Effect of articulatory and acoustic features on the intelligibility of speech in noise: an articulatory synthesis study

Thuanvan Ngo[a,*], Masato Akagi[a], Peter Birkholz[b]

[a] *Graduate School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa, 923-1292, Japan*
[b] *Institute of Acoustics and Speech Communication, TU Dresden, Germany*

Abstract

In noisy conditions, speakers involuntarily change their manner of speaking to enhance the intelligibility of their voices. The increased intelligibility of this so-called Lombard speech is enabled by the change of multiple articulatory and acoustic features. While the major features of Lombard speech are well known from previous studies, little is known about their relative contributions to the intelligibility of speech in noise. This study used an analysis-by-synthesis strategy to explore the contributions of multiple of these features. To this end, an articulatory speech synthesizer was used to synthesize the ten German digit words "Null" to "Neun", for all 16 combinations of four binary features, i.e., modal vs. pressed phonation, normal vs. increased F_1 and F_2 formant frequencies, normal vs. increased f_0 mean and range, and normal vs. increased duration of vowels. Subjects were asked to try to recognize the synthesized words in the presence of strong pink noise and babble noise. Compared to "plain" speech, the word recognition rate was most improved by pressed phonation, followed by an increased f_0 mean and f_0 range, and increased formant frequencies. Increased duration of vowels slightly reduced the recognition rate for pink noise but had no effect for babble noise.

Keywords: Lombard speech, speech intelligibility, articulatory study

1. Introduction

Lombard (1911) realized that humans involuntarily change their way of speaking in noisy conditions. This phenomenon is now called the "Lombard effect" or "Lombard speech" and has been shown to improve the intelligibility of speech in noise (Dreher and O'Neill, 1957, Pittman and Wiley, 2001, Lu and Cooke, 2008). The articulatory and acoustic changes underlying this phenomenon have been thoroughly studied. Compared with "plain" speech (a term adopted from Bradlow and Alexander (2007) to refer to "normal" speech produced in quiet conditions), Lombard speech mainly differs in terms of vocal intensity, spectral tilt, formant frequency, fundamental frequency (f_0), and the duration or speaking rate. Vocal intensity is usually increased in Lombard speech (Junqua, 1993, Summers et al., 1988). The spectral tilt of Lombard speech is normally flatter than for normal speech, i.e., there is more energy at higher frequencies (Davis et al., 2006). With regard to formant frequencies, multiple studies found a systematic increase of F_1 in Lombard speech (Junqua, 1993, Summers et al., 1988, Ngo et al., 2017, Uemura et al., 2010). Some of these studies also reported an increase of F_2, e.g., Uemura et al. (2010), but this increase was smaller and not as systematic as for F_1. With regard to f_0, both the f_0 mean and range increase with Lombard speech (Davis et al., 2006, Junqua, 1993, Uemura et al., 2010). Finally, Junqua (1993) found that, in Lombard speech, the duration of vowels is significantly increased and the duration of consonants is slightly decreased. This leads to an overall increase of word durations and hence a lower speaking rate. Most of these features (spectral tilt, formant frequencies, f_0, duration) were shown to vary continuously with the background noise level (Ngo et al., 2017).

The reasons for the acoustic changes with increasing background noise level are corresponding articulatory changes. Lombard speech is generally hyperarticulated, i.e., the spatial extent and the velocity of tongue, jaw and lip movements are increased (Garnier et al., 2006, Garnier, 2008, Huber and Chandrasekaran, 2006, Simko et al., 2016). The consequence is that the tongue position of vowels in Lombard speech is on average lower than during plain speech (Garnier et al., 2012, Scobbie et al., 2012). Given the general inverse relationship between tongue height and F_1, this explains the increase of F_1 in Lombard speech. Garnier et al. (2006) and Garnier et al. (2018) demonstrated a correlation of the extent of tongue and lip movements not only with F_1 but also with F_2 and f_0. The flattening of the spectral tilt in Lombard speech is most likely explained by a change of the phonation type. According to Stevens (2000), the spectral tilt increases (flattens) by about 6 dB/oct when phonation changes from modal to pressed. With regard to glottal articulation, a more pressed voice quality is achieved by a stronger adduction of the vocal folds.

All the studies mentioned above essentially analyzed the articulatory and acoustic features of naturally produced Lom-

*Corresponding author
Email addresses: vanthuanngo@jaist.ac.jp (Thuanvan Ngo), akagi@jaist.ac.jp (Masato Akagi), peter.birkholz@tu-dresden.de (Peter Birkholz)

bard speech. However, it is also of great interest to learn which of these features contribute to what extent to the enhanced intelligibility of Lombard speech. This knowledge can help in developing suitable methods for synthesizing more intelligible synthetic speech (Raitio et al., 2014, Langner and Black, 2005, Valentini-Botinhao et al., 2012) or modifying natural speech to make it more intelligible (Cooke et al., 2019).

Currently, there are only few studies that clarified the potential intelligibility benefit of typical features of Lombard speech. Lu and Cooke (2009) analyzed to what extent an increased f_0 and a flattened spectral tilt contribute to enhanced intelligibility in noise. With natural speech recordings as the basis, they used the vocoder STRAIGHT (Kawahara et al., 1999) to increase f_0 and a digital filter with a specific magnitude response to flatten the spectral tilt. They found that a flattened spectral tilt had a strong positive effect on the intelligibility, while an increase of mean f_0 had no significant effect. In a similar way, Cooke et al. (2014) analyzed the effect of increased phone durations (besides a flattened spectral tilt) on the recognition of speech in noise. To modify the duration of natural basis material, the PSOLA algorithm implemented in Praat (Boersma and Weenink, 2009) was used. However, no beneficial effects of durational increases were found. In a later study, Cooke and Aubanel (2017) found that increasing durations may still have a positive effect on the intelligibility but only when the background noise is fluctuating (as opposed to stationary). Common to the pioneering studies by Lu and Cooke (2009), Cooke et al. (2014) and Cooke and Aubanel (2017) is that the acoustic features of interest were modified at the *acoustic* level on the basis of natural speech recordings. While this is effective for the manipulation of the features f_0, spectral tilt, and duration, it would be more difficult to explicitly modify individual formants in natural recordings. Furthermore, acoustic manipulations are not explicitly related to articulatory and physical mechanisms.

