The time course of adaptation to distorted speech

Cooke, Martin; Scharenborg, Odette; Meyer, Bernd T.

DOI
10.1121/10.0010235

Publication date
2022

Document Version
Final published version

Published in
Journal of the Acoustical Society of America

The time course of adaptation to distorted speech

Martin Cooke, Odette Scharenborg and Bernd T. Meyer

Citation: The Journal of the Acoustical Society of America **151**, 2636 (2022); doi: 10.1121/10.0010235
View online: https://doi.org/10.1121/10.0010235
View Table of Contents: https://asa.scitation.org/toc/jas/151/4
Published by the Acoustical Society of America

JASA ARTICLE

The time course of adaptation to distorted speech

Martin Cooke,[1,a)] Odette Scharenborg,[2] and Bernd T. Meyer[3]
[1]*Ikerbasque (Basque Science Foundation), Bilbao, Spain*
[2]*Multimedia Computing Group, Delft University of Technology, Netherlands*
[3]*Communication Acoustics and Cluster of Excellence Hearing4all, Carl von Ossietzky University, Oldenburg, Germany*

ABSTRACT:
When confronted with unfamiliar or novel forms of speech, listeners' word recognition performance is known to improve with exposure, but data are lacking on the fine-grained time course of adaptation. The current study aims to fill this gap by investigating the time course of adaptation to several different types of distorted speech. Keyword scores as a function of sentence position in a block of 30 sentences were measured in response to eight forms of distorted speech. Listeners recognised twice as many words in the final sentence compared to the initial sentence with around half of the gain appearing in the first three sentences, followed by gradual gains over the rest of the block. Rapid adaptation was apparent for most of the eight distortion types tested with differences mainly in the gradual phase. Adaptation to sine-wave speech improved if listeners had heard other types of distortion prior to exposure, but no similar facilitation occurred for the other types of distortion. Rapid adaptation is unlikely to be due to procedural learning since listeners had been familiarised with the task and sentence format through exposure to undistorted speech. The mechanisms that underlie rapid adaptation are currently unclear. https://doi.org/10.1121/10.0010235
(Received 15 January 2022; revised 9 March 2022; accepted 25 March 2022; published online 19 April 2022)
[Editor: Melissa Michaud Baese-Berk] Pages: 2636–2646

I. INTRODUCTION

When faced with unfamiliar and challenging forms of speech, listeners have been observed to improve with increasing exposure. This ability to adapt to previously unseen stimuli has been measured in response to many types of speech. Examples include speech colored by unknown reverberation (Brandewie and Zahorik, 2010), noisy speech (Bent *et al.*, 2009), speech produced by multiple talkers (Kato and Kakehi, 1988), speech spoken with a foreign accent (Bradlow and Bent, 2008; Clarke and Garrett, 2004; Melguy and Johnson, 2021), or resulting from text-to-speech synthesis (Lehet *et al.*, 2020). Listeners also adapt to more extreme levels of variation such as those that characterise stimuli that have been severely band limited (Warren *et al.*, 1995), heavily time-compressed (Dupoux and Green, 1997), spectrally shifted (Rosen *et al.*, 1999), noise-vocoded (Davis *et al.*, 2005), tone-vocoded (Hervais-Adelman *et al.*, 2011), reduced to sine waves (Bent *et al.*, 2011), restricted to sparse time-frequency locations (Ahmadi *et al.*, 2013), or spectrally rotated (Azadpour and Balaban, 2015). Adaptation has also been reported in cochlear-implant simulations (Dorman *et al.*, 1997) and for locally time-reversed speech (Saberi and Perrott, 1999).

In the current article, the term "adaptation" will be used to describe any intelligibility gain resulting from increasing exposure, without assuming a specific locus for the process(es) responsible for the increase in intelligibility. This definition is intended to be broad enough to cater for mechanisms, such as the perceptual compensation processes, believed to be responsible for handling changes in room acoustics (e.g., Brandewie and Zahorik, 2010; Ladefoged and Broadbent, 1957), as well as potentially slower processes such as lexically driven perceptual learning (e.g., Norris *et al.*, 2003; Samuel and Kraljic, 2009).

One outstanding issue concerns the speed at which listeners are able to adapt. While adaptation is sometimes described as "rapid" (e.g., Azadpour and Balaban, 2015; Clarke and Garrett, 2004; Rotman *et al.*, 2020) or resulting from "brief" exposure (e.g., Banai and Lavner, 2014), these terms are often used rather loosely to refer to changes between the first and subsequent *blocks* of stimuli. Such blocks have variously consisted of, for instance, 5 sentences (for fast speech; Dupoux and Green, 1997), 10 sentences (for speech heard through narrow spectral slits; Warren *et al.*, 1995), or 20 sentences (for noisy and cochlear-implant simulation speech; Bent *et al.*, 2009). Analysing stimuli by blocks produces a heavily quantised view of the time course of adaptation and risks very fast adaptation going undetected or the amount of adaptation being underestimated (cf. Zhang *et al.*, 2019, in vision). For example, if adaptation occurs solely on the first stimulus in a block, it will make little contribution to the mean score across the entire block. A similar concern is raised by Banai and Lavner (2014) who argue that a typical pretest/exposure/posttest regime will fail to observe any improvements that occur within the pretest phase.

[a)]Electronic mail: m.cooke@ikerbasque.org

https://doi.org/10.1121/10.0010235

Rotman *et al.* (2020) used linear fits to the proportion of words recognised in blocks as small as two sentences in their study of fast speech, but a few studies have reported scores as a function of individual sentence presentation order without blocking subsets of sentences. Davis *et al.* (2005) (experiment 1) demonstrated that adaptation to noise-vocoded speech occurred across a sequence of 30 sentences, but although they analysed outcomes at the level of single sentences, they used 2 counterbalanced orders, which limits the degree to which item- and trial-effects can be separated. Also using noise-vocoded speech, Erb *et al.* (2013) collected sentence scores across 100 trials using a somewhat larger cohort, again finding significant variability, but supporting a linear rather than power-law interpretation of the growth of intelligibility with sentence position. Van Hedger *et al.* (2019) described sentence-by-sentence responses collected from 3 distortion types (noise-vocoded, sine-wave, and time-compressed), as well as accented speech and speech in 12-talker babble, and demonstrated nonlinear adaptation that took a logarithmic form as a function of time, with larger gains in intelligibility for the initial sentences in each block.

