The relationship between talker acoustics, intelligibility, and effort in degraded listening conditions

Maximillian Paulus, Valerie Hazan, and Patti Adank

Citation: The Journal of the Acoustical Society of America **147**, 3348 (2020); doi: 10.1121/10.0001212

View online: https://doi.org/10.1121/10.0001212

View Table of Contents: https://asa.scitation.org/toc/jas/147/5

Published by the Acoustical Society of America

JASA ARTICLE

The relationship between talker acoustics, intelligibility, and effort in degraded listening conditions

Maximillian Paulus,[a] Valerie Hazan, and Patti Adank
Speech, Hearing and Phonetic Sciences, University College London, London, United Kingdom

ABSTRACT:
Listening to degraded speech is associated with decreased intelligibility and increased effort. However, listeners are generally able to adapt to certain types of degradations. While intelligibility of degraded speech is modulated by talker acoustics, it is unclear whether talker acoustics also affect effort and adaptation. Moreover, it has been demonstrated that talker differences are preserved across spectral degradations, but it is not known whether this effect extends to temporal degradations and which acoustic-phonetic characteristics are responsible. In a listening experiment combined with pupillometry, participants were presented with speech in quiet as well as in masking noise, time-compressed, and noise-vocoded speech by 16 Southern British English speakers. Results showed that intelligibility, but not adaptation, was modulated by talker acoustics. Talkers who were more intelligible under noise-vocoding were also more intelligible under masking and time-compression. This effect was linked to acoustic-phonetic profiles with greater vowel space dispersion (VSD) and energy in mid-range frequencies, as well as slower speaking rate. While pupil dilation indicated increasing effort with decreasing intelligibility, this study also linked reduced effort in quiet to talkers with greater VSD. The results emphasize the relevance of talker acoustics for intelligibility and effort in degraded listening conditions. © *2020 Acoustical Society of America*. https://doi.org/10.1121/10.0001212

(Received 10 September 2019; revised 14 April 2020; accepted 20 April 2020; published online 8 May 2020)

[Editor: Tessa C. Bent] Pages: 3348–3359

I. INTRODUCTION

Everyday listening situations are marked by acoustic degradations that can reduce intelligibility and impose higher cognitive demands on the listener, causing discomfort, stress, and fatigue (Pichora-Fuller *et al.*, 2016). Acoustic degradations of the speech signal can be classified into (transmission) channel degradations such as background noise, and source degradations such as anatomical-physiological differences between talkers (Mattys *et al.*, 2012). These two types of degradations are also known to interact: even though speech produced by native speakers is intelligible and perceived effortlessly in optimal listening conditions (McLaughlin and Van Engen, 2020), acoustic-phonetic differences between talkers modulate intelligibility when speech is degraded (Bent *et al.*, 2009; Hazan and Markham, 2004). Listeners have cognitive strategies to effectively deal with degraded speech, including fast adaptation to temporal and spectral degradations such as time-compressed speech or noise vocoding (Davis *et al.*, 2005; Dupoux and Green, 1997). It is currently not clear whether and how acoustic-phonetic profiles associated with anatomical-physiological differences between talkers interact with different spectral and temporal channel degradations. Our study therefore aimed to investigate how combinations of such source and channel degradations affect both intelligibility and listening effort.

A. Source and channel degradations

Acoustic variations between talkers arise due to accent differences (Bradlow and Bent, 2008), but also due to idiosyncratic and anatomical-physiological differences (Hazan and Markham, 2004). Idiosyncratic features such as vowel space, energy in speech-critical bands, and speaking rate (SR) predict intelligibility in quiet or in noise (Bradlow *et al.*, 1996; Hazan and Markham, 2004). An increased vowel space is linked to more precise articulation with a slow speaking style (Hazan and Markham, 2004; Hazan *et al.*, 2018). However, the intelligibility benefit of an increased vowel space can be independent of a talker's SR (Bradlow *et al.*, 1996). It is unclear which combination of acoustic-phonetic features forms an acoustic talker profile that is optimally intelligible in changing listening conditions (Bent *et al.*, 2009). Similar intelligibility differences have been measured for the same talkers, irrespective of whether their speech was undegraded, masked by different types of noise, or passed through cochlear implant speech processors or a simulation thereof (Bent *et al.*, 2009; Green *et al.*, 2007; Hochmuth *et al.*, 2015). A cochlear implant speech processor is usually simulated using the technique of noise-vocoding whereby spectral and temporal details of the source signal are removed and temporal envelope cues are preserved (Shannon *et al.*, 1995). Masking noise and noise-vocoding degrade speech in different ways, and it is therefore conceivable that combinations of acoustic-phonetic features are responsible for preserved talker differences. Green *et al.* (2007) suggested that cues such as amplitude

[a] Electronic mail: m.paulus@ucl.ac.uk

envelope and gross spectral differences that are mainly temporal promote intelligibility since temporal fine structure cues are removed by the speech processor in cochlear implants. In their study, word duration and mean energy in the 1–3 kHz range were both found to be positively correlated with intelligibility. As both acoustic-phonetic measures were also inter-correlated, that might explain why some talkers were more intelligible in both conditions: increased energy in the 1–3 kHz range benefited word recognition in noise while longer word duration benefited word recognition with (simulated) cochlear implant speech processor. However, the set of talkers was a small subset ($N = 6$) taken from an earlier study ($N = 45$) that did not observe this inter-correlation of energy and word duration (Hazan and Markham, 2004).

B. Listening effort and pupillometry

Listening to degraded speech is not only associated with decreased intelligibility but also increased effort. Theoretical frameworks have been proposed to explain the relationship between input demands, motivation, and effort (Pichora-Fuller *et al.*, 2016). Following the definitions by Mattys *et al.* (2012), input demands have been divided into channel and source factors (Pichora-Fuller *et al.*, 2016) and have usually been investigated separately. Listening effort is often quantified using self-report measures. However, individual differences in what is perceived as effortful pose a fundamental challenge for such measures (McGarrigle *et al.*, 2014). An objective approach to quantify listening effort is to measure the task-evoked pupil response (Winn *et al.*, 2018), typically parameterized in peak or mean dilation and peak latency. Larger pupil dilation has been associated with more severe channel degradations such as decreases in the signal-to-noise ratio (Wendt *et al.*, 2018). Even though pupil dilation can be associated with intelligibility, studies have shown masking effects at fixed levels of intelligibility (Wendt *et al.*, 2018). Specifically, speech presented in a competing-speaker background elicited larger pupil dilation than speech presented in noise. With regard to source degradations, both the effects of accented speech and individual acoustic-phonetic differences have been investigated. McLaughlin and Van Engen (2020) observed increased pupil dilation for accented speech over native speech at high levels of intelligibility. Koch and Janse (2016) investigated the effect of SR on listening effort in quiet. Despite increased response times for faster speech, no difference in pupil dilation was found. It is possible that these differences only emerge when speech is perceived in background noise; as channel and source degradations interact, the overall difficulty of the task increases. A recent study found increased intelligibility and decreased pupil dilation in noise for clear speech, i.e., speech produced in noise, when compared to speech produced in quiet (Simantiraki *et al.*, 2018). Apart from this task-related "phasic" component of the pupil dilation, the "tonic" baseline pupil size, has been associated with attention (Unsworth and Robison, 2016). Wagner *et al.* (2019) compared changes in baseline pupil size across trials for normal-hearing and hearing-impaired listeners. They found a slower decrease in baseline pupil size for the hearing-impaired listeners which was interpreted as more sustained attention in the face of increased task demands.

