering that native Dutch speech consisted of different stimuli than non-native English speech), then we would expect to see a difference in plain speech between the Dutch and English stimuli as well, rather than the differences that emerge only in Lombard speech. Combined, these results would indicate that the native Dutch speakers adapt the three acoustic characteristics to a different extent in their Lombard speech in native Dutch and in non-native English. Table 5 below summarized the results of the four acoustic measures in terms of the effect of Lombard speech, nativeness, and their interaction for both the comparison of *English speech* and *Dutch speakers*. Future research should investigate why differences between native Dutch and non-native English emerge for certain measures and not others, as it remains unclear.

Together, the two comparisons (of the native and non-native English and of the native Dutch and non-native English) suggest that the Dutch speakers adapted their non-native English to native English, when producing Lombard speech, and were not influenced by their native language in this respect. This may be surprising considering that past research on median f0 and f0 range in non-native Lombard speech suggested that the non-native English speakers were influenced by their native Dutch (Marcoux & Ernestus, 2019a, 2019b). This may indicate

Table 5
Summary of the English speech and Dutch speakers results in terms of whether there was a statistically significant effect of Lombard speech, nativeness, and their interaction.

	English speech			*Dutch speakers*		
	Lombard	Nativeness	Lombard * Nativeness	Lombard	Nativeness	Lombard * Nativeness
Intensity	Yes	No	No	–	–	–
Intensity: early-focus	–	–	–	Yes	No	No
Intensity: late focus	–	–	–	Yes	No	Yes
Spectral CoG	Yes	No	No	Yes	No	Yes
Duration	Yes	No	No	Yes	Yes	No
VOT	Yes	Yes	No	Yes	Yes	Yes

that it may depend on the acoustic measure whether we see a native language influence or not: one acoustic cue of Lombard speech may be easier to adapt than another one. This calls for further research, into more acoustic measures, which are language specific, such as vowel reduction on prestressed syllables, and spectral CoG of selected phonemes that have been documented to differ in their spectral CoG between the languages, for instance, fricatives such as /θ/ and /s/.

While our data indicate that the Dutch speakers produced non-native English Lombard speech similarly to native English speakers, it is unclear why this the case. Dutch learners of English do not explicitly learn (for instance, at school) how to produce Lombard speech in English. Perhaps they learn this unconsciously, for instance, when watching English spoken movies. Another possibility is that some of the phonological or phonetic differences between Dutch and English trigger differences in how acoustic characteristics are adapted in Lombard speech. Yet another possibility is that there are differences between non-native English as produced by native speakers of Dutch on the one hand and native speakers of English on the other hand, but that we did not find them because there were fewer native English speakers in the corpus compared to native Dutch speakers, which may have decreased statistical power. This calls for further research.

Future research should also extend our research to other types of speakers. First, the DELNN corpus only consists of recordings of women's voices. We know that there are differences between speech produced by women and men such as f0 differences. Our results are therefore not representative of Dutch speakers in general. Second, the corpus only contains native Dutch, native English, and non-native English produced by native speakers of Dutch. We chose to investigate non-native English as produced by native speakers of Dutch, because many Dutch speakers are so proficient in English that they can be expected to produce Lombard speech. Moreover, Dutch and English are very similar to each other, but also show differences, for instance, in VOT, which made it likely that differences between native and non-native speech may be found. Of note, we did not find differences in the extent of decrease in VOT length in Lombard speech for native and non-native English. Future research could focus on language pairs that differ more substantially from each other, which may enlarge the chance that differences between native and non-native Lombard speech will be observed. Additionally, the participants did not hear anything play from the headphones except for noise when Lombard speech was elicited. Therefore, further research could also look at the effect of having the speech fed-back to the participant.

As mentioned, the native Dutch speakers always completed the English session before the Dutch session, presenting a confound. However, these sessions were completed on separate days, and we do not expect that the participants behaved differently during the second session. Additionally, the plain condition always preceded the Lombard condition. Here, we also do not expect the order of the conditions to affect the results as we started with the less demanding condition, but future research could investigate the potential effect of the order of the session.

In the sentences read by the participants, the location of contrastive focus was manipulated, and we therefore included it as a predictor for the four acoustic measures in our analyses. Target words with contrastive focus were longer than those without, in native English, non-native English, and in native Dutch, in line with past research (e.g., Cooper et al., 1985; Sityaev & House, 2003). However, the native English speakers lengthened the words with contrastive focus more so than the non-native English speakers in English. The Dutch native speakers showed even more lengthening for words in focus in Lombard speech, both in Dutch and in non-native English. With respect to VOT, contrastive focus lengthened VOT more in native English than in non-native English, and not in native Dutch. Further, late-focus led to a higher average sentence intensity than early focus in native and non-native English Lombard speech, probably because fewer words in the late-focus condition than in early-focus condition underwent post-focus compression (PFC, e.g., Xu, 2011), where material after the focus is accompanied by a decrease in intensity (e.g., Chen, 2015). Finally, while focus has been shown to affect the distribution of energy in speech (e.g., Campbell, 1995; Campbell & Beckman, 1997), data from spectral CoG of the utterance, our chosen measure of energy, did not show an effect of focus. It could be the case that spectral CoG is affected by whether there is contrastive focus in the sentence, and not so much on where the contrastive focus is located. Combined these data patterns indicate that native and non-native English speakers were implementing focus (slightly) differently, where the non-native speakers were not just implementing focus as they did in their native language.

In conclusion, this article expands upon non-native speakers' production of Lombard speech by examining four distinct acoustic measures: intensity, spectral CoG, word duration, and VOT. We did not observe differences in how native speakers of English and of Dutch adapt their English speech in noisy conditions, indicating that the non-native English are producing Lombard speech similarly to the native English. Importantly, the comparison of the native Dutch and non-native English sentences produced by the same participants nevertheless suggests that, for several acoustic measurements, the Dutch speakers adapt their speech differently in native Dutch than in non-native English. Combined, this would indicate that,

when speaking English, Dutch speakers adapt their way of speaking in noisy conditions to native English.

Acknowledgements

We would like to give special thanks to Dr. Esther Janse for her help in the project since the beginning with feedback and comments. Additionally, we would like to thank Dr. Louis ten Bosch for his technical expertise as well as providing the noise file used to elicit the Lombard speech and for running the Dutch forced alignment. Further, we would like to thank Ton Wempe for his expert advice.

Funding

This project was funded by the European Union's Horizon 2020 research innovation programme (Marie Skłodowska-Curie grant No. 675324).

Appendix A

See Table A1.

Table A1
English target words used in the DELNN corpus per target word category.

Schwa	Voiced Obstruent	/θ/
Balloon	Blood	Theater
Banana	Cab	Theme
Botanical	Club	Theology
Cadaver	Crib	Theory
Computer	Food	Therapist
Gorilla	Lab	Thermal
Massage	Lemonade	Thermodynamics
Parade	Neighborhood	Thermometer
Police	Pub	Thermos
Professor	Rehab	Theta
Tomato	Road	Thriller
Salami	Wood	Throne

Appendix B

See Table B1.

Table B1
Dutch target words used with their English translation.