Intriguingly, some behavioural studies hint at the possibility that listeners are capable of far more rapid adaptation than has been measured to date. A study of adaptation to accented speech showed gains in processing efficiency (reaction times in a cross-modal matching task) after just 2–4 sentences (Clarke and Garrett, 2004). Brandewie and Zahorik (2010) demonstrated that exposure to just two sentences carrying information about room characteristics was sufficient to produce large gains in intelligibility for a subsequent sentence with the same reverberation characteristics. Similarly rapid effects were observed in a subsequent study using high-variability sentences (Srinivasan and Zahorik, 2013). Adaptation to the speech of different talkers has been shown to require a small number of tokens (Kakehi, 1992; Kato and Kakehi, 1988). Using nonsense monosyllables from 100 different talkers, Kato and Kakehi measured the effect of changing the number of consecutive items produced by the same talker within a sequence of monosyllables, discovering that just five were required to reach a performance asymptote. Similarly, García Lecumberri *et al.* (2015) found that per-talker judgments of degree of foreign accent from spontaneous child speech stabilised after exposure to fewer than four words from each talker.

The main aim of this study is to provide a fine-grained characterisation of adaptation to multiple forms of distorted speech with particular emphasis on listeners' responses after very small amounts of exposure. The resulting detailed picture of the time course of adaptation will provide constraints on the types of processes that underlie ongoing improvements as a function of exposure to novel forms of speech.

A further goal is to determine whether listeners display similar forms of adaptation to different types of distortion, since any commonalities in the adaptation response may point to the existence of generic processes used by listeners to handle variation. Apart from Van Hedger *et al.* (2019), few studies have examined multiple distortion types under common listening conditions using the same listeners, speech materials, and task. Here, many distinct forms of distortion are involved, including some for which adaptation has not yet been established and which collectively result in very different forms of degradation to the acoustic content of speech. Testing multiple forms of distortion using the same listeners in a single experimental sitting will also shed light on whether exposure to one form of novel speech is beneficial when processing subsequent forms of speech or whether adaptation performance is independent of prior experience with distorted speech.

In the current study, listeners were first familiarised with the task, talker, and speech materials via undistorted exemplars to remove any procedural learning effects, and then went on to identify keywords in sentences processed by the eight forms of distortion listed in Table I. These forms were chosen as representatives of distortions produced by degrading acoustic information primarily in the temporal (REVERSED, FAST), spectral (NARROWBAND, NOISE-VOCODED, and TONE-VOCODED), and spectro-temporal domains (SINE-WAVE, GLIMPSED, and SCULPTED). All of the listeners heard all eight types of distortion, but the order of presentation was varied across listeners to measure any impact of prior exposure to a different form of distorted speech. Section II defines the choice of parameters and generation procedure for each type of distortion. Section III describes the listening experiment whose outcomes are reported in Sec. IV.

II. DISTORTED SPEECH

A. Choice of distortion parameters

Each of the eight types of distortion listed in Table I represents a family of signals whose intelligibility will depend on the choice of parameters used to generate the distortion. While a natural criterion for parameter selection is to equalise intelligibility across distortions, achieving this goal can be problematic in the context of an adaptation study. One reason is that not all distortion types permit fine-grained parametric control. For sine-wave speech, the choice is effectively one of using either two or three formants, and for the two vocoded conditions, small changes in the number of spectral bands can result in quite large changes in intelligibility. Moreover, it is not clear whether

TABLE I. The types of distorted speech and associated parameter values used in the current study.

Distortion	Parameter settings
FAST	2.5 times speedup
REVERSED	62 ms window
SINE-WAVE	$F1$ and $F2$ frequencies and amplitudes
NOISE-VOCODED	Six-bands, 100–7500 Hz
TONE-VOCODED	Six-bands, 100–7500 Hz
NARROWBAND	Centre, 2000 Hz, fifth-order, 1/3-octave
GLIMPSED	Local SNR, 3 dB; global SNR, 0 dB
SCULPTED	Local SNR, −6 dB; global SNR, 0 dB

the criterion should be to equalise baseline (i.e., initial) intelligibility, mean across-block intelligibility, or asymptotic intelligibility (i.e., an estimated upper bound at the end of the exposure period). There is also the possibility that equalising on mean intelligibility could lead to floor or ceiling effects for the baseline and asymptotic rates, respectively, for some distortion types.

For this study, these considerations motivated a policy of selecting parameters that simply avoided floor and ceiling effects for each individual distortion type, to enable the shape of the adaptation time course to be seen. Where available, initial parameter estimates were chosen based on intelligibility scores from previous studies (detailed in Sec. II B). Subsequently, for all conditions with continuous parameterisations (i.e., all except SINE-WAVE, NOISE-VOCODED, and TONE-VOCODED, where values from the literature were adequate), distorted speech samples for a 25-step continuum of parameter values were generated for a subset of sentences, and a value midway between understanding no words and understanding the entire sentence was chosen by three native Spanish speakers using *SpeechAdjuster* (Simantiraki and Cooke, 2021), a tool that allows the user to explore any parameter space via a simple control knob while continuously evaluating the audio. Finally, to check for floor and ceiling effects, 3 further native Spanish listeners identified keywords in 30 sentences in each of the 8 distortion conditions using the chosen parameter values. None of the participants involved in the pilot stages took part in the main experiment. The final parameters are shown in Table I.

B. Distorted speech generation

The processes used to create each distorted speech condition are detailed below. The source speech materials (described further in Sec. III B) were Spanish sentences from a male talker, sampled at 16 kHz.

For the FAST condition, the rate was increased by a factor of 2.5 (i.e., compressed to 40% of its original duration) using PVOC, a MATLAB implementation of a fast Fourier transform (FFT)-based phase vocoder (Ellis, 2022), with a FFT window of size 512 samples, equivalent to 32 ms.