C. Adaptation

To counter the detrimental effects of channel and source degradations on intelligibility and effort, listeners implement cognitive strategies to deal with such perturbations. Even short-term exposure to degraded speech can thus improve speech recognition, a phenomenon referred to as perceptual learning or adaptation (for a review, see Samuel and Kraljic, 2009). For instance, participants can adapt to sentences time-compressed at 38% of their original duration within a block of 15 sentences (Dupoux and Green, 1997). Similar adaptation effects have been found for noise-vocoded speech (Davis *et al.*, 2005). Both degradations systematically change the underlying signal, allowing for perceptual recalibration (Peelle and Wingfield, 2005). For noise-vocoded speech, it has been shown that adaptation requires attention (Huyck and Johnsrude, 2012). Listeners have been shown to adapt to source degradations as well, such as fast speech (Adank and Janse, 2009) or accented speech (Banks *et al.*, 2015). In addition, it is conceivable that adaptation is modulated by talkers' acoustic-phonetic profiles. It has been suggested that "predictable and consistent deviations (e.g., accented or disordered speech) can cause more fundamental recalibration over time" (Mattys *et al.*, 2012). Eisner and McQueen (2005) showed that while such recalibration is influenced by lexical context, it is also talker-specific.

D. Aims

As channel and source degradations interact, intelligibility decreases. Listening effort has been associated with both degradations, but their combined effect on effort has not been investigated. Specifically, it is unclear how talker-specific acoustic features interact with spectral and temporal degradations, and the extent to which this interaction impacts both intelligibility and effort. Despite listeners' ability to adapt to channel and source degradations, it is not known if and how adaptation to spectral and temporal degradations is modulated by talker-specific acoustic features. We conducted a listening experiment combined with pupillometry, measuring keyword recognition performance, adaptation and task-evoked pupil response for noise-vocoded, time-compressed, and masked speech, as well as speech in quiet. Sentence material from 16 talkers was used. We considered a range of acoustic features that might be linked to intelligibility and effort, as well as adaptation. To expand on previous findings (Bent *et al.*, 2009; Green *et al.*, 2007), we hypothesized that talkers who are more intelligible under degradations that affect the spectral detail of speech (noise-vocoded and masked speech) would also be more intelligible under temporal degradations (time-compressed speech).

This hypothesis is based on the assumption that spectral features such as precise articulation can be linked to temporal features, for instance, slower SRs (Hazan and Markham, 2004). If talker differences were preserved across spectral and temporal degradations, we expected this result to be linked to specific intelligibility-promoting features or combinations of features in acoustic-phonetic talker profiles. Similarly, with respect to listening effort, we hypothesized smaller pupil dilation to be associated with acoustic-phonetic features driving intelligibility benefits. If adaptation was modulated by the talker, we believed acoustic-phonetic features rendering speech more predictable (e.g., less-deviant f_0 fluctuations) to contribute more strongly to adaptation. Furthermore, we expected a slower decline in baseline pupil size to be linked to sustained attention under increased adaptation.

II. METHODS

A. Speech materials

1. Recordings

Sixteen speakers were recorded: eight older adults [four females; $M_{age} = 71$ (5.1) years; $range_{age}$: 61–77 years] and eight younger adults [four females; $M_{age} = 26.8$ (3.2) years; $range_{age}$: 22–33 years]. We sampled speakers across different age groups and both sexes to include a wide range of speaker-related anatomical-physiological variation. All participants were native speakers of Standard Southern British English. Each speaker read aloud 720 Harvard sentences (Institute of Electrical and Electronics Engineers, 1969), which are commonly used in speech perception experiments given their low semantic predictability and normed phonetic structure and length (e.g., Banks *et al.*, 2015). During each recording session, breaks were permitted if needed. Recordings were made in an anechoic chamber using a Bruel & Kjaer 2231 Sound Level Meter fitted with a type 4165 condenser microphone. The signal was digitized with a Focusrite 2i2 USB audio interface at a sampling rate of 44 100 Hz and a bit-depth of 16 bits. Sentences were displayed on a screen facing the participant and the experimenter controlled the timing of sentence presentation. ProRec (Huckvale, 2014) was used for sentence recording and segmentation, including removal of silent parts. Recordings were manually checked and any remaining silent parts at the beginning and end of each sentence were cut at zero crossings using Praat (Boersma and Weenink, 2018). Of all 720 sentences, those with unexpected noise or mispronunciations for any of the speakers were removed from the final set (237 sentences in total). Of the remaining sentences, 192 were randomly selected for the perception experiment. The same subset of sentences was selected from each speaker and only this subset was analysed acoustically. Sentences were converted to mono, down-sampled to 22 050 Hz and high-pass filtered at 50 Hz, removing gross fluctuations. Sentences were root-mean-square (rms) normalized. We automatically annotated and aligned sentences using the Montreal forced aligner (McAuliffe *et al.*, 2017). The aligner is trained on raw speech and word-level transcription of each sentence and outputs aligned text grids with word- and phone-level annotation. Text grids for each speaker were manually checked to ensure that no processing errors occurred. We specifically checked for correct annotation and alignment of vowels since text grids were used exclusively for vowel space analysis.

2. Acoustic analyses

Acoustic analyses were conducted using custom-made scripts in Praat. All acoustic analyses were based on normalized signals. Single measures were obtained for each speaker by averaging across all 192 sentences. For adaptation analyses, standard deviations across all sentences were obtained as well, reflecting acoustic predictability.

Energy in mid-range frequencies (ME13): Mean energy in the 1–3 kHz range has reliably shown a relationship with intelligibility (Green *et al.*, 2007; Hazan and Markham, 2004). First, the long-term average spectrum (LTAS) was obtained for each sentence. We then measured the average intensity of the spectrum in the range from 1 to 3 kHz.

Fundamental frequency (f_0M, f_0SD): Bradlow *et al.* (1996) found mean fundamental frequency (f_0) to be significantly correlated with intelligibility which was driven by increased mean f_0 and higher intelligibility for female talkers. They also found a tendency for a correlation between wider f_0 range and higher intelligibility. An increased dynamic f_0 range corresponds to a clear and slow speaking style (Picheny *et al.*, 1986). We included f_0 median and f_0 standard deviation as acoustic features. Periodicity detection was performed by applying the auto-correlation method implemented in Praat using a 10 ms frame duration. Upper and lower boundaries were set to $q65 * 1.92$ and $q15 * 0.83$ with q representing the respective quantiles. The formulas were optimized to reduce artefacts such as octave jumps (De Looze and Hirst, 2008). Similar to Hazan and Markham (2004), we measured f_0 median (f_0M) and standard deviation in semitones (f_0SD) for each sentence and obtained means across all sentences.

Speaking rate: Even though SR is not consistently linked to intelligibility and effort, we hypothesized that slow speech would be more beneficial than fast speech when speech is time-compressed. We estimated SR by dividing the canonical number of syllables in a sentence by the duration of the sentence. The number of syllables was obtained for each sentence transcription using the package quanteda in R (Benoit, 2018). Syllables per second were then defined as a measure of SR.