Dutch target word	English translation
Ballon	Balloon
Kadaver	Cadaver
Computer	Computer
Gorilla	Gorilla
Banaan	Banana
Massage	Massage
Politie	Police
Professor	Professor
Tomaat	Tomato
Botanische	Botanical
Salami	Salami
Parade	Parade
Universiteit	University
Hoofdgerecht	Main course
Appartement	Apartment
Bibliotheek	Library
Kostuum	Costume
Telefoon	Telephone
Oorbellen	Earrings
Museum	Museum
Artikel	Article
Programma	Program
Rugzak	Backpack
Gladiool	Gladiolus

Appendix C. AudiTon Microphone amplifier MA3

The following description is provided by the engineer, Ton Wempe, owner of the company AudiTon (Wempe, 2023). The text is verbatim except for the paragraph with the examples which we matched with the microphones in our study. This text has been approved by Ton Wempe.

The AudiTon microphone pre-amp MA3 has been designed mainly for high quality speech recording. The unit comprises two independent pre-amps with output volume controls. This features the avoidance of signal clipping due to overloading the inputs of the recording devices. The maximum output voltage of 12 dBu provides for a sufficient level for professional line inputs (e.g. 4 dBu).

The frequency range (70 . . . 18000 Hz) has been limited at the low range to avoid the mostly large background noises which have no components in the speech frequency range. At the high end the range has been somewhat limited to avoid the possible sample noise, generated by some ADC's in recording equipment or sound inputs of computers (e.g. caused by inappropiate filtering).

The (transformerless) pre amplifiers have very low self-noise (-130 dBu). When a microphone is used with a sensitivity of only 2 mV/Pa, the signal to noise ratio (S/N) of the amplifier itself amounts to 81 dB at a SPL of 1 Pa. (When using a microphone of, for example, 10 mV/Pa this S/N becomes about 95 dB.) These values do not take into account the thermal or electronics noise from the microphone used. When a passive dynamic microphone of 2 mV/Pa with an impedance of 200 Ω is used, the thermal noise is about 260 nV. This means that the microphone itself has an S/N of 78 dB at a SPL of 1 Pa. The combined S/N is then 75 dB.

The phantom power needed for some types of condensor microphones is available: it can be switched on for each pre-amp separately. The voltage is limited to 12 V, which does not meet the official standards (24 V or even 48 V), but most modern phantom-powered mics are very tolerant about this voltage.

The calibration tone

In practice when sound is recorded, the recording volume level of the recording device is adjusted for optimal use of its amplitude range (while taking care to avoid clipping of the peaks). Usually, the relation between the acoustic sound pressure level (SPL) and the waveform amplitude presented afterwards has been lost because the exact amplification factors are unknown: in practice, all volume controls, input sensitivities, etc. are uncalibrated.

The AudiTon microphone pre-amplifier MA3 features a sound level reference for calibration of the 'original' acoustic sound pressure level (SPL) at the position of the microphone. As long as the green button on the amplifier is pushed, the microphone signal is replaced with a generated sine wave of approx. 800 Hz. The level of the calibration tone is equivalent to 1 Pascal (or 94 dB) when a microphone of 2 mV/Pascal is used.

The calibration tone's level with respect to the acoustic SPL is independent of the volume control position of the pre-amp and the recording level of the recording device: the micro-

phone signal and the calibration tone signal levels are equally altered by all volume controls, which don't alter their level ratio. The calibration sine wave, afterwards displayed in the waveform used as a reference, offers the possibility to estimate the **absolute** acoustic intensity contour of the recorded sound.

After (re-)adjustment of the recording level or changing the volume control on the pre-amp, the calibration button should be pushed (again) to be able to provide for a new marker with the proper calibration level in the next recording. In this way the absolute acoustic levels of all parts of the recording can be estimated afterwards.

As an example, when a sound editor displays the reference tone part in the waveform to have an intensity of 81 dB, all intensity levels of the signal's waveforms have to be corrected by 94 − 81 dB = +13 dB.

When microphones with different sensitivities are used the calibration tone's level can be computed by the formula:

$$L_{CAL} = \frac{2}{S_M} Pascal$$

where S_M represents the microphone's sensitivity in mV/Pa.

For example, when the mic's sensitivity is 10 mV/Pa, the calibration tone will have a level of 2/10 Pa (or 94 − 14 = 80 dB). Suppose that the sound editor displays the calibration tone in the waveform with an intensity of 81 dB, all intensity levels of the signal's waveforms have to be corrected by 80 − 81 dB = −1 dB. To give another example, when the microphone's sensitivity is 31 mV/Pa, the calibration tone will have a level of 2/31 Pa (or 94 − 23.8 = 70.2 dB). Suppose that the sound editor again displays the calibration tone in the waveform with an intensity of 81 dB, all intensity levels of the signal's waveforms have to be corrected by 70.2 − 81 = −10 .8 dB.

Of course, the tone can also be applied to place markers in the recordings to define the beginnings of specific parts.

Appendix D. MFA evaluation

The English data were annotated at the phone and word level with the Montreal Forced Aligner (MFA; McAuliffe et al., 2017), which uses Kaldi as its basis (Povey et al., 2011). In order to apply forced alignment, the speech signal (WAV files), an orthographic transcription, a pronunciation dictionary, and the phone models were provided. In the orthographic transcription of what the speaker produced, false starts were included for the answers so that the best phonetic annotation possible could be provided. The dictionary provides a map from the words (in the orthographic transcription) to the phones.

Since the English stimuli were recorded by native English as well as native Dutch speakers, Dutch-accented pronunciations of the target words were included in the pronunciation dictionary. We used the Carnegie Mellon University (CMU) Pronouncing Dictionary (*CMU* Pronouncing Dictionary, 2015), which has the American-English pronunciations of words and added Dutch-accented variants for the three target word categories. For the /θ/-initial target words, /t/, /d/, /f/, /v/, /s/, and /z/ were included as alternative pronunciations of /θ/. As for the schwa target words, alternative pronunciations were included where schwa was replaced by /ʌ/, /æ/, /ɑ/, and /ɔ/ if the schwa was orthographically spelled as <a>, and replaced by /ɔ/, /o/, and /ɑ/ if it was spelled with <o>. For the target words with final voiced obstruents, we included variants with /t/ and /p/ at the end of the word instead of /d/ and /b/, respectively.

The acoustic models used for the transcription of the English utterances were English phones trained on the LibriSpeech corpus, with 1000 hours of read speech (Panayotov et al., 2015). First, MFA calculated monophone Gaussian Mixture Models-based (GMM) Hidden Markov Models (HMMs) and then, in order to take the surrounding phones into account, calculated triphone GMM-HMM models (McAuliffe et al., 2017). MFA calculated 13 mel-frequency cepstral coefficients (MFCCs), as well as 13 each for delta and delta-delta, resulting in 39 features per frame. The acoustic models resulted in a total of 68 GMM-HMMs (one for each vowel and consonant, and additional ones for e.g. silences). Cepstral mean and variance normalization (CMVN) was applied per speaker. Speaker adaptation was not implemented since it did not improve phone-level transcription.

In order to evaluate the phone level transcriptions for the English utterances, we had two trained human annotators annotate 25 sentences from 13 non-native English speakers, of which 14 sentences were plain speech and 11 were Lombard speech. We found that overall, the MFA annotation was comparable to that produced by the human annotators (see the all phones section in Table D1).