Therefore, the goal of the present study was to investigate the effects of individual *articulatory* features and their combinations on enhancing the intelligibility of speech in noise. To this end, we used an enhanced version of the articulatory speech synthesizer VocalTractLab (Birkholz, 2017, 2013) to synthesize ten German words for digits in multiple variants that differ with respect to f_0 (mean and range), phonation type, formant frequency, and duration. In a perception experiment, the intelligibility benefit of these features for pink noise and babble noise was evaluated.

2. Method

Using the articulatory speech synthesizer VocalTractLab, each of the ten German words for the digits "0" to "9" (0 /nʊl/, 1 /aɛ̯ns/, 2 /tsv̥aɛ̯/, 3 /dʁaɛ̯/, 4, /fiːᵊ/, 5 /fʏnf/, 6 /zɛks/, 7 /zˈiːbn̩/, 8 /axt/, 9 /nɔœ̯n/) was synthesized in 16 variants. The 16 variants represented all combinations of four binary features, namely, phonation type, formant frequency, f_0, and duration. Each feature had two possible settings, one typical for plain speech and one typical for Lombard speech. Hence, for each digit word, there was one variant with all features of plain speech, one variant with all features of Lombard speech, and 14 variants with a mixture of features typical for plain speech and Lombard speech. To analyze the potential intelligibility benefit of the different feature combinations, a group of listeners was asked to identify the digit words in the presence of pink noise and in the presence of babble noise. In the following, we present in more detail the articulatory speech synthesizer, the creation of the stimuli, and the procedure of the perception experiment.

2.1. Articulatory speech synthesizer VocalTractLab

VocalTractLab (VTL) is an articulatory speech synthesizer that is capable of generating a full range of speech sounds in high quality while providing full control of time-varying glottal and supraglottal articulation. Supraglottal articulation is modeled by means of a 3D geometric model of the vocal tract of an adult male speaker (Birkholz, 2013). This model is controlled by 23 parameters that specify the shape and position of the articulators. The glottis is modeled by means of a geometric vocal fold model (Birkholz et al., 2019), which is a recent extension of VTL 2.2 and allows more precise control of the glottal geometry than the self-oscillating bar-mass model used in previous studies (Birkholz et al., 2011). The vocal fold model is controlled by ten parameters, which specify subglottal pressure, fundamental frequency, the shape of the glottis at rest, and oscillatory features such as the phase lag between the lower and upper vocal fold edges and the skewness of the glottal area pulses. For the synthesis of speech, the models of the vocal tract and the vocal folds are transformed into a unified 1D tube model of the vocal system (also including the subglottal system and the nasal cavity). This tube model is the basis of an aeroacoustic simulation in the time domain (Birkholz and Jackèl, 2004, Birkholz and Pape, 2019).

The parameters of the vocal tract and vocal fold models are controlled by means of a gestural score (similar to a musical score) (Birkholz, 2007), which is a high-level concept for speech movement control based on the ideas of articulatory phonology (Browman and Goldstein, 1992). In these scores, the articulatory gestures required to generate an utterance are specified and temporally coordinated. In VTL, gestural scores are created by means of a graphical editor as shown in Figure 1. The score has eight tiers, five of which define the supraglottal gestures (vowel gestures, lip gestures, tongue tip gestures, tongue body gestures, and velic gestures) and three of which define the laryngeal and pulmonary gestures (glottal shape gestures, f_0 gestures, and lung pressure gestures). As an example, Figure 1 shows the temporal coordination of the gestures required for the German word "acht" (/axt/, engl.: eight).

2.2. Creation of stimuli

The stimuli for the perception experiment were created in three steps:

(1) For each German digit word, a recorded natural utterance of that word spoken with a "plain" speaking style was

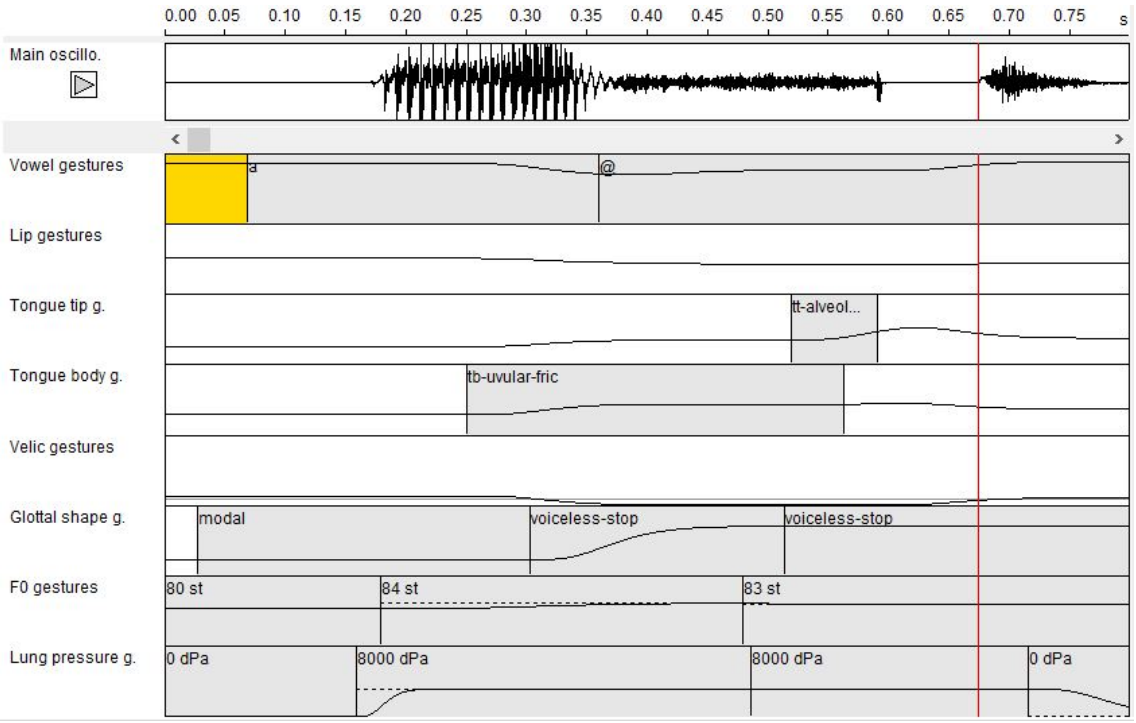


Figure 1: Gestural score and synthesized waveform for the plain speech variant of the German word for digit 8 (/axt/).

resynthesized in terms of a gestural score similar to Birkholz et al. (2017). In the resynthesized utterances, the phone durations and the f_0 contours were closely fitted to those of the natural utterances. The exact acoustic realization of the individual phones was determined by the corresponding predefined settings (shapes) of the vocal tract and vocal fold models. For all ten words, modal phonation and a subglottal pressure of 800 Pa was used. As a result, the resynthesized words had all typical features of plain speech.