Vowel space dispersion (VSD): More peripheral vowel locations in the $F_1 - F_2$ space relate to higher intelligibility (Bradlow *et al.*, 1996). Estimates of a talker's vowel space can be obtained by measuring the range of F_1 and F_2 across vowel realizations, as well as by measuring VSD (Bradlow *et al.*, 1996). Similar to Bradlow *et al.*, we measured VSD as the Euclidean distance of three peripheral vowels (/*aoi*/)

from the geometric center of the vowel space. Only vowels from content words were analyzed ($N_a = 41, N_o = 62, N_i = 89$), following Bradlow *et al.* (1996). Formants were measured at the vowel center, applying short-term spectral analysis with a 25 ms window size. The formant maximum was adjusted for speaker gender (male = 5000 Hz; female = 5500 Hz). Formants were converted to the mel scale (Fant, 1973), following Bradlow *et al.* (1996). The Euclidean distance of each vowel realization (F_1, F_2) from the vowel space center $(\overline{F_1}, \overline{F_2})$ was calculated using the formula in Eq. (1),

$$d(F_1, F_2) = \sqrt{(F_1 - \overline{F_1})^2 + (F_2 - \overline{F_2})^2}. \quad (1)$$

B. Listening experiment

1. Participants

Sixty-four normal-hearing native speakers of British English were recruited for the experiment [40 females; $M_{age} = 22.3$ (4.3) years; $range_{age}$: 18–37 years]. They were either reimbursed for their participation following the guidelines of the Division of Psychology and Language Sciences at the University College London or given course credit. Hearing ability was established by a standardized audiometric test at the beginning of each testing session. Participants had hearing thresholds equal to or better than 25 dB hearing level (HL) at all tested octave frequencies between 0.25 and 8 kHz. This threshold is in line with similar studies (e.g., Wendt *et al.*, 2018). Three participants had hearing thresholds of 30 dB HL at one of the tested frequencies (0.5, 2, and 8 kHz, respectively). We included these participants as well given that their performance was above the mean in the noise condition.

2. Materials

From the 192 sentences, we created four lists of 48 items each. Even though pupil dilation effects during listening can be detected with as few as 20–25 items (Winn *et al.*, 2018), more trials are necessary to sufficiently estimate adaptation to noise-vocoded speech (e.g., Erb *et al.*, 2012). The lists were optimized so that the mean duration was roughly matched across lists [$M_{duration} = 2.246$ s (0.016)]. It was ensured that the same keyword did not appear more than twice within the same list.

3. Listening conditions

We presented sentences in quiet and in three degradations: time-compression, noise-vocoding, and masking (noise). Noise-vocoding and masking have been used in a previous study that found talker differences to be preserved across these conditions (Bent *et al.*, 2009). We expected a similar effect for time-compression. In addition, in accordance with Peelle and Wingfield (2005), time-compression and noise-vocoding were expected to show robust adaptation effects, in contrast to masking. While masking noise is random and only "obscures the speech stimulus" (Peelle and Wingfield, 2005), modifying speech by time-compression and noise-vocoding introduces systematic changes that can be adapted to.

Parameters were chosen based on pilot results indicating that intelligibility was not too low overall, avoiding disengagement effects on the pupil measures. At the same time, it was ensured to leave enough room for possible adaptation effects. Sentences were time-compressed to 37% of their original duration by applying the pitch-synchronous overlap-add implementation in Praat. Pilot data showed that this rate was sufficient to elicit adaptation effects without a significant drop in intelligibility that can be observed when increasing the compression rate further (e.g., Versfeld and Dreschler, 2002). For noise-vocoding, the original signal was divided into six frequency bands spaced according to the cochlear frequency-position function (Greenwood, 1990). Amplitude envelopes were extracted from each band by applying a 4th-order Butterworth low-pass filter with a cutoff frequency at 256 Hz and half-wave rectification. The envelopes were then used to modulate white noise. For masking, speech-shaped noise was created by obtaining the LTAS of a separate set of sentences from a non-experimental talker. Noise was then generated with the same LTAS and added to the sentence at a signal-to-noise ratio of –1 dB.

4. Design and procedure

Each listener was presented all conditions in four blocks of 48 sentences. Blocks were counterbalanced across listeners using a Latin square design. All 48 sentences in one block were spoken by the same talker, as it has been shown that changing talkers interferes with adaptation (e.g., Dupoux and Green, 1997). Talkers were counterbalanced across listeners and blocks so that each talker was heard by 16 listeners in total and by four listeners per condition. Lists and sentences within each list were randomized. The large number of sentences required and the talker change constraint imposed by the adaptation measure limited the number of talkers that could be presented within one testing session. In addition, the acquisition of pupillometry data requires monitoring by the experimenter, making larger-scale studies such as that conducted by Bent *et al.* (2009) unfeasible.

Participants wore headphones (Sennheiser HD 25 SP II) throughout the experiment with output levels at 70 dB sound pressure level (SPL). They were asked to put their head comfortably on a table-mounted chin rest, to minimize head movements. Glasses had to be removed for the duration of the experiment. Pupil recordings were obtained using an EyeLink 1000 table-mounted eye tracker at a distance of 55 cm from the participant's head. A sampling rate of 500 Hz was used. The light level was kept constant at 130 lux, but for participants with very large or very small resting state pupil sizes, the light level was adjusted as required (Wendt *et al.*, 2018). The experiment started with eight practice trials in which sentences in quiet were presented to the participants. The materials were taken from a non-experimental talker whose recordings were not included in the corpus. Each trial followed the same procedure (see Fig. 1): after an inter-stimulus interval of 500 ms, the baseline pupil size was recorded for 2000 ms, which was followed by the sentence onset. After the offset of the sentence, the pupil size was tracked for another 2000 ms in quiet since the dilation usually peaks around 0.7 to 1.2 s after stimulus offset (Winn *et al.*, 2018). The fixation cross changed color to signal the end of the retention period and the start of the response. Participants repeated back words to the experimenter who logged correctly identified words on a separate control screen. A keyword was considered correctly identified despite incorrect suffixes such as plural (-s) or tense (-ed) endings (Banks *et al.*, 2015). Participants were asked to blink as little as possible during the trial up to the point that a response had to be given. The experiment was implemented in MATLAB (R2016a).

5. Dependent variables

Intelligibility and adaptation: To obtain intelligibility scores, the proportion of keywords correctly identified out of five was calculated and averaged across trials. This means that for each listener, recognition averages were based on 48 sentences. To obtain adaptation rates, we compared linear, power-law (Erb *et al.*, 2012) and quadratic function fits to proportion correct of all 48 sentences for each listener and condition. Goodness of fit of each model was determined by the Bayesian information criterion (BIC, Schwarz, 1978). On average, the linear fit resulted in the lowest BIC in all experimental conditions. The slope of the linear fit was therefore used as a measure of adaptation rate.

Pupillometry: We followed the guidelines and functions provided by Geller *et al.* in the GazeR package[1] to preprocess pupil data. We only included data collected between the onset of the inter-stimulus interval and the verbal response prompt due to the possibility that articulatory movements interfered with the measure. Blinks were automatically detected by the EyeLink and marked as missing values. Gaps of missing data were extended to 100 ms before and 100 ms after the gap due to effects of eyelid closure on the pupil size. Trials that contained more than 20% missing data within the specified interval were excluded. This procedure resulted in the inclusion of 47.6 out of 48 trials on average. In order to obtain representative pupil trace averages, blocks with fewer than 24 remaining trials (50%) were removed. Two blocks of one participant were therefore removed. We interpolated missing values linearly. Data was then smoothed using a 5-point moving average.

We determined thresholds for unrealistic pupil sizes by visually inspecting distributions of pupil sizes across all participants. Samples outside these thresholds were removed. Additionally, rapid pupil size disturbances were detected and removed using the median absolute deviation. Divisive baseline correction was applied using the mean of the baseline recorded 1000 ms before the onset of the sentence. Pupil traces were then time-aligned with sentence offset and down-sampled to 20 Hz. They were averaged for each listener and condition. We visually inspected all average pupil traces for anomalies in overall shape and magnitude. One participant was excluded since average pupil traces in each condition showed decreasing pupil size. It is possible that the pupil size for this participant was only affected by the motor response while slowly returning to baseline during the trial.