Since silences influence spectral CoG and intensity values at the sentence level, we needed to remove them and therefore we specifically examined how well MFA annotated silences. The agreement between each of the human transcriptions and the MFA transcriptions was overall lower than the agreement between the two human transcriptions, but this was especially so for the silences' boundaries. In order to investigate this further, the first author annotated silences from 60 randomly selected utterances not considered in the evaluation (a combination of plain and Lombard speech as well as a combination of native English and non-native English), resulting in manual annotation of 117 silence start boundaries and 55 silence end boundaries (there were more start boundaries than end boundaries because many end boundaries were sentence final and therefore not annotated). While the annotations of the silence start boundaries did not show a consistent pattern and could therefore not be improved, these annotations suggested that the MFA silence end boundaries were on average 30 ms late, unless they were sentence final. We therefore lengthened the non-sentence final silences by 25 ms (5 ms less than the average difference in order to ensure that only in few cases the annotated silence lasted in the next phone). This change improved the agreement between the MFA and human annotators, increasing the number of silence end boundaries that were within 25 ms of each other.

Table D1 illustrates the agreement of the MFA transcription with each of the two human transcribers and the agreement between the two human transcribers themselves. It shows the effect of lengthening the non-utterance final silence boundaries by 25 ms. A 25 ms window was chosen as to be able to compare MFA performance's on our data to its evaluation of other corpora (see McAuliffe et al., 2017). Table D1 indicates that after lengthening the non-utterance final silence boundaries, of the phones with the same labels, 75.4 % and 75.1 % were less than 25 ms off from Human 1 and Human

Table D1
Statistics on label and boundary agreement between MFA, Human 1 and Human 2, for pre- and post-moving of the silence end boundary. 'All labels' is the combined number of phones that the two annotators transcribed for the specified category (silence, or all phones). 'Same labels (%)' is the number of phones for which the two annotators agreed upon the labeling followed by its percentage of the 'All labels', in parenthesis. The 'Boundaries within 25 ms (%)' is the number of boundaries that the two annotators labeled the same (Same labels) and were within 25 ms of each other, followed by the percentage in parenthesis.

	Pre Move			Post move		
	All labels	Same labels (%)	Boundaries within 25 ms (%)	All labels	Same labels (%)	Boundaries within 25 ms (%)
Silence start boundary						
MFA-Human 1	56	39 (69.6)	25 (64.1)	56	39 (69.6)	25 (64.1)
MFA-Human 2	51	40 (78.4)	25 (62.5)	51	40 (78.4)	25 (62.5)
Human 1–Human 2	47	37 (78.7)	33 (89.2)	47	37 (78.7)	33 (89.2)
Silence end boundary						
MFA-Human 1	56	27 (48.2)	6 (22.2)	57	27 (47.4)	21 (77.8)
MFA-Human 2	51	28 (54.9)	8 (28.6)	52	28 (53.9)	22 (78.6)
Human 1–Human 2	47	37 (78.7)	36 (97.3)	47	37 (78.7)	36 (97.3)
All phones						
MFA-Human 1	803	712 (88.7)	525 (73.7)	804	710 (88.3)	535 (75.4)
MFA-Human 2	798	645 (80.8)	476 (73.8)	799	643 (80.5)	483 (75.1)
Human 1–Human 2	792	676 (85.4)	593 (87.7)	792	676 (85.4)	593 (87.7)

Table D2
Statistics on label and boundary agreement between Kaldi (Povey et al., 2011), Human 1 and Human 2, for pre- and post-moving of the silence start and end boundaries. 'All labels' is the combined number of phones that the two annotators transcribed for the specified category (silence, or all phones). 'Same labels (%)' is the number of phones which the two annotators agreed upon the labeling, followed by its percentage of the 'All labels', in parenthesis. The 'Boundaries within 25 ms (%)' is the number of boundaries that the two annotators labeled the same ('Same labels') and were within 25 ms of each other, followed by the percentage in parenthesis.

	Pre Move			Post Move		
	All labels	Same labels (%)	Boundaries within 25 ms (%)	All labels	Same labels (%)	Boundaries within 25 ms (%)
Silence start boundary						
Kaldi-Human 1	44	36 (81.8)	17 (47.2)	44	36 (81.8)	24 (66.7)
Kaldi-Human 2	38	30 (79.0)	15 (50.0)	38	30 (79.0)	24 (80.0)
Human 1–Human 2	42	30 (71.4)	22 (73.3)	42	30 (71.4)	22 (73.3)
Silence end boundary						
Kaldi-Human 1	44	26 (59.1)	3 (11.5)	44	26 (59.1)	11 (42.3)
Kaldi-Human 2	38	23 (60.5)	4 (17.4)	38	23 (60.5)	18 (78.3)
Human 1–Human 2	42	33 (78.6)	25 (75.8)	42	33 (78.6)	25 (75.8)
All phones						
Kaldi-Human 1	1045	885 (84.7)	711 (80.3)	1045	885 (84.7)	727 (82.2)
Kaldi-Human 2	1041	860 (82.6)	723 (84.1)	1041	860 (82.6)	745 (86.6)
Human 1–Human 2	988	893 (90.4)	818 (91.6)	988	893 (90.4)	818 (91.6)

2′s annotations, respectively. These figures are in the same range as the forced aligners trained by other researchers, for instance, McAuliffe et al. (2017) found that 77 % of the aligned phone boundaries were less than 25 ms off from the gold-standard annotations for the Buckeye corpus and 72 % for the Phonsay corpus.

The Dutch data were directly annotated with Kaldi (Povey et al., 2011) since there were no acoustic models for Dutch built into MFA. As with the English annotation, the orthographic transcription of what the speaker produced included false starts for the answers. The pronunciation dictionary was created from a combination of Celex (Baayen, Piepenbrock, & Gulikers, 1995) and the Spoken Dutch Corpus (CGN; Dutch Language Institute, 2014). The acoustic models were trained on the complete CGN except for the telephone recordings, which have a low acoustic quality. As was the case with MFA (McAuliffe et al., 2017), Kaldi (Povey et al., 2011) computed 13 MFCCs and the delta and delta-delta, for a total of 39 features per frame. The training resulted in 50 nnet3 triphone models (DNNs) for vowels, consonants and other speech sounds. Cepstral mean and variance normalization (CMVN) was applied per utterance.

In order to evaluate Kaldi's (Povey et al., 2011) transcription, two different human annotators annotated 25 Dutch utterances, 15 plain utterances from a separate corpus and 10 Lombard utterances from the DELNN corpus. The results are presented in Table D1. As with the English transcriptions, we found that the annotations of the Dutch silence boundaries could be improved. In order to calculate how the silence boundaries should be adjusted for improvement, the first author annotated 30 answers (20 plain and 10 Lombard) from the DELNN corpus not included in the evaluation. This led to a total of 86 silence start and 57 silence end boundary annotations (excluding utterance final boundaries). These annotations indicated that the silence start boundary should be moved forward by 20 ms and the silence end boundary should be lengthened by 20 ms. These changes improved the transcription, as can be seen in Table D1 below. This resulted in Human 1 and Kaldi having 82.2 % of all phones (that had the same label) within 25 ms of each other and 86.6 % for Human 2 and Kaldi. Here we see that the evaluation of the Dutch transcriptions which used Kaldi, is even better than the evaluation of the English transcriptions above for MFA, which is similar to the evaluation of the forced aligner by McAuliffe et al. (2017).