Following the original study into the intelligibility of locally time-reversed speech (Saberi and Perrott, 1999), the REVERSED condition was generated by time-reversing successive nonoverlapping segments of speech, unwindowed, using a segment duration of 62 ms.

For the SINE-WAVE condition, sine-wave formants were generated using first and second formant ($F1$,$F2$) frequencies and their amplitudes. The frequencies of the lowest five formants were estimated every 10 ms using the Burg algorithm as implemented in Praat (Boersma, 2001) with a maximum formant frequency of 5500 Hz. The formant amplitudes for $F1$ and $F2$ were derived by looking up values at the corresponding spectro-temporal locations in a broadband spectrogram with a frame size of 16 ms. Each sine-wave formant was then constructed by (i) generating instantaneous frequency and amplitude estimates at the target 16 kHz sampling frequency via linear interpolation, (ii) computing the sine of the cumulative sum of the instantaneous frequency, and (iii) weighting the resulting frequency-varying sinusoid by the instantaneous amplitude.

The NOISE-VOCODED condition was generated based on a procedure similar to that of Davis *et al.* (2005). Speech was filtered into six bands with cutoff frequencies equally spaced on an equivalent rectangular bandwidth (ERB)-rate scale in the range of 100–7500 Hz, leading to band edges at 100, 328, 713, 1365, 2469, 4338, and 7500 Hz. Each filter was a zero-phase, fifth-order Butterworth implemented in second-order sections using the butter and sosfilt methods from the Python scipy.signal module. Instantaneous envelopes were computed by convolving the squared output from each filter with a 64 ms Kaiser window with $\beta = 20$ and taking the square root. The envelope was then multiplied by a noise carrier that resulted from passing a Gaussian random noise signal through the same filter. Finally, NOISE-VOCODED stimuli were formed by summing across the six filters.

Generation of the TONE-VOCODED condition followed the same procedure as that for noise-vocoding apart from the use of a tonal carrier instead of a noise carrier. The frequency of the tonal carrier was set to the geometric mean of the upper and lower cutoff frequencies in each filter.

A narrow spectral band of speech that constituted the NARROWBAND condition was generated by filtering speech through a third-octave filter centred at 2 kHz. The filter was a fifth-order zero-phase Butterworth, implemented as described in the NOISE-VOCODED condition.

The GLIMPSED condition involved resynthesising spectro-temporal regions of speech that survived masking by a speech-shaped noise signal when mixed at a global signal-to-noise ratio (SNR) of 0 dB. Glimpse synthesis involves two steps (Cooke and García Lecumberri, 2020). In the analysis step, auditory spectrograms for the speech signal and masker are generated independently by passing these signals through a 55-channel bank of gammatone filters with centre frequencies equally spaced on an ERB-rate scale in the range 100–7500 Hz. The Hilbert envelope at the output of each filter is computed, smoothed with a temporal integrator with an 8 ms time constant, downsampled to 10 ms frames. Here, glimpses were defined as those time-frequency regions in the auditory spectrogram where the speech exceeds the masker by at least 3 dB. In the resynthesis step, the speech-plus-noise mixture is passed through the same gammatone filterbank and a triangular window is applied to each filter centred on those time frames whenever a glimpse exists. The windowed signals are then summed to produce the resynthesised signal. To remove artefacts due to differential phase delays across filters, the signal is passed through the filterbank, reversed, passed through a second time, and reversed.

Finally, sculpted speech is a form of distortion derived by passing an arbitrary signal (typically nonspeech) through a time-frequency mask, a process that for certain carrier signals results in intelligible but clearly distorted speech (Cooke and García Lecumberri, 2020). In this way, the target speech information is conveyed solely by the glimpse

https://doi.org/10.1121/10.0010235

locations rather than the speech signal itself. In the current study, a musical carrier was chosen as that results in stimuli that are less speechlike than produced, for example, by stationary noise carriers. The resynthesis stage used randomly chosen fragments from a 171 s recording of an operatic work extracted from a digital archive.[1] The procedure for generating the SCULPTED condition was similar to that for the GLIMPSED speech. The analysis phase was identical but following pilot testing the criterion for selecting which regions are considered as glimpsed was taken to be those where the local SNR exceeded −6 dB rather than +3 dB.

Figure 1 provides an illustration of the eight distorted forms alongside the original undistorted sentence.

III. EXPERIMENT: TIME COURSE OF ADAPTATION TO DISTORTED SPEECH

A. Listeners

A cohort of 69 listeners participated in the experiment. A sample size in the range of 60–70 was indicated by a power analysis using sentence-level standard deviation (s.d.) estimates from our earlier study into sculpted speech (Cooke and García Lecumberri, 2020), one of the conditions of this study, under the assumption that similar variability would be present for the other distortion conditions. A subsequent analysis indicated that this assumption was valid.

All of the listeners were native speakers of Spanish or bilingual in Spanish and Basque. The listeners were paid for their participation and audiometrically screened at octave frequencies in the range 125–8000 Hz in each ear with a criterion of 20 dB hearing level (HL) or better at each frequency. The results from one participant were later excluded from analysis due to elevated thresholds. The remaining 68 participants had ages in the range 19–29 years old (mean 21.4 years old) and 8 were male.

B. Stimuli

Undistorted speech materials were drawn from the male talker of the Sharvard Corpus (Aubanel *et al.*, 2014). Sharvard sentences are Spanish analogues to the Harvard sentences (Rothauser *et al.*, 1969) and each contains exactly five keywords that are used in intelligibility scoring. A typical sentence is "Lleva el cubo a la pared para regar las flores" ("carry the bucket to the wall to water the flowers"), where the keywords are underlined. All signals were sampled at 16 kHz. Each of the 8 distortion methods described in Sec. II was applied to 240 sentences (Sharvard sentence numbers 241–480). These sentences have a mean duration of 2.46 s (range, 1.69–3.64; s.d., 0.32 s). In all conditions, the sentences were presented in the absence of added noise. The stimuli were calibrated to a presentation level of 72.3 ± 0.3 dB(A) in all conditions using a B and K sound level meter (model 2250-L, Brüel & Kjær, Nærum, Denmark) and B and K Artificial Ear (model 4153). All stimuli are available for downloading at a permanent repository.[2]

C. Procedure

The stimuli were presented blocked by distortion condition, with 30 sentences per block. Previous studies (e.g., Davis *et al.*, 2005; Dupoux and Green, 1997; Warren *et al.*, 1995) have shown significant adaptation effects across 20–30 sentences, and although it is possible that adaptation continues beyond this point for some types of distortion, the main features of the time course of adaptation were felt to be observable within a block of this size. Block ordering followed a balanced Latin square design and the participants were assigned to one of the resulting eight block orderings in consecutive fashion. The sentence order within each block was randomised.