Several dependent measures were obtained, following Winn *et al.* (2018). The peak dilation is the maximum value of each average trace within a specified time window. Since sentence duration varied largely between time-compressed speech and all other conditions, we first inspected the latency of dilation maximums for each participant and condition. The search space ranged from –1418 to +2000 ms (quiet, masking, and noise-vocoding) and –525 to +2000 ms (time-compression) with respect to sentence offset. The lower boundary was the respective duration of the shortest sentence and the upper boundary was the onset of the response. The interquartile range for peak latencies in all conditions was located within the retention period from 0 to 2000 ms. This is a typical observation for pupillometry studies (Winn *et al.*, 2018). We extracted peak dilation and latency, as well as mean dilation with respect to the specified time window. All measures were expected to reflect listening effort.

Following Wagner *et al.* (2019), we analyzed changes in baseline pupil size across trials, hypothesizing that sustained attention as indexed by a slower decline in baseline pupil size would relate to adaptation. Changes in baseline pupil size were calculated as percentage change from the mean baseline of the first two trials. Similar to Wagner *et al.* (2019), we additionally analyzed changes in mean pupil dilation across trials, as reflecting fatigue. Mean instead of peak pupil dilation was chosen as peaks are usually obtained from average pupil traces.

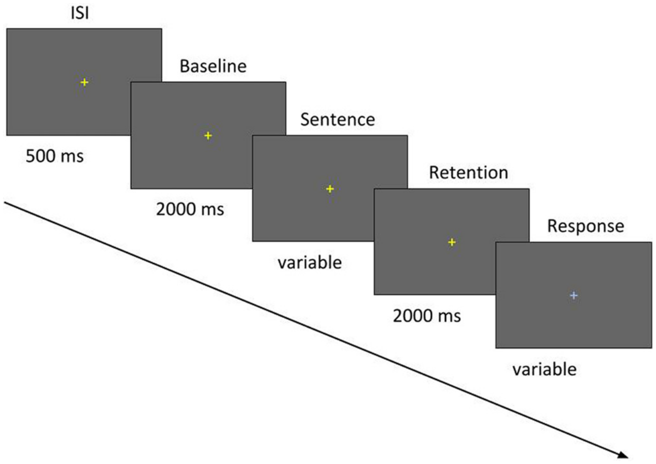

FIG. 1. (Color online) Trial events with duration. Rectangles represent displays with central fixation cross.

6. Statistical analysis and data aggregation

For differences between listening conditions, we used linear mixed effects models (Bates *et al.*, 2015) with Satterthwaite's degrees of freedom method. In all models, we allowed random intercepts for listeners. For pairwise comparisons, Bonferroni adjustments were made. Similar models were used to analyze changes in baseline pupil size and mean pupil dilation across trials. We included linear and quadratic time terms as predictors following the growth curve analysis approach (Mirman, 2014). This type of analysis allows for an independent analysis of overall pupil size (intercept), slope (linear term), and rise and fall rate around the inflection point (quadratic term) (Wendt *et al.*, 2018).

The relationship between dependent measures and acoustic-phonetic characteristics was investigated by means of Pearson's product moment correlations across the 16 talkers. Dependent measures were aggregated for each talker and condition so that each talker average was based on four listeners and 192 sentences in total. Dependent measures from one listener in the noise-vocoding condition were not included in the talker averages since almost no keywords were recognized correctly [$M = 2.5\%$]. Due to the exclusion of average pupil traces for some listeners during preprocessing, talker averages for one talker in the time-compression and quiet condition, two talkers in the masking condition and three talkers in the noise-vocoding condition were based on three listeners only.

III. RESULTS

A. Effect of channel degradation

1. Intelligibility and adaptation

First, we investigated the effect of degradation type on keyword recognition. There was a main effect of condition [$F(3, 188.3) = 155.33, p < 0.001$]. Pairwise comparisons showed that recognition was poorer for all types of degradations compared to quiet ($p < 0.001$). Recognition for noise-vocoding was poorer than for masking and time-compression ($p < 0.001$), while recognition for masking was poorer than for time-compression ($p < 0.001$) (Fig. 2). These differences should not be understood as an effect of degradation type, but rather as reflecting the degree of degradation chosen *a priori* for each condition. In part B of this results section, we will show how talker differences can explain the variances observed in each condition.

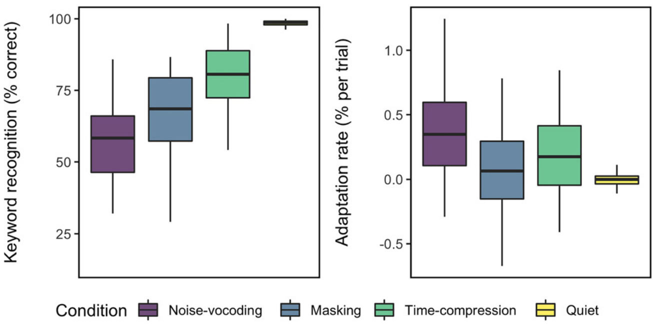

FIG. 2. (Color online) Distributions of the average and rate of keywords recognized correctly in all conditions. Boxes represent values from the first to the third quartile with the median indicated by a black line. Whiskers extend up to 1.5 times the interquartile range.

We compared adaptation rates between conditions to confirm that listeners adapted to noise-vocoded and time-compressed speech, but not to masked speech. For speech in quiet, we expected rates to be close to 0 due to ceiling intelligibility. We found a main effect of condition [$F(3, 251) = 21.05, p < 0.001$]. Pairwise comparisons showed that adaptation rates were higher for noise-vocoding ($p < 0.001$) and time-compression ($p < 0.05$) compared to quiet for which adaptation rates were close to 0 [$M = -0.003(0.07)$], indicating lack of adaptation due to a ceiling effect. This result indicated that listeners adapted to noise-vocoded and time-compressed speech. Adaptation rates were also larger for noise-vocoding compared to masking ($p < 0.001$) and time-compression ($p < 0.05$). However, there was no difference between adaptation rates for masking and time-compression. This result might be related to overall higher intelligibility for time-compressed speech allowing for less improvement overall.

2. Pupillometry

We investigated whether pupil dilation measures followed the same trend as recognition scores, reflecting increased effort for degraded speech (Fig. 3). There was a main effect of condition for peak dilation [$F(3, 183.17) = 33.17, p < 0.001$]. Pairwise comparisons showed that peak dilation was larger for all three degradations compared to quiet ($p < 0.001$), but there was no difference between degradations. There was a main effect of condition for mean dilation [$F(3, 183.09) = 32.73, p < 0.001$] with pairwise comparisons indicating that mean dilation was larger for all three degradations compared to quiet ($p < 0.001$). Mean dilation was also larger for noise-vocoding compared to time-compression ($p < 0.001$), but not compared to

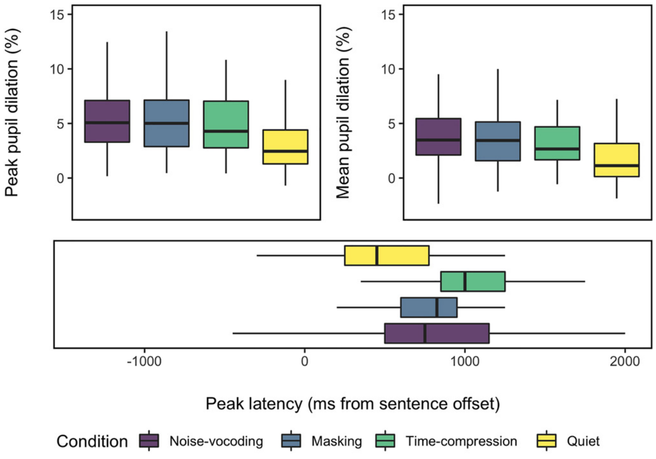

FIG. 3. (Color online) Distributions of peak pupil dilation, mean pupil dilation, and peak latency in all conditions. Boxes represent values from the first to the third quartile with the median indicated by a black line. Whiskers extend up to 1.5 times the interquartile range.

masking. Peak latency also showed a significant main effect of condition [$F(3, 183.9) = 15.41, p < 0.001$]. Pairwise comparisons indicated larger peak latency for masking ($p < 0.05$), noise-vocoding ($p < 0.01$) and time-compression ($p < 0.001$) compared to quiet. These results indicate that the task-evoked pupil response peaked later when speech was degraded, reflecting increased effort. Pairwise comparisons also indicated larger peak latency for time-compression compared to masking ($p < 0.01$) and noise-vocoding ($p < 0.05$). It has to be noted that pupil traces were aligned to sentence offset so that these results indicate a delayed peak response for time-compressed speech measured from the end of the sentence. Overall, the results presented above indicate that listening effort was inversely related to recognition performance. At the same time, pupil dilation measures were less sensitive to differences between conditions, possibly driven by larger variability between listeners.