Appendix E

See Fig. E1

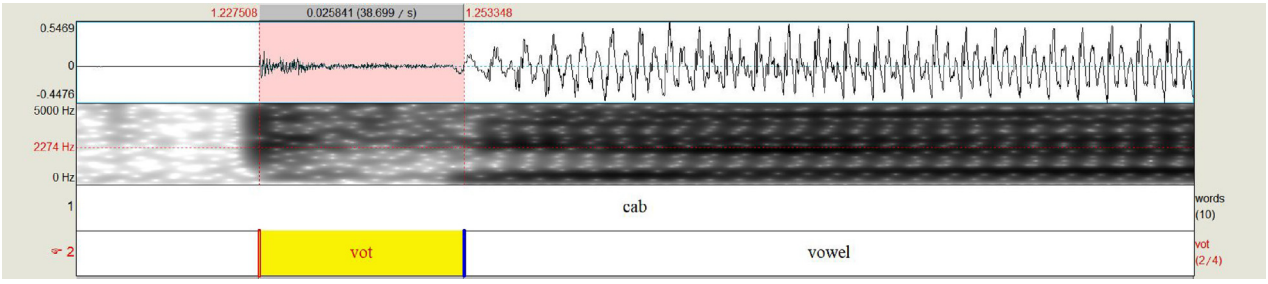


Fig. E1. Example of a VOT annotation.

Appendix F. VOT: Interrater agreement

There were three annotators, of which one annotated tokens of all the target words while each of the two others only annotated parts of the target words, not overlapping. Human 1 and Human 2 both annotated English *computer*, *cadaver*, *parade*, *professor*, *police*, and *cab*. To determine the interrater agreement, we had a total of 60 annotations from each annotator (Human 1 and Human 2), nine per target word except for *pub,* which had six. Human 2 and Human 3 both annotated the English *crib* and *club*, and the Dutch *kadaver*, *kostuum*, *computer*, *parade*, and *politie*. For determining the interrater agreement, we had a total of 62 annotations from each annotator (Human 2 and Human 3), nine per target word except for *crib*, which had eight. Table F1 below shows the results of this comparison. The values in the table indicate that the differences between Human 1 and Human 2 and between Human 2 and Human 3 are not statistically significant. This can be seen in the values of the mean and the standard deviation. When the standard deviation is added or subtracted from the mean, it includes zero, indicating that the human annotators in each comparison are not significantly different from each other.

Table F1
The average and standard deviation of the difference between the two human annotators for VOT start boundary, VOT end boundary and vowel end boundary.

	Human 1- Human 2		Human 2 – Human 3	
Boundary	*M*	*SD*	*M*	*SD*
VOT start	−1.3 ms	11.5 ms	0.0 ms	1.6 ms
VOT end	0.2 ms	2.8 ms	−1.8 ms	2.1 ms
Vowel end	0.4 ms	15.0 ms	−6.1 ms	12.7 ms

Appendix G. Native Dutch and native English speech comparison

Here we present the results of the comparison of native Dutch and native English speech. We follow the same model selection procedure as discussed in Section 3.3. The difference is that we replaced the predictor of interest Nativeness by Language, since, in both cases, the speech is produced by native speakers and differs in language. Further, for this model, instead of having non-native speaker (Nativeness) on the intercept, English (Language) is on the intercept.

The *English speech* and the *Dutch speakers* analyses reported in the body of the paper, did not show a difference between how the opposition between plain and Lombard

Table G1
Lmer models of native Dutch and native English intensity.

Fixed Effects	*β*	*t*	*p*
(Intercept)	67.04	143.12	<0.001
Speech Style: Lombard	8.03	14.81	<0.001
Contrastive: Focus	0.46	2.38	0.019
Trial Number	0.34	4.55	<0.001
Random Effects			*SD*
Speaker (Intercept)			2.76
Speech Style by Speaker			3.22
Contrastive by Speaker			0.69
Answer (Intercept)			1.06
Residual			1.05

Table G2
Lmer models of native Dutch and native English spectral CoG.

Fixed Effects	*β*	*t*	*p*
(Intercept)	833.70	34.15	<0.001
Speech Style: Lombard	219.50	9.41	<0.001
Trial Number	22.28	4.75	<0.001
Random Effects			*SD*
Speaker (Intercept)			130.71
Speech Style by Speaker			133.87
Answer (Intercept)			146.96
Speech Style by Answer			45.96
Residual			128.74

Table G3
Lmer models of native Dutch and native English target word durations.

Fixed Effects	*β*	*t*	*p*
(Intercept)	394.29	15.91	<0.001
Speech Style: Lombard	54.14	4.66	<0.001
Language: Dutch	67.80	2.09	0.039
Contrastive: Focus	147.56	10.90	<0.001
Occurrence	−18.18	−2.91	0.006
Language: Dutch * Contrastive: Focus	−64.82	−4.20	<0.001
Random Effects			*SD*
Speaker (Intercept)			53.12
Speech Style by Speaker			17.51
Focus by Speaker			39.81
Target word (Intercept)			93.83
Residual			51.22

Table G4
Lmer models of native Dutch and native English VOT.

Fixed Effects	*β*	*t*	*p*
(Intercept)	65.68	13.83	<0.001
Speech Style: Lombard	−4.06	−2.60	0.009
Language: Dutch	−23.58	−4.02	0.001
Contrastive: Focus	15.36	12.33	<0.001
Stress: unstressed	−20.28	−3.38	0.006
Trial Number	−6.45	−5.27	<0.001
Occurrence	4.67	3.51	<0.001
Language: Dutch * Contrastive: Focus	−17.35	−11.47	<0.001
Random Effects			*SD*
Speaker (Intercept)			3.56
Target word (Intercept)			8.74
Residual			10.34

speech affects intensity, spectral CoG, and VOT in native English speakers and in non-native English speakers, while these acoustic measures differed between non-native English Lombard speech and native Dutch Lombard speech for three acoustic measures. Based on this, we could expect an interaction of Language and Speech Style for intensity, spectral CoG, and VOT. The results presented below, in Tables G1–G4, are not in line with this expectation. This may be due to the small number of native English speakers in combination with the differences in stimuli (due to the two different languages) read aloud by the participants. This may have resulted in more noise and lower power compared to the analyses in the main body of the paper.

Appendix H. Figures of the data with the outliers removed

See Figs. H1–H8.