Prior to the experiment, participants were told that they would hear distorted Spanish sentences in eight distinct blocks and that each block would contain a different type of distortion. They were encouraged to type as many words as they could make out even if they were not certain.

While most speech perception experiments make use of familiarisation sessions to eliminate learning effects, our requirement was the opposite: a participant's initial encounter with each distortion type had to occur within the experimental block itself. However, some form of familiarisation was necessary to ensure that any observed gains in intelligibility

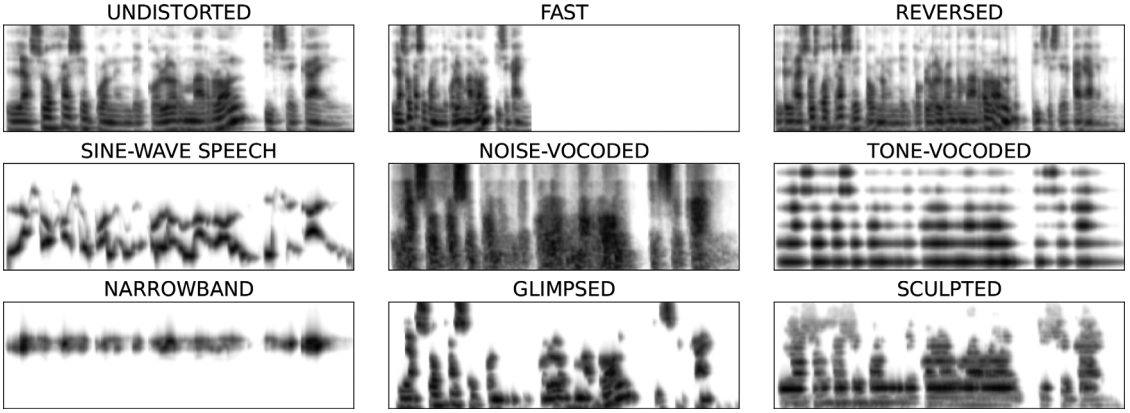

FIG. 1. Auditory spectrograms for undistorted and eight distorted speech conditions for an example sentence, "Las hojas se ponen de color marrón y se caen" ("The leaves turn brown and fall"). The frequency axis runs from 0 to 7500 Hz on an ERB-rate scale while the time axis corresponds to 2.625 s.

were due to adaptation on the stimuli themselves and not based on procedural or task learning, such as learning how to use the interface or becoming familiar with the type of sentences or the voice of the speaker (e.g., Ortiz and Wright, 2009; Robinson and Summerfield, 2006). For this reason, immediately prior to the main experiment, listeners were familiarised with the task, interface, and speaker by responding to a block of 15 undistorted sentences (Sharvard sentence numbers 481–495). No feedback on responses was provided at any point; whereas feedback has been shown to increase the speed of learning (Davis *et al.*, 2005), adaptation is still possible in its absence (Erb *et al.*, 2012); see also a recent review of adaptation by Bieber and Gordon-Salant (2021) in which studies that involved "passive" (no feedback) and "active" (with feedback) learning are distinguished.

The experiment took place in a sound-attenuating speech perception facility in the Language and Speech Laboratory at the University of the Basque Country (Alava Campus, Vitoria-Gasteiz, Spain). Between one and six participants were tested simultaneously in acoustically separated workstations within the laboratory. Listeners heard stimuli over Sennheiser HD380 Pro headphones (Wedemark, Germany) and typed their responses into an on-screen text input box. Stimulus presentation and response collection were under the control of a custom software application running on MacMini computers (Cupertino, CA). Listeners were able to visualise the number of remaining stimuli in each block of the experiment via a progress bar. The experiment was self-paced, and listeners were encouraged to take a short break between each of the blocks. The experiment required around 45 min to complete.

D. Postprocessing

The final dataset consists of 16 320 responses made up of 240 sentence responses from 68 listeners. All responses were manually inspected and any very obvious typographic errors were corrected. Punctuation and vowel stress marks were removed prior to automatic scoring. Each response resulted in an integer score in the range of 0–5, representing the number of correctly identified keywords. Where not specified explicitly, percentages referred to in the text represent averages (rather than medians) across listeners or sentences.

IV. RESULTS

A. Familiarisation phase

On average, listeners identified 96% of words in the first familiarisation sentence correctly, a figure rising to 99% by the second sentence and remaining at close to ceiling for the remainder of the 15-sentence familiarisation phase. This outcome suggests that the task induced negligible procedural learning and speaker adaptation effects.

B. Intelligibility for each type of distorted speech

Taking all conditions together, listeners identified 59.1% of all keywords (range, 34.1–76.5; s.d., 8.7). Across each block, intelligibility was highest in the FAST condition (82.2%) and lowest in the SINE-WAVE condition (40.7%). Figure 2 reveals that large individual differences in mean scores are present, particularly in the SINE-WAVE and TONE-VOCODED conditions, where some participants identified very few keywords in the entire block while others correctly identified around four in every five keywords. Figure 2 also encodes the order in which participants heard the block containing each distortion.

C. Sentence position

When combined across distortion conditions, keyword scores increased with sentence position (Fig. 3). Listeners recognised twice as many keywords in the final sentence of the block compared to the initial sentence (67.5% vs 33.7%). Intelligibility increased rapidly over the first three or four sentences and then rose more gradually with sentence position. We will refer to these as the "rapid" and "gradual" phases. Note that these are purely descriptive terms and should not be taken to imply the existence of different underlying mechanisms.