We conducted a growth curve analysis to investigate changes in baseline pupil size and mean dilation across trials. Baseline pupil size and mean pupil dilation generally declined over the course of the experiment (see Fig. 4). As the curves showed a flatter response towards the end of the block, we included both linear and quadratic terms in the models. However, the best fitting model, obtained using backward elimination, did not include effects on the quadratic term. The overall magnitude of the baseline pupil size was larger for masking ($\beta = 1.78, t = 7.83, p < 0.001$; $\beta = 1.63, t = 7.12, p < 0.001$) and time-compression ($\beta = 1.51, t = 6.63, p < 0.001$, $\beta = 1.36, t = 5.94, p < 0.001$) compared to quiet and noise-vocoding, respectively. There was no significant difference between noise-vocoding and quiet, as well as masking and time-compression. There was a significantly slower linear decline for masking ($\beta = 50.73, t = 2.04, p < 0.05$), time-compression ($\beta = 75.17, t = 3.03, p < 0.01$), and noise-vocoding ($\beta = 55.00, t = 2.20, p < 0.05$) compared to quiet. There was no significant difference between degradations, despite an initially faster decline for noise-vocoded speech. Increasing baseline pupil size towards the end of the block might explain the similar slopes. Overall, baseline pupil size was smallest for noise-vocoded speech and speech in quiet. Mean pupil dilation generally declined across trials, but there was no significant difference between conditions on the linear term. Instead, we observed differences in the overall mean pupil dilation in accordance with the inverse relationship of intelligibility and pupil dilation reported above. Degraded speech was associated with overall larger mean dilation compared to quiet (masking:$\beta = 2.14, t = 13.75, p < 0.001$; time-compression: $\beta = 1.67, t = 10.70, p < 0.001$; noise-vocoding: $\beta = 2.82$, $t = 17.94, p < 0.001$). Overall mean dilation was also larger for noise-vocoding than for masking ($\beta = 0.68$, $t = 4.31, p < 0.001$) and time-compression ($\beta = 1.16$, $t = 7.35, p < 0.001$). Furthermore, mean dilation was larger for masking than for time-compression ($\beta = 0.48$, $t = 3.07$, $p < 0.01$).

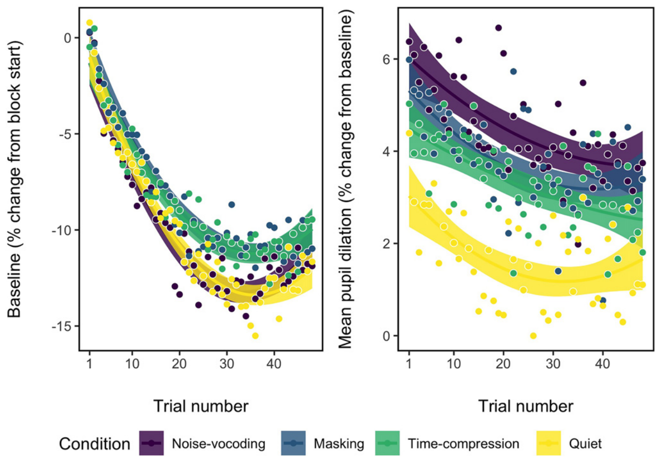

FIG. 4. (Color online) Change in baseline pupil size and mean pupil dilation across trials. Ribbons indicate the 95% confidence interval.

3. Summary

Intelligibility differed between listening conditions and was optimal for speech in quiet. We observed adaptation to noise-vocoded and time-compressed speech only. Intelligibility was inversely related to pupil dilation, indicating more effortful processing for less intelligible speech. Degraded speech generally elicited a peak pupil dilation that occurred later compared to speech in quiet. Overall baseline pupil size was smaller for noise-vocoded speech and speech in quiet, possibly reflecting inattentiveness. We will link the dependent measures presented in this section to acoustic-phonetic talker differences.

B. Effect of talker

1. Correlations between acoustic-phonetic measures

We investigated relationships amongst acoustic features by conducting correlation analyses. There was a negative correlation between SR and f_0 standard deviation (f_0SD) ($r = -0.50, p < 0.05$), indicating that talkers with slower SR showed larger f_0 fluctuations. There was a negative correlation between SR and VSD that did not reach significance ($r = -0.40$). Talker group averages are included as supplemental materials.[2] To avoid multicollinearity in later analyses, we determined the variance inflation factor (VIF). All VIF scores were below 5—a VIF score larger than 5 or 10 indicates multicollinearity (Menard, 1995).

Individual recognition scores for listeners were aggregated by talker and condition. The following analyses aimed at (1) investigating whether talker intelligibility would be similar across conditions and (2) identifying the talker acoustics contributing to intelligibility differences in each condition and across conditions. Talker-aggregated recognition scores for quiet were at ceiling (96%–99%) and were therefore excluded from intelligibility analyses.

2. Intelligibility across degradations

To investigate talker intelligibility across listening conditions, we conducted correlation analyses for each pair of conditions. There was a significant correlation for per-talker recognition averages in noise-vocoding and masking ($r = 0.66,\ \ p < 0.01$) and noise-vocoding and time-compression ($r = 0.54, p < 0.05$) (see Fig. 5). These results indicate that talkers who were intelligible under noise-vocoding were also intelligible under masking and time-compression. For the masking/time-compression pair, we found a correlation at $r = 0.43$ which did not reach significance. We conducted further analyses with acoustic-phonetic features to identify possible candidate features or feature combinations that explain why some talkers were more intelligible across degradations.

3. Correlations of acoustic-phonetic measures with intelligibility and adaptation

For each acoustic feature, we conducted correlation analyses with intelligibility data (Table I). For masking, there was a significant correlation of recognition scores and ME13 ($r = 0.61, p < 0.05$), indicating that talkers with increased spectral energy measured for frequencies between 1 and 3 kHz were more intelligible under masking. For noise-vocoding, there was a significant correlation of recognition scores and VSD ($r = 0.56, p < 0.05$). This result indicated that talkers with greater articulatory distances between vowel center and peripheral vowel realizations were more intelligible when speech was noise-vocoded. For time-compression, there was a significant negative correlation with SR ($r = -0.55, p < 0.05$), indicating that talkers with slower SRs were more intelligible when speech was time-compressed.

We assessed whether changes in intelligibility across trials (adaptation) depended on a talker's acoustic characteristics by conducting correlation analyses for each acoustic feature. We included means as well as standard deviations of each feature across all sentences. Standard deviations were assumed to reflect acoustic predictability. We did not find significant correlations in any of the conditions, indicating that there was no systematic relationship between talker acoustics and adaptation over time.