(2) The gestural scores created in (1) were used to generate the remaining 15 variants of each digit word by changing the phonation type, f_0, formant frequencies, and phone duration (either individually or jointly) to a setting typical for Lombard speech. How exactly the features were adjusted is detailed in Secs. 2.2.1 to 2.2.4. Table 1 gives an overview of the plain speech settings and the Lombard speech settings for the four features. All speech items were synthesized as 16-bit mono signals with a sampling frequency of 22,050 Hz. The amplitude of the synthetic items was *not* normalized so that the inherent amplitude differences between the items due to the different feature settings (e.g., modal vs. pressed voice) were maintained.

(3) All 160 speech items (10 digit words × 16 variants) were combined with two types of noise as detailed in Sec. 2.2.5 to create the stimuli for the perception experiment.

2.2.1. Adjustment of phonation type

In Lombard speech, the spectral tilt is flatter than in plain speech, i.e., the higher-frequency components are enhanced. To cause spectral flattening for the synthetic words, the parameters of the vocal fold model (Birkholz et al., 2019) were adjusted to generate a more pressed voice quality. The main vocal fold model parameters that affect the voice quality on the continuum from a modal to a pressed voice are the (pre-phonatory) rest displacement x_{rest} of the vocal folds at the level of the arytenoids, the area A_{chink} of a permanent glottal chink between the arytenoids, and the subglottal pressure P_{sub}. The settings for modal phonation (for plain speech) were the "standard" values $x_{\text{rest}} = 0.3$ mm, $A_{\text{chink}} = 2$ mm^2, and $P_{\text{sub}} = 800$ Pa. In contrast, pressed phonation was generated with $x_{\text{rest}} = 0$ mm, $A_{\text{chink}} = 0$ mm^2, and $P_{\text{sub}} = 1600$ Pa, i.e., with no glottal rest area and twice the subglottal pressure used for modal phonation. In the absence of any published measurements of subglottal pressure during Lombard speech, the value of 1600 Pa was chosen to be clearly higher than that of plain speech to reflect the higher vocal effort, but also clearly below the maximum lung pressures of around 6 kPa that humans can produce in extreme situations (Titze, 1994).

With these settings, the average spectral tilt across all ten of the digit words (as approximated by a regression line to the long-term average spectral magnitude in dB between 0 and 4 kHz) was -9.23 dB/oct for the modal voice and -3.55 dB/oct for the pressed voice. Hence, from modal to pressed voice, the spectral tilt increased by 5.68 dB/oct, which is close to the typical difference of 6 dB between these phonation types (Stevens, 2000). Furthermore, the average sound pressure level (SPL) of the digit words synthesized with pressed voice increased by 10.05 dB compared with modal voice, which is in the range of an 8 to 15-dB difference in SPL between plain and Lombard speech as observed by Kubo et al. (2016).

Table 1: Plain speech settings and Lombard speech settings of four examined features used for articulatory speech synthesis.

Feature	Setting for plain speech	Setting for Lombard speech
Phonation type	Modal voice & 800 Pa lung pressure	Pressed voice & 1600 Pa lung pressure
Formants	Standard formant values in VTL	F_1 increased by 25%, F_2 increased by 5%
Fundamental frequency	Reproduced from natural plain speech	f_0 mean increased by 5 st, f_0 range increased by the factor 1.3
Phone durations	Reproduced from natural plain speech	Durations of vowels increased by 30%

2.2.2. Adjustment of fundamental frequency

Compared to plain speech, Lombard speech is characterized by both an increased f_0 mean and range. The data of Ngo et al. (2017, Fig. 2b) indicate that for very high background noise levels, f_0 increases by about 5 st compared with plain speech. Furthermore, the data by Davis et al. (2006) show that the f_0 range of Lombard speech is 1.2 to 1.8 times the f_0 range of plain speech (on the Hz scale). Accordingly, to model the change in f_0 due to the Lombard effect in the synthesizer, the mean f_0 and the f_0 range of the reference gestural scores were increased by 5 st and a factor of 1.3, respectively. This change was implemented by modifying the f_0 target offsets of the target approximation model (Prom-on et al., 2009), which is the f_0 model used in VTL.

2.2.3. Adjustment of formant frequencies

Multiple studies found that Lombard speech has increased frequencies of F_1 and F_2 compared with plain speech. The values estimated from Ngo et al. (2017) and used here are a 25% increase of F_1 and a 5% increase of F_2. To implement a change of the formant frequencies in VTL, we had to adjust the vocal tract target shapes of the corresponding vowels. The target shapes of the vowels occurring in the ten digit words were adjusted in two steps. First, starting with the standard shapes of the vowels, the mouth cavity was opened more (as for shouting) by manually adjusting the vocal tract parameters of the jaw, the lips, and the tongue. The jaw angle was decreased by 3° (parameter JA), the distance between the lower and upper lips was increased by 5 mm (LD), and the tongue position was lowered by 5 mm (TTY, TBY, TCY). While this manual adjustment increased F_1 for all vowels, the increase was never exactly 25%, and the change of F_2 was rather unpredictable. Therefore, as the 2^{nd} step, a greedy optimization algorithm (Birkholz, 2013) was used to fine-tune all vocal tract model parameters in such a way that the resulting formant frequencies assumed the intended values. After optimization, the deviation of the formant frequencies from the intended values was 1.5% ± 1.5% across all vowels. As an example, Figure 2 shows the standard vocal tract shape (corresponding to plain speech) of the vowel /a/ in gray and the adjusted shape with F_1 increased by 25% and F_2 increased by 5% in black.