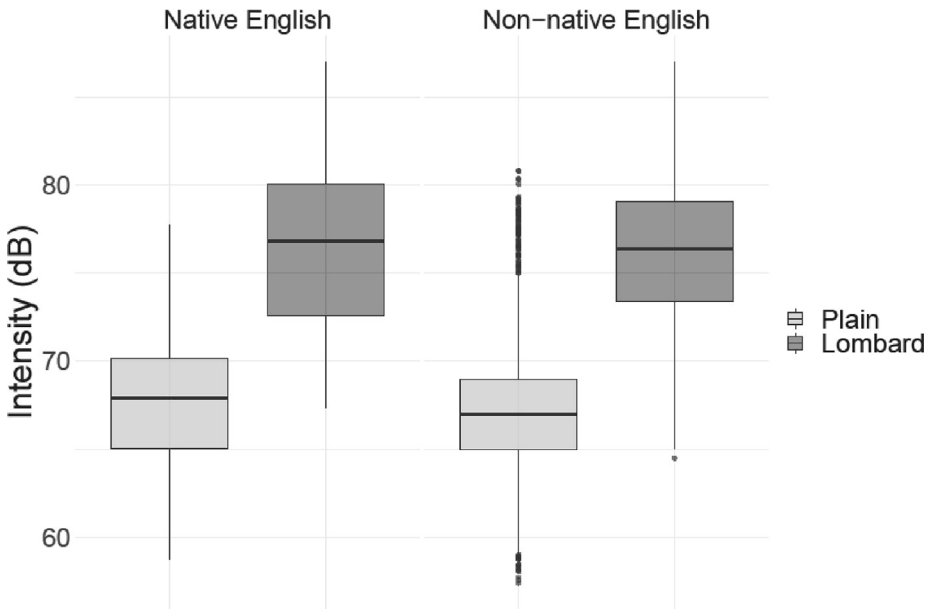


Fig. H1. Average intensity data for native English and non-native English split by speech style. The data are visualized after outliers were removed for the statistical analysis.

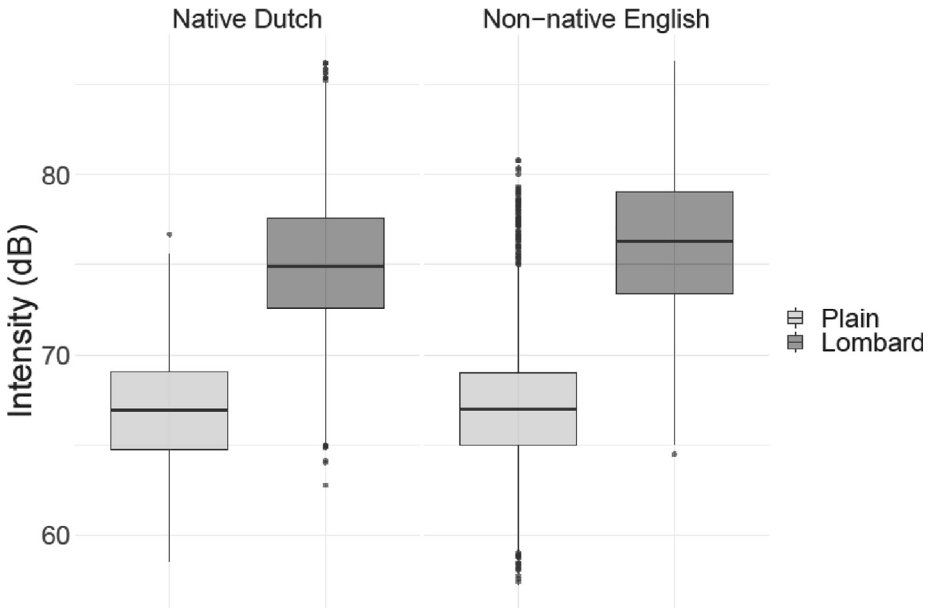


Fig. H2. Average intensity data for non-native English and native Dutch split by speech style. The data are visualized after outliers were removed for the statistical analysis.

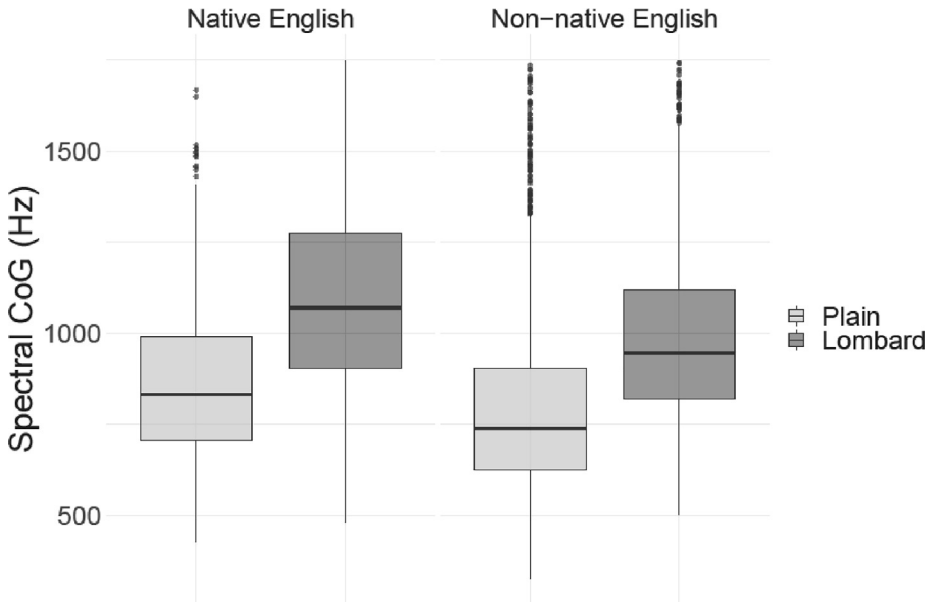


Fig. H3. Spectral CoG data for native English and non-native English split by speech style. The data are visualized after outliers were removed for the statistical analysis.

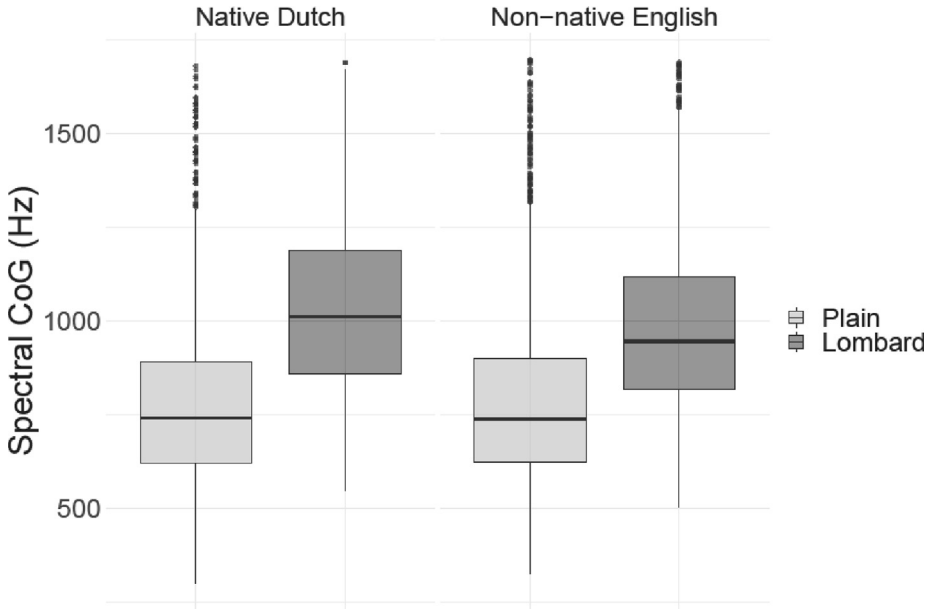


Fig. H4. Spectral CoG data for non-native English and native Dutch split by speech style. The data are visualized after outliers were removed for the statistical analysis.