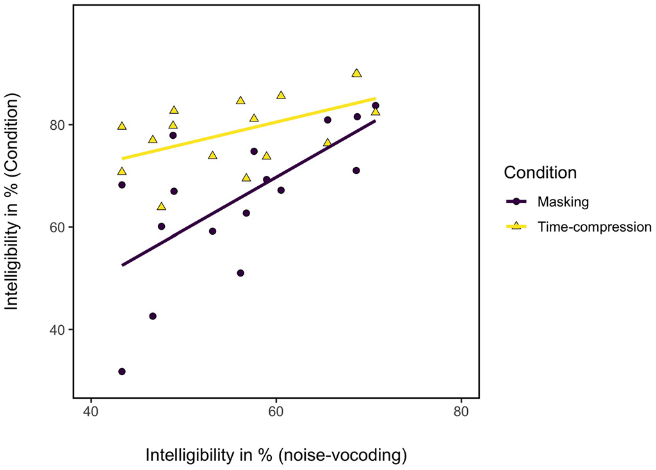

FIG. 5. (Color online) Intelligibility, i.e., average proportion of keywords recognized correctly aggregated by talker and experimental condition. Each symbol therefore represents one talker average. Noise-vocoding is plotted against masking and time-compression.

TABLE I. Correlation analyses between acoustic features and per-talker recognition scores in each condition. ME13, mean energy; F_0M, f_0 median; F_0SD, f_0 standard deviation; SR, VSD. * = significant values ($p < 0.05$).

	Masking	Time-compression	Noise-vocoding
ME13	0.61*	0.05	0.10
F_0M	–0.03	0.10	0.06
F_0SD	0.25	0.09	0.27
SR	–0.12	–0.55*	–0.38
VSD	0.10	0.36	0.56*

Even though only single features correlated significantly with intelligibility in each degradation, we expected to find combinations of features to explain why some talkers were more intelligible across degradations. We investigated which combination of features would best predict intelligibility by constructing a linear model for each condition. We applied step-wise regression with forward and backward selection as implemented in R's MASS package. The best models were determined based on the Akaike information criterion (AIC). R^2 was adjusted for the number of predictors. In addition, we determined the relative importance (relaimp) of each feature in each model by applying the lmg method from R's relaimp package (Grömping, 2006). The final model for masking contained mean energy ($relaimp = 0.85$) and VSD ($relaimp = 0.15$) ($R^2 = 0.46$). The final model for noise-vocoding contained VSD ($relaimp = 0.84$) and mean energy ($relaimp = 0.16$) ($R^2 = 0.37$). The final model for time-compression contained SR ($R^2 = 0.25$). The results show that VSD and mean energy emerged as common features between models for noise-vocoded and masked speech.

4. Correlations of acoustic-phonetic measures with listening effort

We conducted correlation analyses for each acoustic feature with peak dilation, mean dilation, and peak latency. For quiet, mean pupil dilation ($r = -0.55, p < 0.05$) and peak latency ($r = -0.62, p < 0.01$) were negatively correlated with VSD, indicating reduced listening effort for well-articulated speech. Furthermore, SR was correlated with mean ($r = 0.50, p < 0.05$) and peak pupil dilation for time-compression ($r = 0.62, p < 0.05$), indicating reduced effort for slower talkers under time-compression. Note that SR also predicted intelligibility of time-compressed speech. We did not find a relationship between pupil dilation measures and talker acoustics for noise-vocoded and masked speech. Overall intelligibility levels were lower for noise-vocoded and masked speech, but higher for time-compressed speech and speech in quiet. Lower overall intelligibility levels might therefore have had an influence on the detection of talker differences with respect to the non-linearity of the pupil response (Wendt *et al.*, 2018). It has to be noted that individual listener differences were relatively large (see Table II), which might have affected the reported correlation analyses.

5. Summary

We linked acoustic-phonetic talker differences to intelligibility, adaptation, and effort. The results demonstrated that individual acoustic-phonetic features were beneficial for intelligibility when speech was noise-vocoded (VSD), time-compressed (SR), or masked (mean energy). We found that talkers who were more intelligible under noise-vocoding were also more intelligible under masking and time-compression. Increased VSD and mean energy could be linked to intelligibility under masking and noise-vocoding. Acoustic-phonetic talker differences did not have an effect on adaptation. However, we linked some talker differences to pupil dilation measures associated with listening effort. Talkers with slower SR or larger VSD were respectively associated with reduced listening effort under time-compression and in quiet.

IV. DISCUSSION

The aim of our study was to establish the combined effect of channel and source degradations on intelligibility and listening effort. In particular, our focus was on the interaction of talker-specific acoustic features with spectral and temporal degradations. As listeners can adapt to both channel and source degradations, we hypothesized that adaptation to spectral and temporal degradations is modulated by talker-specific acoustic features.

A. Intelligibility and listening effort

Average intelligibility differed across listening conditions, reflecting the parameters chosen for each degradation type. Intelligibility was optimal when speech was presented in quiet. Adaptation rates, i.e., improvements of intelligibility over time, were higher for time-compressed and noise-vocoded speech, compared to speech in quiet. Adaptation rates for masked speech were not different from speech in quiet, and it is likely that adaptation was degradation-specific and not due to task familiarity (Peelle and Wingfield, 2005). We found higher adaptation rates for noise-vocoded than for time-compressed speech. Noise-vocoded speech was also less intelligible, thus providing more “room for improvement.”

TABLE II. To analyze talker effects, dependent measures from listeners were aggregated by talker and condition. Condition-wise grand means and means of standard deviations (in brackets) are displayed.

	Recognition (%)	Adaptation (% / trial)	Peak dilation (%)
Noise-v.	55.9 (11.9)	0.4 (0.3)	6.4 (3.7)
Masking	65.5 (7.3)	0.1 (0.3)	5.5 (2.7)
Time-c.	78.6 (10.7)	0.2 (0.3)	5.6 (3.3)
Quiet	98.3 (1.3)	−0.0 (0.1)	3.1 (2.2)

Peak and mean pupil dilation were larger for degraded speech than for speech in quiet, which was attributable to lower intelligibility and higher effort (Wendt *et al.*, 2018). A growth curve analysis of mean dilation change across trials also indicated intelligibility-related differences between noise-vocoded, masked, and time-compressed speech. These statistical differences were not detectable when averaging across trials, which was possibly influenced by the variability of the pupil dilation measure. Latency of the peak pupil dilation was larger for degraded speech compared to speech in quiet, indicating higher effort for degraded speech. Latency was also larger for time-compressed speech compared to all other conditions. Since intelligibility was overall high for time-compressed speech, this effect might not solely be due to increased demands. Even though dilation peaks usually appear with a delay of 0.7–1.2 s after sentence offset (Winn *et al.*, 2018), it seems that a shorter sentence duration prolongs the peak. Despite lower demands for time-compressed than for noise-vocoded speech, complete sentence processing might occur later. This finding should be considered in listening effort frameworks that tend to emphasize the overall magnitude of the pupil dilation.

We observed that mean pupil dilation and baseline pupil size generally declined over the course of a block (48 sentences). This decline might reflect task familiarization, but also fatigue (Wagner *et al.*, 2019). Given the role of attention in adapting to noise-vocoded speech (Huyck and Johnsrude, 2012), we expected to see a more sustained baseline pupil size across trials. Contrarily, there was a more rapid decline and overall smaller baseline pupil size for noise-vocoded speech and speech in quiet. This finding might be linked to overall intelligibility differences observed between conditions. As noise-vocoded speech and speech in quiet were respectively the most and least challenging conditions, a faster decline in baseline pupil size might reflect disengagement or inattentiveness (Unsworth and Robison, 2016). However, a question arises of how inattentiveness while processing noise-vocoded speech can explain the strong adaptation effects and the consistently larger mean dilation. One explanation might be the speed of adaptation (Erb *et al.*, 2012): fast adaptation at the beginning of a block led to an earlier onset of fatigue. As recognition performance increased towards the end of a block, fatigue decreased. This hypothesis is corroborated by the finding that the overall slope of decline in baseline pupil size was not different for noise-vocoded, masked and time-compressed speech.