2.2.4. Adjustment of duration

In Lombard speech, vowels are on average longer, and consonants tend to be slightly shorter than in plain speech (Ngo et al., 2017, Junqua, 1993). According to Ngo et al. (2017) (Fig. 2a), the increase of vowel duration converges to about 30% for very high background noise levels. As this change is much greater than the change of consonantal durations, we chose here to model the durational changes due to the Lombard effect by increasing the durations of all vowels in the gestural scores by 30%. This was achieved by "stretching" the scores by the appropriate durations around the acoustic midpoints of the vowels contained in the utterances.

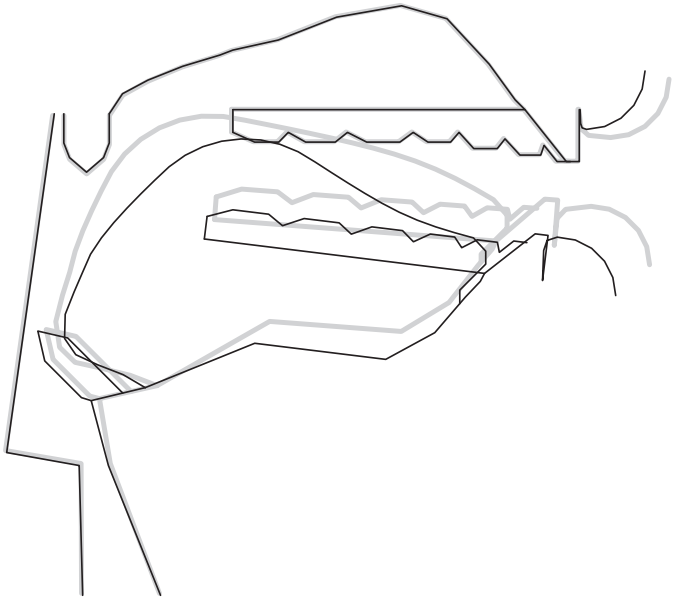


Figure 2: Midsagittal shapes of the vocal tract model for the vowel /a/ with normal (standard) articulation (gray lines) and Lombard articulation (black lines).

2.2.5. Addition of noise

As each German digit word was synthesized in 16 variants, there were 160 (clean) synthetic speech items in total. A second set of 160 items was generated by adding pink noise (Pink-Noise, 1984) to the clean speech items, and a third set of 160 items was generated by adding babble noise (Babble-Noise, 1990) to the clean speech items. Both types of noise are common in daily life and have a speech-like overall spectral shape (see Figure 3). Furthermore, they represent both a kind of stationary noise (pink noise) and a kind of non-stationary noise (babble noise). The babble noise was generated by 100 people speaking in a cafeteria with individual voices still slightly audible. For each kind of noise, the amplitude was chosen in such a way that the sound pressure level (SPL) of the noise was 20 dB higher than the average SPL of the ten synthesized digits in the plain speaking style. The SPL difference of 20 dB is equiva-

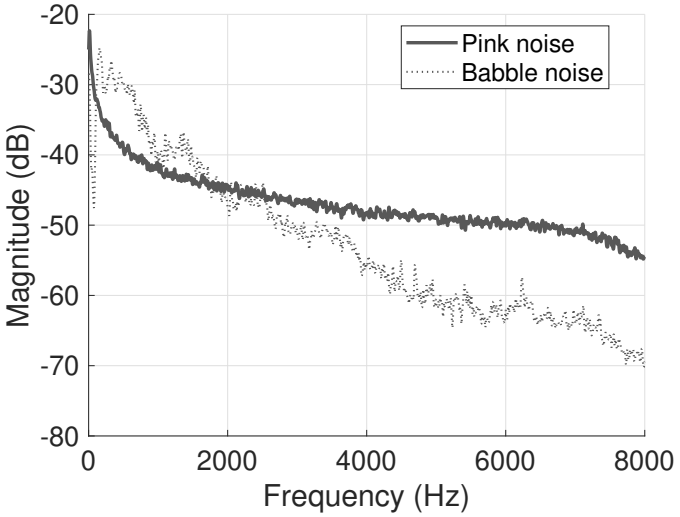


Figure 3: Long-term average spectra of additive noise from 0-8000 Hz with frequency resolution of 15 Hz.

lent to the 84 dB absolute noise level that was used to induce Lombard speech in the database of Ngo et al. (2017) and Kubo et al. (2016). For the perception experiment, the audio files of the speech items superimposed with noise had a total length of 2 s with the target words embedded in the middle.

2.3. Perception experiment

The perception experiment had two tasks. The first task was an evaluation of the naturalness of the synthetic utterances (without background noise), and the second task was a test for intelligibility. Seventeen native Germans, including 13 men with an average age of 32.5 years and a standard deviation of 9.8 years, and 4 women with an average age of 40.3 years and a standard deviation of 11.1 years, participated in the tests. All participants gave informed consent and reported no hearing problems. Each participant performed the two tasks in two consecutive sessions.

For the evaluation of the naturalness of the synthesis (first task), each participant listened to the 160 stimuli without added background noise in random order using high-quality headphones (AKG K240) connected to a desktop computer via a FireWire audio interface (MOTU 896HD) in a soundproof room. The volume for the headphones was adjusted to make all stimuli without added noise comfortably audible for the subjects. After each stimulus was played, the participants were asked to rate the naturalness among four options, 1 - unnatural, 2 - rather unnatural, 3 - rather natural, or 4 - natural, by clicking on one of four buttons with the respective labels. After choosing an answer to the current stimulus, the next stimulus was automatically played (repetitions were not possible). The session took about 12 minutes.

After a short break, the participants started the second session for the second task. In this task, each participant listened to all 480 stimuli (160 speech items in three conditions: pink noise, babble noise, and no noise) in random order by using the same equipment as in the first session. After each stimulus was played, the participants had to click on one of ten buttons on the computer screen that represented the perceived digit word. If they could not clearly understand the spoken digit, they were allowed to randomly choose one of the digits. After choosing the answer to the current stimulus, the next stimulus was automatically played (repetitions were not possible). Halfway through this session, the participants took a short break of 2 minutes. The whole session lasted about 35 minutes.