The rapid-then-gradual pattern is reflected in most of the individual types of distortion (Fig. 4) but is absent in the GLIMPSED condition. Although intelligibility gains for most distortions have levelled off by the end of the block, intelligibility in the TONE-VOCODED condition, and to a lesser extent in the SINE-WAVE condition, continues to rise. The largest overall gains (quantified in Table II) are observed for these two most difficult conditions, but there is no clear relationship between the amount of adaptation and ultimate

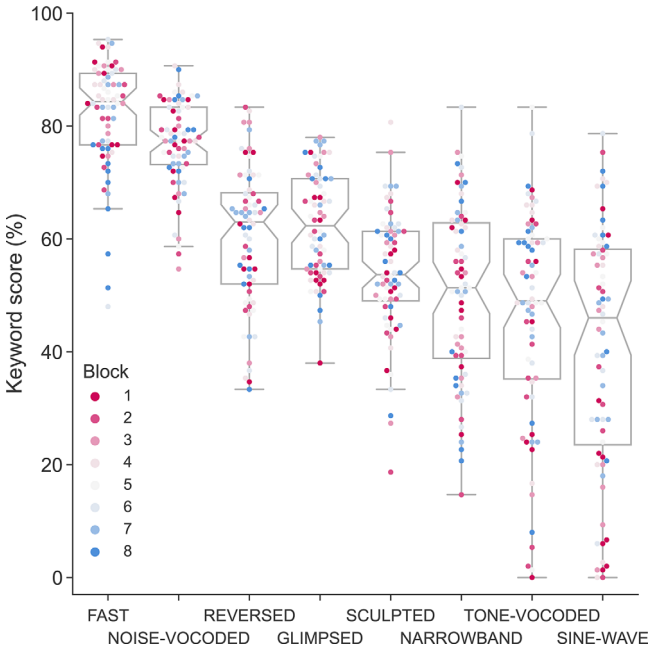

FIG. 2. (Color online) Keyword score statistics for each type of distortion. The boxes depict quartiles of the score distribution, whiskers extend to 1.5 of the inter-quartile range, and notches denote 95% confidence intervals. Each dot represents the score for one participant; the dot color indicates the block in which the participant heard the condition.

https://doi.org/10.1121/10.0010235

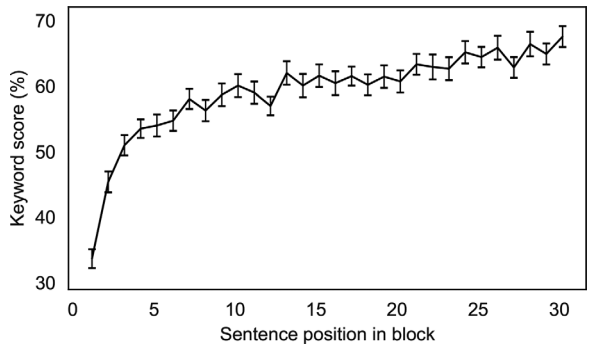

FIG. 3. Keyword scores averaged across distortion conditions as a function of sentence position (solid line). The error bars indicate ±1 standard error.

intelligibility with larger across-block gains for the two conditions with the highest rate at the end of the block (FAST and NOISE-VOCODED) than for three of the conditions with intermediate gains (REVERSED, GLIMPSED, and SCULPTED).

The time courses of adaptation overall and for the individual distortion types were analysed by comparing goodness-of-fits for a range of three-parameter, continuous growth curve functions (piecewise-linear, power, exponential error decay, hyperbolic, and logarithmic). Fits were evaluated using the Levenberg-Marquardt algorithm as implemented in the scipy.optimize.curve_fit method in Python, taking account of the variance at each data point. In all cases, a logarithmic function provided the best or equally best fit. Additionally, to define the rapid and gradual phases in what follows, we determined the best two-component piecewise-linear fit for each distortion, computed with the pwlf package (Jekel and Venter, 2019). The fit was weighted by the inverse of the s.d. at each point. The vertical lines in Fig. 4 indicate the locations of the breakpoints. In the FAST condition, the fact that the location of the breakpoint does not correspond with a visual impression is because of the use of a weighted fit: an unweighted fit would place the breakpoint at around sentence 3.

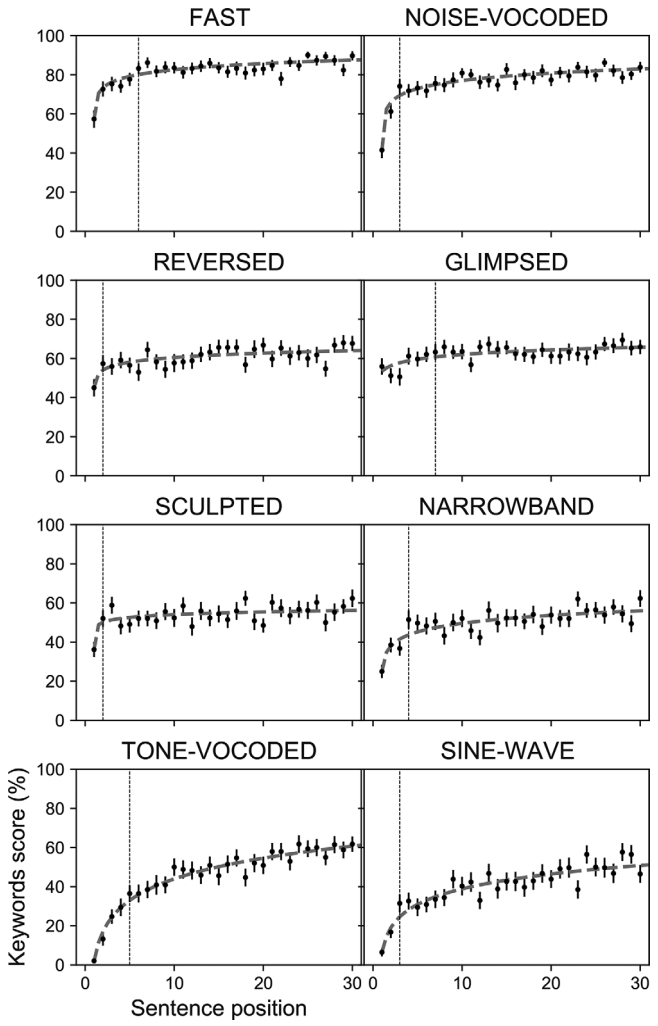

FIG. 4. The time course of adaptation to each individual form of distorted speech. The error bars indicate ±1 standard error. The best logarithmic fit to each distortion is shown as a gray dashed line. Vertical dotted lines indicates the breakpoint for the optimal two-component piecewise-linear fit.