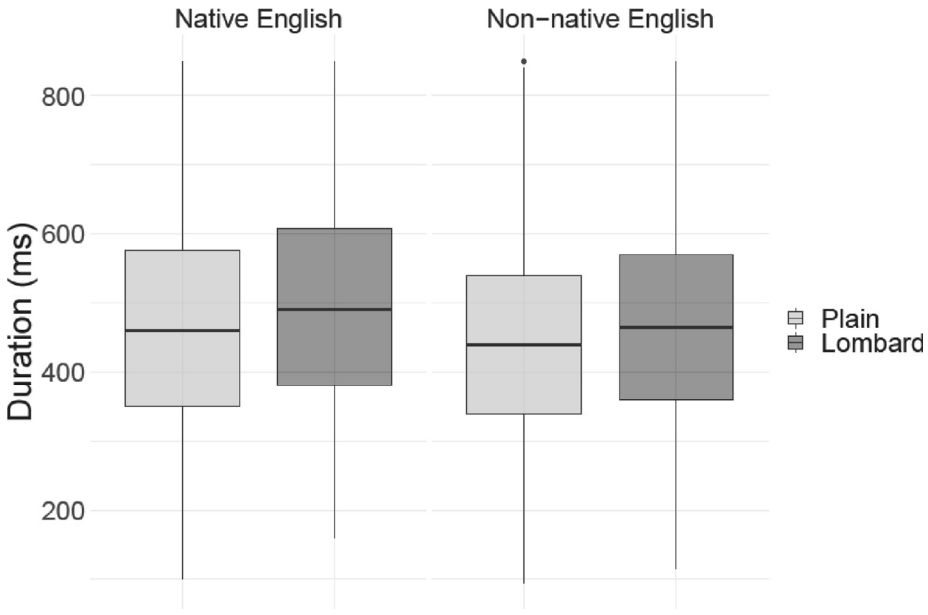


Fig. H5. The durations of target words produced by native English and non-native English split by speech style. The data are visualized after outliers were removed for the statistical analysis.

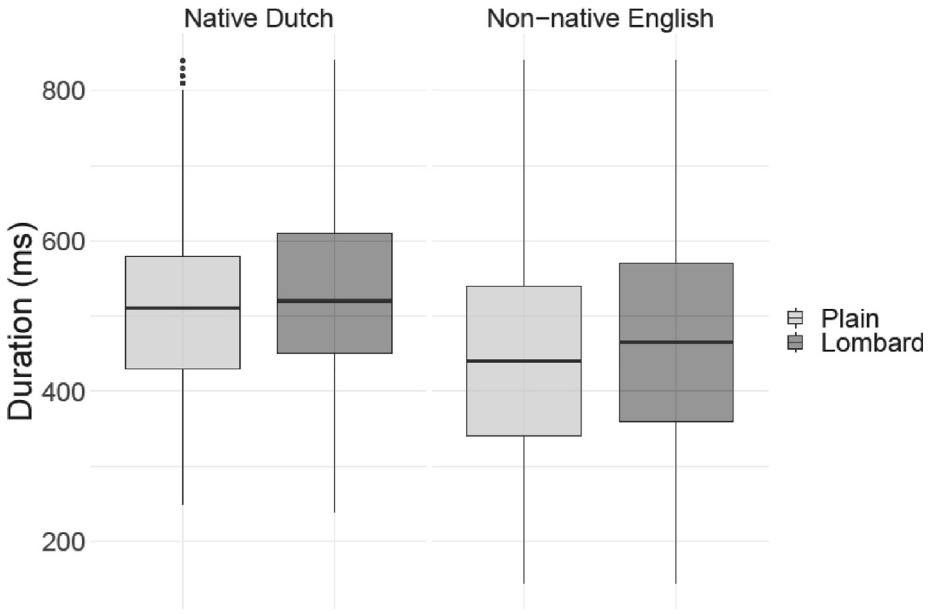


Fig. H6. The durations of target words produced by non-native English and native Dutch split by speech style. The data are visualized after outliers were removed for the statistical analysis.

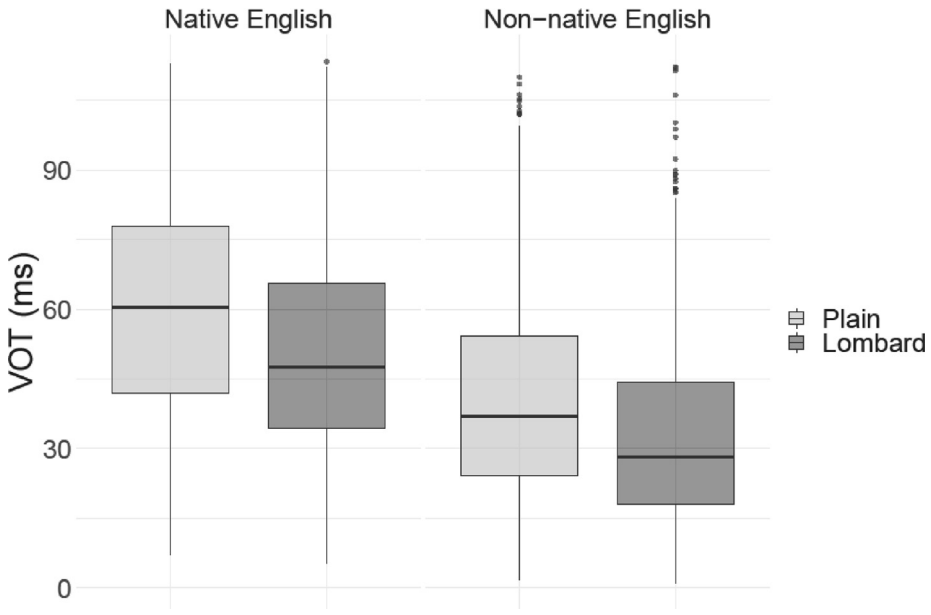


Fig. H7. The VOT of /p/ and /k/ produced by native English and non-native English split by Speech Style. The data are visualized after outliers were removed for the statistical analysis.