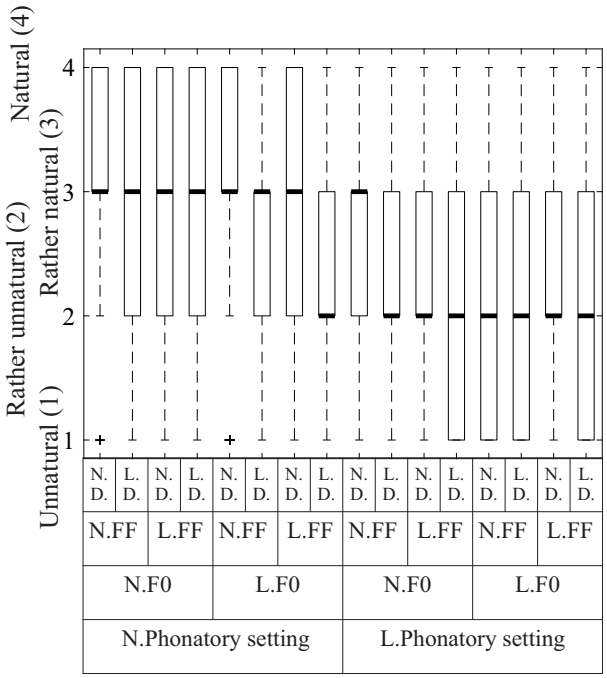


Figure 4: Box plots of naturalness ratings of all 16 word variants, i.e., feature combinations. The labels below boxes indicate feature combinations, where "N." stands for the neutral setting of a feature (as in plain speech), "L." stands for the Lombard setting of a feature, "D." stands for duration, "FF" for formant frequency, and "F0" for fundamental frequency.

3. Results and discussion

3.1. Perceptual test of naturalness

The perceptual ratings of the naturalness of the synthetic stimuli are shown in Figure 4, with one boxplot for each of the 16 feature combinations. The leftmost boxplot represents the ratings of the stimuli with all features of plain speech (median = 3), and the rightmost boxplot represents the ratings of the stimuli with all features of Lombard speech (median = 2). As can be seen, the feature combinations affected the ratings, and the stimuli with the settings for plain speech were among the most natural sounding items.

To test the effect of individual features on the naturalness ratings, we formed five groups of stimuli (see Table 2). Four two-tailed Mann-Whitney U tests were performed to compare the response distributions of groups A vs. B, A vs. C, A vs. D, and A vs. E. The response distributions of groups A and B differed significantly (Mann-Whitney $U = 9739.0$, $N1 = N2 = 170$, $p < 0.001$), i.e., the stimuli with pressed phonation were perceived to be more unnatural than the stimuli with modal phonation. The response distributions of groups A and C did

Table 2: Feature combinations of the groups of stimuli that were compared with respect to the perceived naturalness of the stimuli. The stimuli in group A represent plain speech. The stimuli in the groups B, C, D, and E differ in one feature each from group A.

Group	Type of stimuli
A	Modal phonation, normal f_0, normal formants, and normal durations.
B	**Pressed phonation**, normal f_0, normal formants, and normal durations.
C	Modal phonation, **Lombard f_0**, normal formants, and normal durations.
D	Modal phonation, normal f_0, **Lombard formants** and normal durations.
E	Modal phonation, normal f_0, normal formants, and **Lombard durations**.

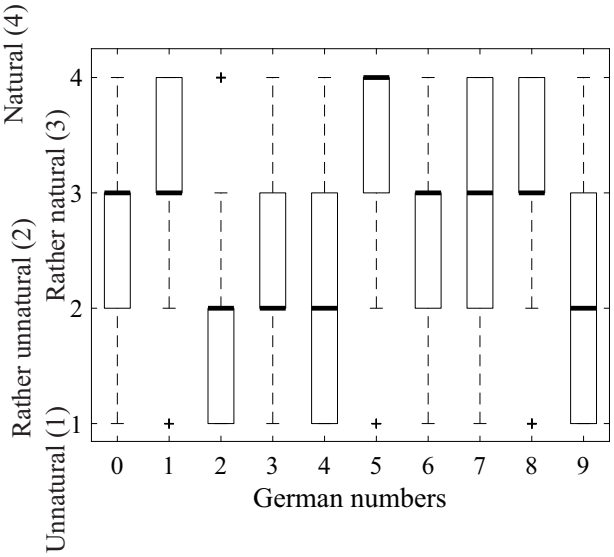


Figure 5: Naturalness ratings of individual digit words, pooled across all 16 variants.

not differ significantly (Mann-Whitney $U = 13393.0$, $N1 = N2 = 170$, $p > 0.05$), i.e., the f_0 setting had no effect on the naturalness. The response distributions of groups A and D (Mann-Whitney $U = 12494.5$, $N1 = N2 = 170$, $p < 0.022$) and groups A and E (Mann-Whitney $U = 12537.5$, $N1 = N2 = 170$, $p < 0.025$) were both significantly different; hence, the Lombard-typical settings of the formants and durations also caused a slight decrease of the naturalness.

The decreased naturalness of the stimuli with the Lombard-typical phonation type, formants, and durations may be in part due to the fact that the raters listened to the stimuli in the absence of noise, while they are normally used to perceive Lombard-typical features under noisy conditions. Another reason may be a non-perfect simulation of the according features.

Figure 5 shows how the naturalness varied across the individual digit words. According to the median response values, six of the ten words were perceived as "natural" or "rather natural", while four words were perceived as "rather unnatural". The reasons for the ratings of the words for 2, 3, 4, and 9 as rather unnatural are hard to tell. A retrospective informal comparison of the synthetic words indicated that the modeled f_0 contour of the four words with the lower ratings might have been somewhat atypical.