TABLE II. Quantitative summary of the time course of adaptation. "First" represents the score for the initial sentence. Gain columns show percentage point (p.p.) improvements overall and during the rapid and gradual phases. "Break" indicates the sentence position where the two-component linear fit is optimally broken. Slopes for the rapid and gradual phases are in units of p.p./sentence. The t_{50} value is the sentence number at which 50% of the overall gain is reached, estimated using the logarithmic fit. RMSE indicates the root mean square errors for the log and piecewise-linear (*two-lin*) fits. The final columns provide coefficients of the logarithmic fit for a function of the form $y = a\ \log_e(x + b) + c$, where t represents the sentence position and y is the score at that position.

		Gain in p.p.				Slope			RMSE		log		
Type	First (%)	All	Rapid	Gradual	Break	Rapid	Gradual	t_{50}	log	two-lin	a	b	c
ALL	33.6	31.4	16.6	14.8	3	10.0	0.47	2.7	1.04	1.44	5.7	−0.89	46.0
FAST	57.3	30.2	22.8	7.4	6	3.5	0.23	1.8	2.13	3.02	4.2	−0.98	73.3
NOISE-VOCODED	41.4	41.5	27.9	13.6	3	15.6	0.35	1.5	2.04	2.42	5.1	−0.99	65.8
REVERSED	45.1	18.8	8.9	9.9	2	12.2	0.29	2.2	3.02	3.51	3.0	−0.95	53.9
GLIMPSED	53.4	12.3	7.4	4.9	7	1.8	0.10	4.9	2.57	2.75	3.3	−0.27	54.4
SCULPTED	36.2	20.0	13.9	6.0	2	15.3	0.22	1.1	3.08	3.65	1.8	−1.00	50.1
NARROWBAND	25.0	31.0	18.7	12.3	4	6.7	0.39	2.6	3.08	3.69	5.5	−0.90	37.4
TONE-VOCODED	1.9	58.8	31.0	27.7	5	9.3	0.91	4.5	2.35	3.13	14.8	−0.44	10.4
SINE-WAVE	6.4	44.4	18.2	26.2	3	12.6	0.84	4.1	3.48	4.09	10.5	−0.57	15.2

Table II provides a numerical summary of the logarithmic and piecewise-linear fits, along with a number of parameters derived from the split into rapid and gradual phases. In particular, the overall change in intelligibility in each condition is divided into components for the rapid and gradual phases, and the slopes of the two phases (in percentage points per sentence) are estimated.

All distortion types except SINE-WAVE and REVERSED resulted in numerically larger gains in the rapid phase than in the gradual phase in spite of the much smaller number of sentences during the rapid phase, whose boundary is shown by the "break" column in Table II. Across all distortions, the scores improved by 10.0 percentage points for each sentence heard during the rapid phase compared to 0.47 points per sentence in the gradual phase.

A generalised linear mixed-effects model was constructed to predict the proportion of keywords identified correctly in each trial. The model contained fixed effects of CONDITION and BLOCK, by-subject random intercepts and per-condition slopes, and by-sentence random intercepts. Since the log-transformed position of the sentence in the block was a highly significant predictor in its own right ($\chi^2(1) = 1804,\ p < 0.001$), it was added as a covariate to all of the models. Model estimation was via the glmer function of the lme4 package (Bates *et al.*, 2015) in R (R Core Team, 2021), using the nloptwrap optimizer setting. The importance of retaining factors in the interactions and main effects was determined by model comparison using the anova function. The resulting minimally adequate model contained a significant effect of CONDITION ($\chi^2(7) = 174,\ p < 0.001$) and a CONDITION by BLOCK interaction ($\chi^2(49) = 105,\ p < 0.001$) but no overall block effect ($\chi^2(7) = 4.7,\ p = 0.70$). *Post hoc* comparison of scores in the first and last blocks indicated that only the SINE-WAVE ($p < 0.01$) and FAST ($p < 0.05$) conditions were influenced by prior exposure to other distortion types, with a detrimental effect in the latter case (Fig. 5). These outcomes are supported by Fig. 2: 7 of the 17 scores in the bottom quartile for the SINE-WAVE condition came from participants who were exposed to this condition in the first or second block. Conversely, 8 of 15 lower quartile scores for FAST speech occurred in the final block.

D. Per-listener rapid and gradual phase slopes

The fact that, by design, each listener heard only one sentence at each position in the block precludes robust per-listener characterisation of the rapid and gradual phases for each distortion. However, more robust piecewise-linear fits describing each individual's adaptation pattern can be obtained by combining that listener's results across distortion types. For consistency of interpretation, rather than using a different breakpoint for each individual's data, a common breakpoint at sentence 3 was adopted, corresponding to the cohort mean across distortion breakpoint (Table II).

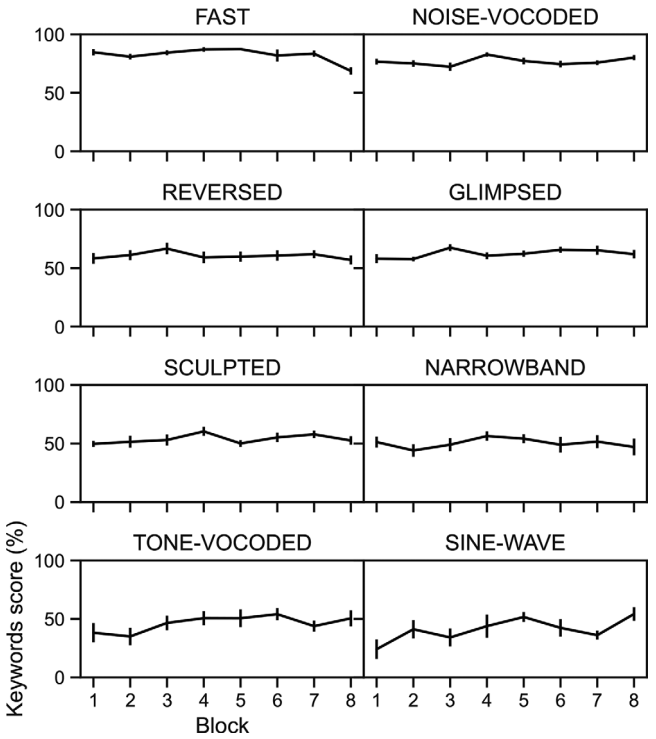

FIG. 5. Keyword scores as a function of the position of the block in the experiment in which listeners heard them. The error bars indicate ±1 standard errors.