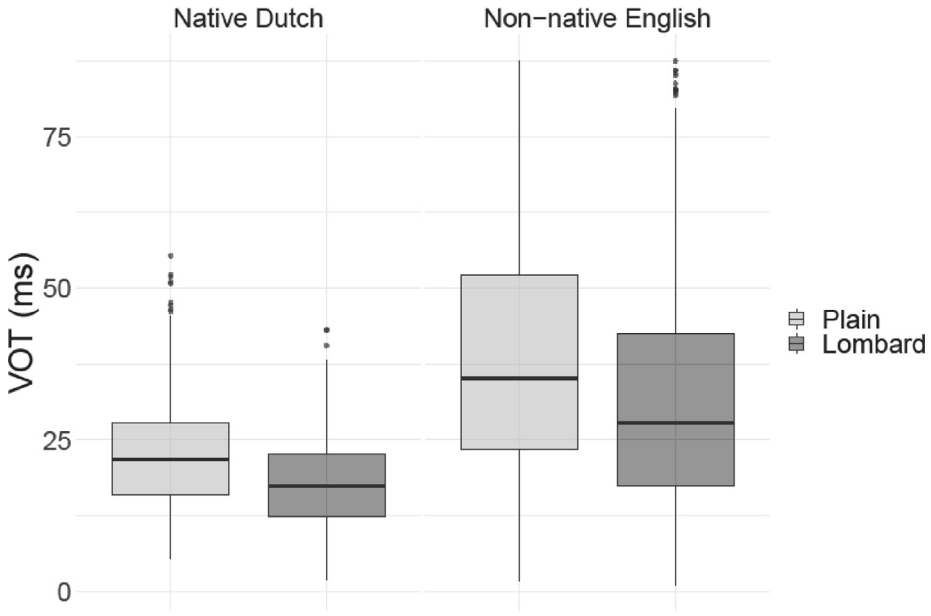


Fig. H8. The VOT of /p/ and /k/ produced by non-native English and native Dutch split by Speech Style. The data are visualized after outliers were removed for the statistical analysis.

References

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX Lexical Database (CD-ROM)* (Release 2, Dutch Version 3.1) [Data set]. Linguistic Data Consortium.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01.

Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language, 60*(1), 92–1111. https://doi.org/10.1016/j.jml.2008.06.003.

Berendsen, E. (1986). The phonology of Dutch cliticization. In W. de Gruyter (Ed.), *The Phonology of Cliticization* (pp. 35–98). Foris Publications.

Boersma, P., & Weenink, D. (2018). Praat: doing phonetics by computer (Version 6.0.37) [Computer software]. http://www.praat.org/.

Booij, G. (1985). Lexical phonology, final devoicing and subject pronouns in Dutch. *Linguistics in the Netherlands*, 21–26. https://doi.org/10.1515/9783112330128-005.

Booij, G. (1999). *The phonology of Dutch*. Oxford University Press.

Bosker, H. R., & Cooke, M. (2018). Talkers produce more pronounced amplitude modulations when speaking in noise. *The Journal of the Acoustical Society of America, 143*(2), 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. *The Journal of the Acoustical Society of America, 147*(2), 721–730. https://doi.org/10.1121/10.0000646.

Burgos, P., Cucchiarini, C., van Hout, R., & Strik, H. (2013). Pronunciation errors by Spanish learners of Dutch: A data-driven study for ASR-based pronunciation training. In F. Bimbot, C. Cerisara, C. Fougeron, G. Gravier, L. Lamel, F. Pellegrino, and P. Perrier (Eds.) *Proceedings of the 14th Annual Conference of the International Speech Communication Association (INTERSPEECH 2013)* (pp. 2385–2389).

Cai, X., Yin, Y., & Zhang, Q. (2020). A cross-language study on feedforward and feedback control of voice intensity in Chinese-English bilinguals. *Applied Psycholinguistics, 41*(4), 771–795. https://doi.org/10.1017/S0142716420000223.

Cai, X., Yin, Y., & Zhang, Q. (2021). Online control of voice intensity in late bilinguals' First and second language speech production: Evidence from unexpected and brief noise masking. *Journal of Speech, Language, and Hearing Research, 64*(5), 1471–1489. https://doi.org/10.1044/2021_JSLHR-20-00330.

Campbell, W. N. (1995). Loudness, spectral tilt, and perceived prominence in dialogues. In K. Elenius, & Branderud (Eds.) *Proceedings of the 13th International Congress of Phonetic Sciences* (pp. 676–679).

Campbell, N., & Beckman, M. (1997). Stress, prominence, and spectral tilt. In A. Botinis (Ed.) *Proceedings of Intonation: Theory, Models and Applications*.

Castellanos, A., Benedí, J. M., & Casacuberta, F. (1996). An analysis of general acoustic-phonetic features for Spanish speech produced with the Lombard effect. *Speech Communication, 20*(1–2), 23–35. https://doi.org/10.1016/S0167-6393(96)00042-8.

Chen, Y. (2015). Post-fous compression in English by Mandarin learners. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences*. The University of Glasgow.

Cho, T., & McQueen, J. M. (2005). Prosodic influences on consonant production in Dutch: Effects of prosodic boundaries, phrasal accent and lexical stress. *Journal of Phonetics, 33*(2), 121–157. https://doi.org/10.1016/j.wocn.2005.01.001.

Choi, H. (2003). Prosody-induced acoustic variation in English stop consonants. In M. J. Solé, D. Recasens, and J. Romero (Eds.), *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 2661–2664).

CMU Pronouncing Dictionary (2015). (Version 0.7b). http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

Collins, B., & Mees, I. (1996). *The phonetics of English and Dutch* (3rd. rev.). Brill.

Cooke, M., King, S., Garnier, M., & Aubanel, V. (2014). The listening talker: A review of human and algorithmic context-induced modifications of speech. *Computer Speech & Language, 28*(2), 543–571. https://doi.org/10.1016/j.csl.2013.08.003.

Cooper, W. E., Eady, S. J., & Mueller, P. R. (1985). Acoustical aspects of contrastive stress in question–answer contexts. *The Journal of the Acoustical Society of America, 77*(6), 2142–2156. https://doi.org/10.1121/1.392372.

Council of Europe (2001). Council for Cultural Co-operation. Education Committee. Modern Languages Division (Strasbourg). *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. *The Journal of the Acoustical Society of America, 29*(12), 1320–1323. https://doi.org/10.1121/1.1908780.

Dutch Language Institute (2014). *Corpus Gesproken Nederlands - CGN* (Version 2.0.3). http://hdl.handle.net/10032/tm-a2-k6.

Elsendoorn, B. A. (1985). Production and perception of Dutch foreign vowel duration in English monosyllabic words. *Language and Speech, 28*(3), 231–254.

Flege, J. E., & Eefting, W. (1987). Cross-language switching in stop consonant perception and production by Dutch speakers of English. *Speech Communication, 6*(3), 185–202. https://doi.org/10.1016/0167-6393(87)90025-2.

Fowler, C. A., & Housum, J. (1987). Talkers' signaling of "new" and "old" words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language, 26*(5), 489–504. https://doi.org/10.1016/0749-596X(87)90136-7.

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3rd ed.). Sage Publications.

Garnier, M., & Henrich, N. (2014). Speaking in noise: How does the Lombard effect improve acoustic contrasts between speech and ambient noise? *Computer Speech & Language, 28*(2), 580–597. https://doi.org/10.1016/J.CSL.2013.07.005.

Gramming, P., Sundberg, J., Ternström, S., Leanderson, R., & Perkins, W. H. (1988). Relationship between changes in voice pitch and loudness. *Journal of Voice, 2*(2), 118–126. https://doi.org/10.1016/S0892-1997(88)80067-5.

Gussenhoven, C., & Broeders, A. (1997). *English pronunciation for student teachers* (2nd ed.). Groningen: Wolters-Noordhoff.

Hanssen, J. E. G., Peters, J., & Gussenhoven, C. (2008). Prosodic effects of focus in Dutch declaratives. In Plínio A. Barbosa, Sandra Madureira, and Cesar Reis (Eds.), *Proceedings of Speech Prosody 2008* (pp. 609–612).