3.2. Perceptual test of intelligibility

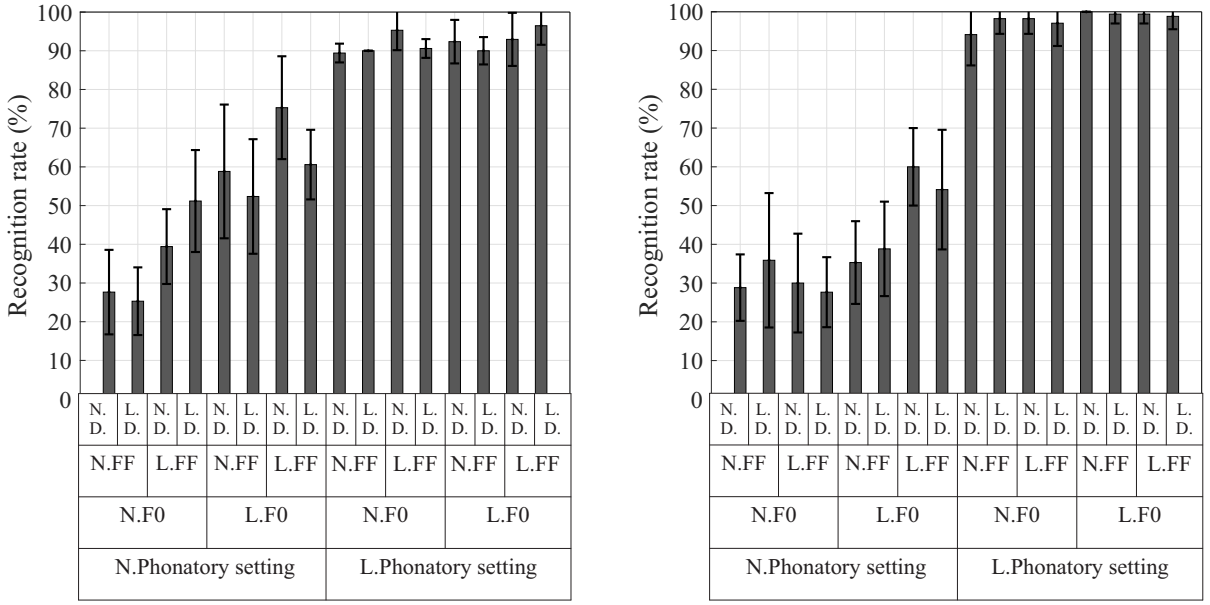
The results of the intelligibility test are shown in Figures 6 and 7 in terms of recognition rates. Figure 6 shows how the 16 different feature combinations affected the recognition rates in the presence of pink noise and babble noise, respectively. The recognition rates of the digit words without background noise are not explicitly shown as they were very close to 100%. In general, the more Lombard-typical features the stimuli contained under the noisy conditions, the higher the recognition rates.

To study the effects of the four features on the recognition rate in more detail, a four-way repeated measures ANOVA was performed. The features phonation type, f_0, formant, and duration were the four factors, each having two levels (the normal setting and the Lombard-typical setting). The dependent factor was the recognition rate. Two individual ANOVAs were performed, one for the case of pink background noise and one for the case of babble background noise.

In the case of *pink noise*, there were significant main effects for all four factors, that is, phonation type [$F(1, 16) = 724.5$, $p < 0.001$], f_0 [$F(1, 16) = 101.5$, $p < 0.001$], formant [$F(1, 16) = 82.3$, $p < 0.001$], and duration [$F(1, 16) = 4.7$, $p = 0.045$]. However, while the Lombard settings of phonation type, f_0, and formant had a *positive* effect on the recognition rate, the Lombard-typical (i.e. longer) durations *reduced* the mean recognition rate (from a mean of 71.40% across all samples with normal vowel durations to a mean of 69.56% across all samples with increased vowel durations). The effect size was strongest for phonation type ($\eta^2 = 0.978$), followed by f_0 ($\eta^2 = 0.864$), formant ($\eta^2 = 0.837$), and duration ($\eta^2 = 0.229$). In addition, there were multiple significant interactions between factors, namely, between phonation type and any of the f_0, formants, and durations, between f_0 and duration, and between phonation type, formant, f_0, and duration. Among these, the interaction between phonation type and f_0 was strongest ($F(1, 16) = 108.0$, $p < 0.001$).

In the case of *babble noise*, there were significant main effects for three of the factors, that is, phonation type [$F(1, 16) = 1284.8$, $p < 0.001$], f_0 [$F(1, 16) = 124.4$, $p < 0.005$], and formant [$F(1, 16) = 17.8$, $p = 0.001$], i.e., their Lombard-typical settings increased the recognition rate. However, in contrast to the pink noise case, there was no significant effect for the factor duration ($p > 0.5$). The effect size was strongest for phonation type ($\eta^2 = 0.988$), followed by f_0 ($\eta^2 = 0.886$) and formant ($\eta^2 = 0.527$). In addition, there were significant interactions between phonation type and formant, between phonation type and f_0, between formant and duration, and between f_0 and phonation type. Among these, the interaction between phonation type and f_0 was strongest ($F(1, 16) = 75.4$, $p < 0.001$).

Figure 7 shows the recognition rates separated by the digit words and the noise conditions. It illustrates that the digit words were almost perfectly recognized without background noise and



(a) Pink noise

(b) Babble noise

Figure 6: Recognition rates for all 16 feature combinations in presence of pink noise and babble noise, pooled across all digit words and listeners. The labels indicate feature combinations, where "N." stands for the neutral setting of a feature (as in plain speech), "L." stands for the Lombard setting of a feature, "D." stands for duration, "FF" for formant frequency, and "F0" for fundamental frequency.