B. Talker-dependent intelligibility

Talkers with slower SRs showed larger fluctuations in fundamental frequency. A wider range in fundamental frequency is a characteristic of clear speaking styles (Picheny *et al.*, 1986), especially for read speech materials (Hazan and Baker, 2011). As talkers showing these characteristics were primarily older adult talkers, the f_0 range might have been exaggerated as a number of older adults used a more "theatrical" reading style. Talkers who were more intelligible under noise-vocoding were also more intelligible under masking and time-compression. This finding replicates and extends results from Bent *et al.* (2009), who found that talkers intelligible under noise-vocoding were also more intelligible under babble noise. The authors argued that both types of degradations affected the spectral characteristics of talkers so that similar acoustic-phonetic features were responsible for the effect. However, this explanation does not extend to time-compression because the signal processing technique used in this experiment (pitch synchronous overlap and add, Moulines and Charpentier, 1991) does ideally not change the spectral properties of speech. Furthermore, the acoustic-phonetic predictors with highest relative importance were in our case distinct for the three degradations. For masked speech, mean energy in the 1–3 kHz region contributed most to intelligibility, similar to previous studies (e.g., Green *et al.*, 2007; Hazan and Markham, 2004). This result is expected, given the predictions of the speech intelligibility index (American National Standards Institute, 1997). For masked speech, VSD was also marginally relevant. For noise-vocoded speech, VSD contributed most to intelligibility, with a marginal relevance of mean energy. For time-compressed speech, only SR contributed to intelligibility.

The lack of a single "catch-all" feature suggests that a combination of features is more likely to explain why some talkers are more intelligible under different degradations (Hazan and Markham, 2004). We observed that both VSD and mean energy contributed to intelligibility under noise-vocoding and masking. This finding suggests that talkers ranking higher specifically on these features were more likely to be intelligible in both conditions. We did not find such common features between noise-vocoded and time-compressed speech. At the same time, VSD was moderately correlated with SR. Even though this correlation was not significant, it suggests that at least some talkers with slower SRs also exhibited greater VSD and were therefore more intelligible under noise-vocoding and time-compression.

C. Talker-dependent adaptation and listening effort

We found that listeners adapted to noise-vocoded and time-compressed speech, but we were not able to link adaptation slopes to acoustic-phonetic measures. Furthermore, we did not find a systematic relationship between adaptation slopes and talker intelligibility. Contrary to our results, a previous study investigating adaptation to accented speech found faster adaptation to talkers with higher baseline intelligibility (Bradlow and Bent, 2008). Talkers in our study were from the same accent group and talkers and listeners shared the same native language background. Therefore, it seems likely that listeners were familiar with the accent- or language-specific acoustic-phonetic characteristics so that talker-specific adaptation was not required. We therefore suggest that adaptation mainly functioned to overcome the channel degradation and not the source degradation. Systematic changes in the signal, introduced by time-compression and noise-vocoding, allowed for perceptual learning to occur.

We related acoustic-phonetic features to pupil dilation measures, investigating the effect of talker differences on listening effort. For speech in quiet, talkers with greater VSD were associated with a faster and more attenuated pupil response, indicating reduced listening effort. VSD also emerged as a relevant feature for noise-vocoded and masked speech. Even though intelligibility for speech in quiet was optimal, it appears that the pupil dilation indicated ease of processing for generally more intelligible talkers. Previously, also subjective ratings of listening effort have been shown to be more sensitive at higher intelligibility levels (Morimoto *et al.*, 2004; Rennies *et al.*, 2019). For time-compressed speech, we found that slower talkers were associated with more attenuated peak pupil dilation and higher intelligibility. SR is not by default linked to intelligibility and a recent study showed that differences in SR were not reflected in the pupil dilation response to speech in quiet (Koch and Janse, 2016). We suggest that time-compression amplified the effect of SR on both measures.

Our results indicated an effect of degradation level on the sensitivity of the pupil dilation measure. Intelligibility was on average low for noise-vocoded speech (56%) and masked speech (66%), and high for time-compressed speech (79%) and speech in quiet (98%). For stationary maskers, Wendt *et al.* (2018) found that peak dilation remained large when increasing the signal-to-noise ratio (SNR) from –8 to –4 dB, while improvement in sentence recognition was steepest (∼30%–80%). Increasing the SNR from 0 to 4 dB resulted in a significant decrease in peak dilation, but virtually no difference in intelligibility, due to a ceiling effect. The non-linearity of the pupil dilation response can explain why talker differences were not apparent for noise-vocoded and masked speech in our experiment: even though talker differences contributed largely to intelligibility in these conditions, the pupil dilation remained at a maximum. For time-compressed speech and speech in quiet, even small differences in intelligibility associated with the acoustic-phonetic characteristics of the talkers were reflected in the pupil dilation response.

D. Limitations

First, intelligibility was not equal across conditions. Therefore, direct comparisons of pupil dilation between conditions should also consider the impact of differing intelligibility levels on pupil dilation. Another limitation was the small number of listeners assigned to each talker. This decision was due to constraints imposed on the experimental design by adaptation and pupillometry measures, as outlined in Sec. II. In particular, measures of pupil size are subject to listener variability, which could have affected talker averages and subsequently the correlation analyses. Variability between listeners for each measure is shown in Table II.

V. CONCLUSION

Our results show that talkers intelligible under noise-vocoding were also intelligible under time-compression. We therefore extend previous research that found this effect for noise-vocoded and masked speech (Bent *et al.*, 2009). We associated intelligibility with acoustic-phonetic talker profiles and found that VSD, mean energy in mid-range frequencies, and SR predicted intelligibility for noise-vocoded, masked, and time-compressed speech, respectively. Even though adaptation to noise-vocoded and time-compressed speech was observed, talker differences did not modulate the effect. Pupillometry findings indicated that some acoustic-phonetic features associated with intelligibility also related to listening effort. However, these findings were dependent on a condition's baseline intelligibility so that stronger correlations were only found in conditions with higher intelligibility, i.e., time-compressed speech and speech in quiet. Baseline pupil size changes were also affected by overall intelligibility, indicating a faster decline of attention for conditions with lowest and highest intelligibility, i.e., noise-vocoded speech and speech in quiet, respectively. The limitations of the current study should be taken into account in future studies. Such studies should either target specific intelligibility levels by employing adaptive procedures or cover a range of parameters for each degradation type, representing low and high intelligibility levels, respectively. In addition, a fewer number of trials in each block would allow for a larger number of talkers to be presented to each listener. This procedure might result in more accurate averages, possibly eliminating individual listener differences from talker-specific analyses.

ACKNOWLEDGMENTS

The authors would like to thank Outi Tuomainen for her help with acoustic analyses. The first author has received funding from the EU's H2020 research and innovation programme under the MSCA GA 675324 (ENRICH: www.enrich-etn.eu).

[1] For more information, see https://github.com/dmirman/gazer.

[2] See Supplementary Material at https://doi.org/10.1121/10.0001212 for talker group averages per acoustic feature.

Adank, P., and Janse, E. (**2009**). "Perceptual learning of time-compressed and natural fast speech," J. Acoust. Soc. Am. **126**(5), 2649–2659.

American National Standards Institute (**1997**). ANSI S3.79-1997, *Method for the Calculation of the Speech Intelligibility Index* (ANSI, New York).

Banks, B., Gowen, E., Munro, K. J., and Adank, P. (**2015**). "Cognitive predictors of perceptual adaptation to accented speech," J. Acoust. Soc. Am. **137**(4), 2015–2024.

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

Benoit, K. (**2018**). "quanteda: Quantitative Analysis of Textual Data," http://quanteda.io, doi:10.5281/zenodo.1004683 (Last viewed 5/4/2020).