Hanulikova, A., & Weber, A. (2010). Production of English interdental fricatives by Dutch, German, and English speakers. In K. Dziubalska-Kołaczyk, M. Wrembel, & M. Kul (Eds.), *Proceedings of the 6th International Symposium on the Acquisition of Second Language Speech, New Sounds 2010* (pp. 173–178). Adam Mickiewicz University.

Hazan, V., Grynpas, J., & Baker, R. (2012). Is clear speech tailored to counter the effect of specific adverse listening conditions? *The Journal of the Acoustical Society of America, 132*(5), EL371–EL377. https://doi.org/10.1121/1.4757698.

Hillenbrand, J. M., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America, 97*(5), 3099–3111. https://doi.org/10.1121/1.411872.

House, A. S. (1961). On vowel duration in English. *The Journal of the Acoustical Society of America, 33*(9), 1174–1178. https://doi.org/10.1121/1.1908941.

Johnson, K., & Babel, M. (2010). On the perceptual basis of distinctive features: Evidence from the perception of fricatives by Dutch and English speakers. *Journal of Phonetics, 38*(1), 127–136. https://doi.org/10.1016/j.wocn.2009.11.001.

Junqua, J. C., Fincke, S., & Field, K. (1999). The Lombard effect: A reflex to better communicate with others in noise. In *Proceedings of the 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP99 (Cat. No. 99CH36258)* (pp. 2083–2086). IEEE.

Junqua, J. C. (1993). The Lombard reflex and its role on human listeners and automatic speech recognizers. *The Journal of the Acoustical Society of America, 93*(1), 510–524. https://doi.org/10.1121/1.405631.

Kormos, J. (2006). Monitoring. In *Speech Production and Second Language Acquisition* (pp. 122–136). Lawrence Erlbaum Associates.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13.

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods, 44*(2), 325–343. https://doi.org/10.3758/s13428-011-0146-0.

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word, 20*(3), 384–422. https://doi.org/10.1080/00437956.1964.11659830.

Lisker, L., & Abramson, A. S. (1967). Some effects of context on voice onset time in English stops. *Language and Speech, 10*(1), 1–28. https://doi.org/10.1177/002383096701000101.

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

Lu, Y., & Cooke, M. (2008). Speech production modifications produced by competing talkers, babble, and stationary noise. *The Journal of the Acoustical Society of America, 124*(5), 3261–3275. https://doi.org/10.1121/1.2990705.

Lu, Y., & Cooke, M. (2009). Speech production modifications produced in the presence of low-pass and high-pass filtered noise. *The Journal of the Acoustical Society of America, 126*(3), 1495–1499. https://doi.org/10.1121/1.3179668.

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

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

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. In *Proceedings of the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH 2017)* (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 *Proceedings of the 9th International Conference on Speech Prosody 2018* (pp. 848–852). https://doi.org/10.21437/SpeechProsody.2018-171.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. In *Proceedings of* the 2015 *IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 5206–5210). IEEE.

Pick, H. L., Siegel, G. M., Fox, P. W., Garber, S. R., & Kearney, J. K. (1989). Inhibiting the Lombard effect. *The Journal of the Acoustical Society of America, 85*(2), 894–900. https://doi.org/10.1121/1.397561.

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

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., ... Vesely., K. (2011). The Kaldi speech recognition toolkit. In *Proceedings of IEEE 2011 Workshop*

on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society.

Quené, H., Orr, R., & van Leeuwen, D. (2017). Phonetic similarity of /s/ in native and second language: Individual differences in learning curves. *The Journal of the Acoustical Society of America, 142*(6), EL519–EL524. https://doi.org/10.1121/1.5013149.

R Core Team (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rump, H. H., & Collier, R. (1996). Focus conditions and the prominence of pitch-accented syllables. *Language and Speech, 39*(1), 1–17. https://doi.org/10.1177/002383099603900101.

Segalowitz, N. (2010). Second language cognitive fluency. In *Cognitive bases of second language fluency* (pp. 74–106). Routledge.

Simon, E. (2010). Phonological transfer of voicing and devoicing rules: Evidence from L1 Dutch and L2 English conversational speech. *Language Sciences, 32*(1), 63–86. https://doi.org/10.1016/J.LANGSCI.2008.10.001.

Simon, E., & Leuschner, T. (2010). Laryngeal systems in Dutch, English, and German: A contrastive phonological study on second and third language acquisition. *Journal of Germanic Linguistics, 22*(4), 403–424. https://doi.org/10.1017/S1470542710000127.

Simonet, M., Casillas, J. V, & Díaz, Y. (2014). The effects of stress/accent on VOT depend on language (English, Spanish), consonant (/d/,/t/) and linguistic experience (monolinguals, bilinguals). In *Proceedings of the 7th International Conference on Speech Prosody* (pp. 202–206).

Sityaev, D., & House, R. (2003). Phonetic and phonological correlates of broad, narrow and contrastive focus in English. In M. J. Solé, D. Recasens, and J. Romero (Eds.), *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 1819–1822).

van Bergem, D. R. (1993). Acoustic vowel reduction as a function of sentence accent, word stress, and word class. *Speech Communication, 12*(1), 1–23. https://doi.org/10.1016/0167-6393(93)90015-D.

van Maastricht, L., Krahmer, E., & Swerts, M. (2016). Prominence patterns in a second language: Intonational transfer from Dutch to Spanish and vice versa. *Language Learning, 66*(1), 124–158. https://doi.org/10.1111/lang.12141.

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. *The Journal of the Acoustical Society of America, 84*(3), 917–928. https://doi.org/10.1121/1.396660.

Varadarajan, V. S., & Hansen, J. H. L. (2006). Analysis of Lombard effect under different types and levels of noise with application to in-set speaker ID systems. In *Proceedings of the Ninth International Conference on Spoken Language Processing (Interspeech 2006 – ICSLP).*

Villegas, J., Perkins, J., & Wilson, I. (2021). Effects of task and language nativeness on the Lombard effect and on its onset and offset timing. *The Journal of the Acoustical Society of America, 149*(3), 1855–1865. https://doi.org/10.1121/10.0003772.

Wester, M., García Lecumberri, L., Cooke, M. (2014). DIAPIX-FL: A symmetric corpus of problem-solving dialogues in first and second languages. In *Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association (INTERSPEECH 2014)* (pp. 509–513).

Welby, P. (2006). Intonational differences in Lombard speech: Looking beyond F0 range. In *Proceedings of the Third International Conference on Speech Prosody* (pp. 763–766).

Wempe, T. (2023). *AudiTon Microphone amplifier MA3.* AudiTon.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag.

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

Xu, Y., & Xu, C. X. (2005). Phonetic realization of focus in English declarative intonation. *Journal of Phonetics, 33*(2), 159–197. https://doi.org/10.1016/j.wocn.2004.11.001.

Yao, Y. (2009). Understanding VOT variation in spontaneous speech. *UC Berkley PhonLab Annual Report, 5*(5), 29–43.

Zollinger, S. A., & Brumm, H. (2011). The Lombard effect. *Current Biology, 21*(16), R614–R615. https://doi.org/10.1016/j.cub.2011.06.003.