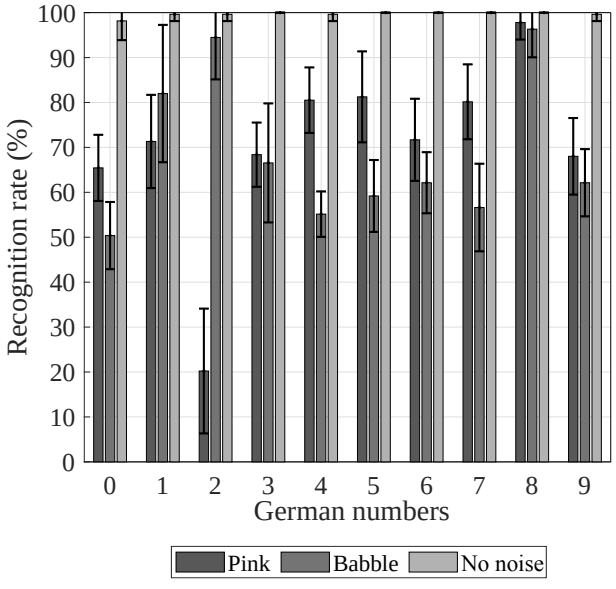


Figure 7: Recognition rates of individual German digit words across all speech variants, separated by noise conditions: pink noise, babble noise, and no noise. Bar heights indicate mean values, and error bars indicate standard deviations.

that the recognition rates were generally higher in the presence of pink noise than in babble noise. A notable exception is the word /tsy͡aɛ̯/ (digit 2), which was badly recognized in pink noise. The reason is probably that the initial consonant cluster /tsy/ mainly consists of wideband noise which is hard to perceive in pink noise due to similar noise characteristics.

In summary, we found that the change of phonation type from a modal to a pressed voice improved the intelligibility of the words most, independently from the type of background noise. Given that the effect of a more pressed voice quality is a flattening of the spectral tilt, this finding is in line with the previous findings of Lu and Cooke (2009) and Cooke et al. (2014). However, we also found an increase of f_0 to be highly effective at increasing the intelligibility in both types of noise, which contradicts the findings by Lu and Cooke (2009). The reason may be that the greatest mean f_0 increase examined by Lu and Cooke (2009) was only 2.5 st (from 148 Hz to 171 Hz), while we used 5 st based on the data by Ngo et al. (2017). The present study also proved that there was a beneficial effect of an increase of the first two formant frequencies of vowels, which had so far not been explicitly shown in analysis-by-synthesis experiments. Finally we found that an increase of the duration of voiced sounds did not improve word intelligibility in noise. In fact, in pink noise, the durational increase even led to slightly worse intelligibility. However, what is the reason that increased durations are frequently observed in natural Lombard speech then? One reason is probably the wider extent of the articulatory movements in Lombard speech (e.g., generally lower tongue positions in vowels), which takes more time. Another

reason is probably that longer phone durations *can* improve the intelligibility in noise but only for certain types of fluctuating noise (Cooke and Aubanel, 2017).

The main limitations of the present study are the following:

1. Only two values per feature were analyzed (one value for plain speech and one value typical for Lombard speech). This led to a ceiling effect of the recognition rates for certain feature combinations. For example, according to Figure 6b, the recognition rate was almost 100% for all stimuli with pressed phonation in the presence of babble noise, independently from the other feature values. Future studies could investigate the intelligibility benefit of multiple values in smaller steps along each feature dimension in more detail, or use multiple different levels of additive noise.

2. The speech material was limited to the ten German digit words, so the results may not directly translate to longer utterances or different languages. However, the used words do contain 8 different vowels and 11 different consonants, which cover roughly half of the German phonemes. Therefore we would expect similar results for languages with a phoneme system similar to German. For longer utterances, e.g., sentences, we would generally expect better recognition rates, because more context helps to disambiguate individual words that are strongly masked by noise.

4. Conclusion

The present study used articulatory speech synthesis to generate synthetic words with different combinations of articulatory-acoustic features and explored their individual and combined effects on the intelligibility of the words in pink noise and babble noise. It was found that using a pressed voice quality (i.e., flattening the spectral tilt), increasing f_0, and increasing F_1 and F_2 all enhance the intelligibility to different degrees. Furthermore, the beneficial effect of these features is generally additive, e.g., increasing both f_0 and formant frequencies improves the intelligibility more than either feature alone. However, increasing vowel durations has no positive effect. These results suggest how to adapt synthetic speech to varying background noise conditions such that a generated utterance always remains intelligible. While the results can be most directly applied to parametric forms of speech synthesis, they can also be adapted to enhance the intelligibility of other types of synthesis, such as unit-selection synthesis.

5. Acknowledgments

This research was supported by an off-campus research grant by Japan Advanced Institute of Science and Technology, SECOM Science and Technology foundation and JST-Mirai Program (JPMJMI18D1) and the Technische Universität Dresden.

References

Babble-Noise. *Noisex*. NOISE-ROM-0, NATO: AC243/(Panel 3)/RSG10, 1990.

P. Birkholz. Modeling consonant-vowel coarticulation for articulatory speech synthesis. *PLoS ONE*, 8(4):e60603, 2013.

P. Birkholz. VocalTractLab [software], 2017. URL `http://www.vocaltractlab.de`.

P. Birkholz, L. Martin, Y. Xu, S. Scherbaum, and C. Neuschaefer-Rube. Manipulation of the prosodic features of vocal tract length, nasality and articulatory precision using articulatory synthesis. *Computer Speech & Language*, 41: 116–127, 2017.

Peter Birkholz. Control of an articulatory speech synthesizer based on dynamic approximation of spatial articulatory targets. In *Interspeech 2007 - Eurospeech*, pages 2865–2868, Antwerp, Belgium, 2007.

Peter Birkholz and Dietmar Jackèl. Influence of temporal discretization schemes on formant frequencies and bandwidths in time domain simulations of the vocal tract system. In *Interspeech 2004-ICSLP*, pages 1125–1128, Jeju, Korea, 2004.

Peter Birkholz and Daniel Pape. How modeling entrance loss and flow separation in a two-mass model affects the oscillation and synthesis quality. *Speech Communication*, 110:108–116, 2019.

Peter Birkholz, Bernd J. Kröger, and Christiane Neuschaefer-Rube. Synthesis of breathy, normal, and pressed phonation using a two-mass model with a triangular glottis. In *Interspeech 2011*, pages 2681–2684, Florence, Italy, 2011.

Peter Birkholz, Susanne Drechsel, and Simon Stone. Perceptual optimization of an enhanced geometric vocal fold model for articulatory speech synthesis. In *Interspeech 2019*, pages 3765–3769, Graz, Austria, 2019.

Paul Boersma and David Weenink. Praat: doing phonetics by computer (version 5.1.13), 2009. URL `http://www.praat.org`.

Ann R Bradlow and Jennifer A Alexander. Semantic and phonetic enhancements for speech-in-noise recognition by native and non-native listeners. *The Journal of the Acoustical Society of America*, 121(4):2339–2349, 2007.