Bent, T., Buchwald, A., and Pisoni, D. B. (**2009**). "Perceptual adaptation and intelligibility of multiple talkers for two types of degraded speech," J. Acoust. Soc. Am. **126**(5), 2660–2669.

Boersma, P., and Weenink, D. (**2018**). "Praat: Doing phonetics by computer (version 6.0.40) [computer software]," http://www.praat.org/ (Last viewed 5/4/2020).

Bradlow, A. R., and Bent, T. (**2008**). "Perceptual adaptation to non-native speech," Cognition **106**(2), 707–729.

Bradlow, A. R., Torretta, G. M., and Pisoni, D. B. (**1996**). "Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics," Speech Commun. **20**(3), 255–272.

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., and McGettigan, C. (**2005**). "Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences," J. Exp. Psychol. General **134**(2), 222–241.

De Looze, C., and Hirst, D. (**2008**). "Detecting changes in key and range for the automatic modelling and coding of intonation," in *Proceedings of Speech Prosody*, May 6–9, Campinas, Brazil, pp. 135–138.

Dupoux, E., and Green, K. (**1997**). "Perceptual adjustment to highly compressed speech: Effects of talker and rate changes," J. Exp. Psychol. Hum. Percept. Perform. **23**(3), 914–927.

Eisner, F., and McQueen, J. M. (**2005**). "The specificity of perceptual learning in speech processing," Percept. Psychophys. **67**(2), 224–238.

Erb, J., Henry, M. J., Eisner, F., and Obleser, J. (**2012**). "Auditory skills and brain morphology predict individual differences in adaptation to degraded speech," Neuropsychologia **50**(9), 2154–2164.

Fant, G. (**1973**). *Speech Sounds and Features* (MIT Press, Cambridge, MA).

Green, T., Katiri, S., Faulkner, A., and Rosen, S. (**2007**). "Talker intelligibility differences in cochlear implant listeners," J. Acoust. Soc. Am. **121**(6), EL223–EL229.

Greenwood, D. D. (**1990**). "A cochlear frequency position function for several species—29 years later," J. Acoust. Soc. Am. **87**(6), 2592–2605.

Grömping, U. (**2006**). "Relative importance for linear regression in R: The package relaimpo," J. Stat. Softw. **17**(1), 1–27.

Hazan, V., and Baker, R. (**2011**). "Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions," J. Acoust. Soc. Am. **130**(4), 2139–2152.

Hazan, V., and Markham, D. (**2004**). "Acoustic-phonetic correlates of talker intelligibility for adults and children," J. Acoust. Soc. Am. **116**(5), 3108–3118.

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

Hochmuth, S., Jürgens, T., Brand, T., and Kollmeier, B. (**2015**). "Talker- and language-specific effects on speech intelligibility in noise assessed with bilingual talkers: Which language is more robust against noise and reverberation?," Int. J. Audiol. **54**, 23–34.

Huckvale, M. (**2014**). "ProRec: A program for field workers (version 1.45) [computer software]," https://www.phon.ucl.ac.uk/ (Last viewed 5/4/2020).

Huyck, J. J., and Johnsrude, I. S. (**2012**). "Rapid perceptual learning of noise-vocoded speech requires attention," J. Acoust. Soc. Am. **131**(3), EL236–EL242.

Institute of Electrical and Electronics Engineers (**1969**). "IEEE recommended practices for speech quality measurements," IEEE Trans. Aud. Electroacoust **17**, 227–246.

Koch, X., and Janse, E. (**2016**). "Speech rate effects on the processing of conversational speech across the adult life span," J. Acoust. Soc. Am. **139**(4), 1618–1636.

Mattys, S. L., Davis, M. H., Bradlow, A. R., and Scott, S. K. (**2012**). "Speech recognition in adverse conditions: A review," Lang. Cogn. Process. **27**(7–8), 953–978.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. (**2017**). "Montreal forced aligner: Trainable text-speech alignment using Kaldi," in *Proceedings of Interspeech*, August 20–24, Stockholm, Sweden, pp. 498–502.

McGarrigle, R., Munro, K. J., Dawes, P., Stewart, A. J., Moore, D. R., Barry, J. G., and Amitay, S. (**2014**). "Listening effort and fatigue: What exactly are we measuring? A British Society of Audiology Cognition in Hearing Special Interest Group 'white paper'," Int. J. Audiol. **53**(7), 433–440.

McLaughlin, D. J., and Van Engen, K. J. (**2020**). "Task-evoked pupil response for accurately recognized accented speech," J. Acoust. Soc. Am. **147**(2), EL151–EL156.

Menard, S. (**1995**). *Applied Logistic Regression Analysis* (Sage, Thousand Oaks, CA).

Mirman, D. (**2014**). *Growth Curve Analysis and Visualization Using R* (Chapman and Hall/CRC, London).

Morimoto, M., Sato, H., and Kobayashi, M. (**2004**). "Listening difficulty as a subjective measure for evaluation of speech transmission performance in public spaces," J. Acoust. Soc. Am. **116**(3), 1607–1613.

Moulines, E., and Charpentier, F. (**1990**). "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," Speech Commun. **9**, 453–467.

Peelle, J. E., and Wingfield, A. (**2005**). "Dissociations in perceptual learning revealed by adult age differences in adaptation to time-compressed speech," J. Exp. Psychol. Hum. Percept. Perform. **31**(6), 1315–1330.

Picheny, M. A., Durlach, N. I., and Braida, L. D. (**1986**). "Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech," J. Acoust. Soc. Am. **29**(1), 434–446.

Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W., Humes, L. E., Lemke, U., Lunner, T., Matthen, M., Mackersie, C. L., Naylor, G., Phillips, N. A., Richter, M., Rudner, M., Sommers, M. S., Tremblay, K. L., and Wingfield, A. (**2016**). "Hearing impairment and cognitive energy: The framework for understanding effortful listening (FUEL)," Ear Hear. **37**, 5S–27S.

Rennies, J., Best, V., Roverud, E., and Kidd, G. (**2019**). "Energetic and informational components of speech-on-speech masking in binaural speech intelligibility and perceived listening effort," Trends Hear. **23**, 1–21.

Samuel, A. G., and Kraljic, T. (**2009**). "Perceptual learning for speech," Atten. Percept. Psychophys. **71**(6), 1207–1218.

Schwarz, G. (**1978**). "Estimating the dimension of a model," Ann. Stat. **6**(2), 461–464.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**(5234), 303–304.

Simantiraki, O., Cooke, M., and King, S. (**2018**). "Impact of different speech types on listening effort," in *Proceedings of Interspeech*, September 2–6, Hyderabad, India, pp. 2267–2271.

Unsworth, N., and Robison, M. K. (**2016**). "Pupillary correlates of lapses of sustained attention," Cogn. Affect. Behav. Neurosci. **16**(4), 601–615.

Versfeld, N. J., and Dreschler, W. A. (**2002**). "The relationship between the intelligibility of time-compressed speech and speech in noise in young and elderly listeners," J. Acoust. Soc. Am. **111**(1), 401–408.

Wagner, A. E., Nagels, L., Toffanin, P., Opie, J. M., and Başkent, D. (**2019**). "Individual variations in effort: Assessing pupillometry for the hearing impaired," Trends Hear. **23**, 1–18.

Wendt, D., Koelewijn, T., Książek, P., Kramer, S. E., and Lunner, T. (**2018**). "Toward a more comprehensive understanding of the impact of masker type and signal-to-noise ratio on the pupillary response while performing a speech-in-noise test," Hear. Res. **369**, 67–78.

Winn, M. B., Wendt, D., Koelewijn, T., and Kuchinsky, S. E. (**2018**). "Best practices and advice for using pupillometry to measure listening effort: An introduction for those who want to get started," Trends Hear. **22**, 1–32.