Figure 6 summarises the resulting slopes of the rapid and gradual components of the fit. Apart from three participants, the slope during the rapid phase was substantially larger than during the gradual phase (note the differing scales on the axes), indicating that although a high degree of individual variability is present in response to distorted speech (e.g., as manifest in Fig. 2), the rapid-then-gradual property is a robust effect at the individual as well as the

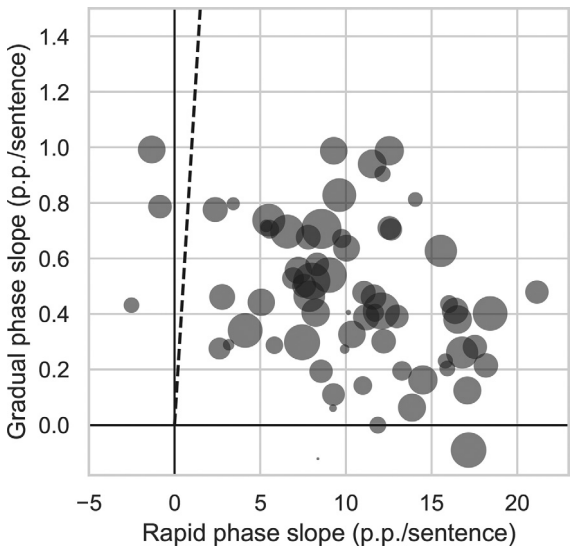

FIG. 6. Rapid vs gradual phase slopes. Each disk corresponds to an individual participant; disk area is proportional to that participant's mean score. The region to the right of the diagonal line corresponds to locations where the rapid phase slope is greater than that in the gradual phase.

https://doi.org/10.1121/10.0010235

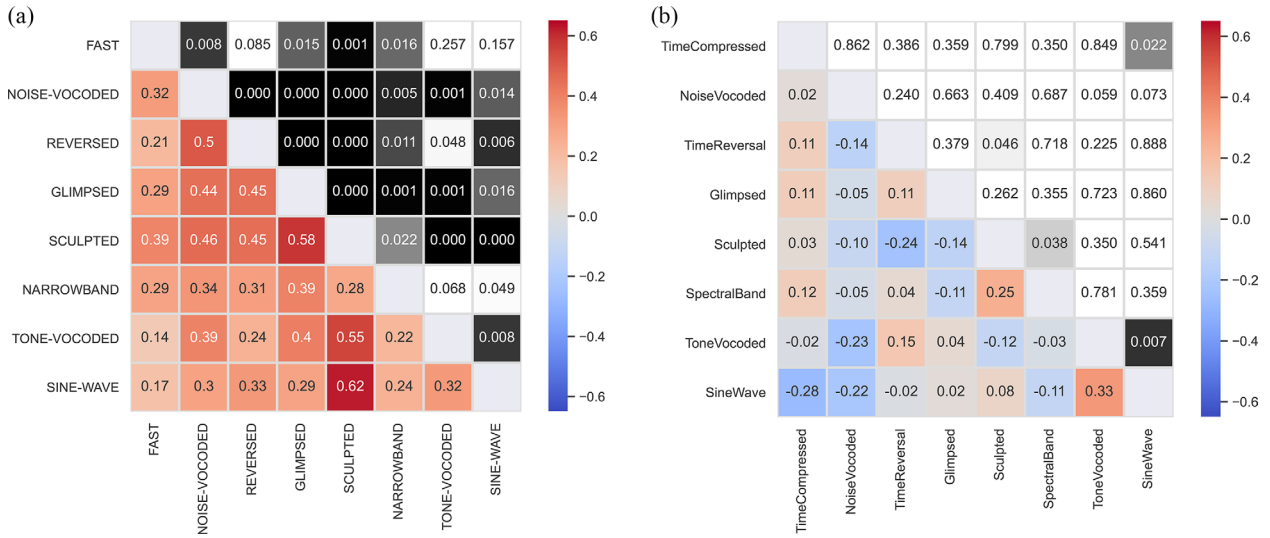

FIG. 7. (Color online) Pearson correlation coefficients (lower triangle) and corresponding p-values (upper triangle) for absolute scores (upper panel) and slopes (lower panel).

cohort level. Mean scores per participant did not correlate with the slope in either the rapid ($r = 0.11, p = 0.38$) or gradual ($r = 0.21, p = 0.09$) phases, suggesting that the shape of the adaptation curve is independent of absolute intelligibility.

E. Correlations between distortion types

Absolute keywords correct scores were positively correlated in all 28 pairs of conditions (upper panel of Fig. 7), suggesting that listeners who cope well with one form of distortion are also better able to identify words in other distortion types. However, because the main focus here is on whether adaptation to one form of distortion was predictive of adaptation to another, a correlation analysis was performed on the slopes of the best linear fit of by-participant scores as a function of the log of sentence position. The resulting correlation coefficients (lower panel of Fig. 7) are generally small and not statistically significant. Three of the four pairs with p-values below 0.05 cannot safely be regarded as statistically significant under any scheme for correcting multiple comparisons. The remaining case hints at a positive correlation between the SINE-WAVE and TONE-VOCODED conditions.

V. DISCUSSION

A. Speed and degree of adaptation

When faced with previously unseen forms of speech, listeners' word identification performance improved with increasing exposure, confirming earlier reports of adaptation to specific forms of distortion, viz., speech filtered into a narrow spectral band (Warren *et al.*, 1995), fast speech (Dupoux and Green, 1997), noise-vocoded speech (Davis *et al.*, 2005), sine-wave speech (Bent *et al.*, 2011), and tone-vocoded speech (Hervais-Adelman *et al.*, 2011). Additionally, we examined distortions where the presence of exposure has not been tested previously, finding adaptation to locally time-reversed and sculpted speech.