Catherine P. Browman and Louis Goldstein. Articulatory phonology: An overview. *Phonetica*, 49:155–180, 1992.

M. Cooke and V. Aubanel. Effects of linear and nonlinear speech rate changes on speech intelligibility in stationary and fluctuating maskers. *The Journal of the Acoustical Society of America*, 141:4126–4135, 2017.

M. Cooke, C. Mayo, and J. Villegas. The contribution of durational and spectral changes to the lombard speech intelligibility benefit. *The Journal of the Acoustical Society of America*, 135(2):874–883, 2014.

Martin Cooke, Vincent Aubanel, and Mara Luisa Garca Lecumberri. Combining spectral and temporal modification techniques for speech intelligibility enhancement. *Computer Speech & Language*, 55:26 – 39, 2019. ISSN 0885-2308. doi: https://doi.org/10.1016/j.csl.2018.10.003. URL `http://www.sciencedirect.com/science/article/pii/S0885230818300676`.

C. Davis, J. Kim, K. Grauwinkel, and H. Mixdorff. Lombard speech: Auditory (a), visual (v) and av effects. In *Proceedings of the Third International Conference on Speech Prosody*, pages 248–252, 2006.

John J Dreher and John O'Neill. Effects of ambient noise on speaker intelligibility for words and phrases. *The Journal of the Acoustical Society of America*, 29(12):1320–1323, 1957.

M. Garnier. May speech modifications in noise contribute to enhance audio-visible cues to segment perception? In *AVSP*, pages 95–100, 2008.

M. Garnier, L. Bailly, M. Dohen, P. Welby, and H. Loevenbruck. An acoustic and articulatory study of lombard speech: Global effects on the utterance. In *Interspeech/ICSLP 2006*, pages 2246–2249, Pittsburgh, United States, 2006.

M. Garnier, L. Mnard, and G. Richard. Effect of being seen on the production of visible speech cues. a pilot study on lombard speech. In *13th Annual Conference of the International Speech Communication Association (InterSpeech 2012)*, pages 611–614, Portland, United States, 2012.

M. Garnier, L. Mnard, and B. Alexandre. Hyper-articulation in lombard speech: An active communicative strategy to enhance visible speech cues? *The Journal of the Acoustical Society of America*, 144:1509–1074, 2018.

J. E. Huber and B. Chandrasekaran. Effects of increasing sound pressure level on lip and jaw movement parameters and consistency in young adults. *J. Speech Language Hearing Res.*, 49:1368–1379, 2006.

J. C. Junqua. The lombard reflex and its role on human listeners and automatic speech recognizers. *The Journal of the Acoustical Society of America*, 93: 510–524, 1993.

Hideki Kawahara, Ikuyo Masuda-Katsuse, and Alain de Cheveign. Restructuring speech representations using a pitch-adaptive timefrequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication*, 27(3):187 – 207, 1999. ISSN 0167-6393. doi: https://doi.org/10.1016/S0167-6393(98)00085-5. URL http://www.sciencedirect.com/science/article/pii/S0167639398000855.

R. Kubo, D. Morikawa, and M. Akagi. Effects of speaker's and listener's acoustic environments on speech intelligibility and annoyance. In *Inter-Noise 2016*, pages 171–176, Hamburg, Germany, 2016.

Brian Langner and Alan W Black. Improving the understandability of speech synthesis by modeling speech in noise. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05)*, volume 1, pages I–265, 2005.

E. Lombard. Le signe de l'lvation de la voix. *Annales des Maladies de L'Oreille et du Larynx*, 37:101–119, 1911.

Y. Lu and M. Cooke. The contribution of changes in f0 and spectral tilt to increased intelligibility of speech produced in noise. *Speech Communication*, 51:1253–1262, 2009.

Youyi Lu and Martin Cooke. Speech production modifications produced by competing talkers, babble, and stationary noise. *The Journal of the Acoustical Society of America*, 124(5):3261–3275, 2008.

T. V. Ngo, R. Kubo, D. Morikawa, and M. Akagi. Acoustical analyses of tendencies of intelligibility in lombard speech with different background noise levels. *Journal of Signal Processing*, 21:171–174, 2017.

Pink-Noise. Various - audio test CD-1 - 91 test signals for home and laboratory use, 1984. URL https://www.discogs.com/.

Andrea L Pittman and Terry L Wiley. Recognition of speech produced in noise. *Journal of Speech, Language, and Hearing Research*, 2001.

Santhitam Prom-on, Yi Xu, and Bundit Thipakorn. Modeling tone and intonation in Mandarin and English as a process of target approximation. 125(1): 405–424, 2009.

Tuomo Raitio, Antti Suni, Martti Vainio, and Paavo Alku. Synthesis and perception of breathy, normal, and lombard speech in the presence of noise. *Computer Speech & Language*, 28(2):648 – 664, 2014. ISSN 0885-2308. doi: https://doi.org/10.1016/j.csl.2013.03.003. URL http://www.sciencedirect.com/science/article/pii/S0885230813000193.

J. Scobbie, J. Ma, and J. White. The tongue and lips in lombard speech: A pilot study of vowel-space expansion. Casl, 2012.

J. Simko, S. Benus, and M. Vainio. Hyperarticulation in lombard speech: Global coordination of the jaw, lips and the tongue. *The Journal of the Acoustical Society of America*, 139:151–162, 2016.

K. N. Stevens. *Acoustic Phonetics*. The MIT Press, 2000.

W. V. Summers, D. B. Pison, R. H. Bernacki, R. I. Pedlow, , and M. A. Stokes. Effects of noise on speech production: Acoustic and perceptual analyses. *The Journal of the Acoustical Society of America*, 84:917–928, 1988.

Ingo R. Titze. *Principles of Voice Production*. Prentice Hall, 1994.

Y. Uemura, M. Morise, and T. Nishiura. The lombard speech recognition based on the voice conversion towards neutral speech. *ICA2010*, 167, 2010.

Cassia Valentini-Botinhao, Junichi Yamagishi, and Simon King. Evaluating speech intelligibility enhancement for hmm-based synthetic speech in noise. In *SAPA-SCALE Conference*, 2012.