However, the principal contribution of the current study is the finding that adaptation can be both very rapid and substantial in degree. The speed of adaptation is demonstrated by the finding that, overall, listeners required just 2.7 sentences, equivalent to 6.6 s of speech exposure, to achieve half the total gain observed across each 30-sentence block (parameter t_{50} in Table II). Concerning the amount of adaptation across a block, after being exposed to not much more than 1 min of speech, correct keyword identifications rose from 1 in 3 in the first sentence to 2 out of 3 by the 30th sentence.

Figure 8 illustrates why estimation of the speed and degree of adaptation depends critically on the number of sentences used as the size of the analysis unit when the time pattern of adaptation is nonlinear. The lower curve shows per-sentence scores (redrawn from Fig. 3) while the other curves (offset for clarity) show scores derived by using increasingly larger numbers of sentences in the analysis unit (e.g., 15 groups of 2 sentences, 10 groups of 3 sentences, etc). The presence of rapid adaptation is harder to gauge as this number increases above 5; moreover, the total amount of adaptation (indicated by the "gain" column) across the 30-sentence block is increasingly underestimated as the number of sentences used for the analysis increases.

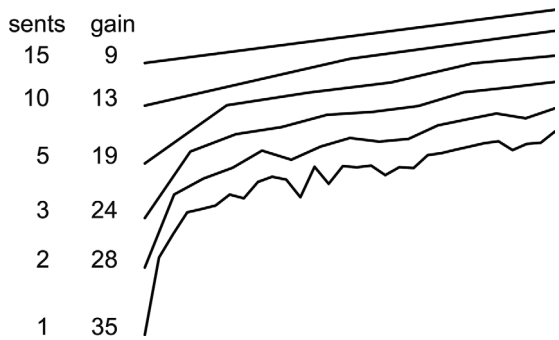

FIG. 8. The effect of the number of sentences used as the analysis unit ("sents") on apparent adaptation.

Our findings are consistent with the preliminary report of Van Hedger *et al*. (2019) for the common conditions of the two studies (sine-wave, tone-vocoded, and fast speech), although the correlations between distortion types were somewhat weaker in the present study. Whereas Van Hedger *et al*. (2019) do not explicitly refer to the very rapid initial adaptation in their data, an initial jump in performance from the first sentence to subsequent sentences is evident in their sinewave, fast, and babble speech conditions.

B. Role of prior exposure to distorted speech

We found no evidence of a generalised benefit of prior exposure to other forms of distortion. Although few studies have tested multiple forms of distorted speech with the same cohort, this outcome is in line with Casaponsa *et al*. (2019), who reported that exposure to amplitude modulated tones did not improve subsequent performance on tone-vocoded speech, and in partial agreement with Hervais-Adelman *et al*. (2011), who found incomplete transfer of perceptual learning from noise- to tone-vocoded speech (or vice versa). Our finding that adaptation was observed irrespective of where in the task listeners heard the block strengthens the notion that adaptation is not a consequence of procedural learning, which would be expected to diminish the adaptation effects in later blocks.

In this study, however, prior exposure to other distortions did help in the specific case of sine-wave speech. When first encountering this form of distortion, many listeners are unable to treat it as speechlike (Remez *et al*., 1981). Here, several listeners recognised no more than a handful of words in the entire block (Fig. 2). For those listeners who heard sine-wave sentences in a later block, the prior presence of other types of distortion is likely to have predisposed listeners to treat the "odd-sounding" signals that they encountered in subsequent blocks as speechlike.

C. Effect of distortion type

The time course of adaptation differs across distortion types. Whereas some of the variation may be due to the amount of acoustic-phonetic information that survives distortion, leading to the expectation of less adaptation in the "easier" conditions, there is evidence that the pattern of improvement is not strongly dependent on the intrinsic difficulty of the condition. This assertion is evidenced by the finding that six of the eight conditions produce final intelligibility levels in a fairly narrow range (51%–66%), indicating a similar level of intrinsic difficulty post-adaptation, yet, these conditions have initial intelligibilities covering a much wider range from just 2% to 53% (all values taken from logarithmic fits). Thus, the speed of adaptation does not depend solely on the degree to which the distorted stimuli are intrinsically capable of supporting speech perception following adaptation.

It is not yet clear which stimulus properties control the speed of adaptation. The only potential relationship suggested by correlation of gains during adaptation is between the TONE-VOCODED and SINE-WAVE conditions, although given the number of comparisons, this finding needs to be replicated before it can be considered robust. However, it is possible to speculate whether some shared characteristic may be responsible for their slower adaptation rate. In contrast to the other forms of distortion whose formant structure is broad and more representative of natural speech, TONE-VOCODED and SINE-WAVE are based on tonal carriers. Listeners typically use nonspeech terms such as "whistles" or "birdsong" to describe sine-wave speech on first hearing it (Remez *et al*., 1981). Simply broadening the formants by amplitude comodulation is sufficient for listeners to hear sine-wave sentences as more naturally speechlike (Carrell and Opie, 1992); indeed, listeners continue to report that such stimuli are "very much like speech" even when they are rendered unintelligible through the use of spectral and amplitude modulation taken from two different sentences (Rosen *et al*., 2011).

D. Potential adaptation mechanisms

Various proposals have been put forward to explain how listeners might adapt to distorted forms of speech. Broadly, these mechanisms fall into three categories: transformation, mapping, and selection. Here, we consider the degree to which such mechanisms can account for the range of adaptation responses observed here and, in particular, to the issue of how very rapid adaptation is achieved.

Transformational mechanisms attempt to undo the effect of distortion by a process of compensation or inverse transformation. Compensatory mechanisms have been proposed mainly in the context of adaptation to durational changes (e.g., Miller and Liberman, 1979) and room acoustics (e.g., Watkins, 2005). In the latter case, adaptation is indeed rapid with improvements taking place within a few sentences (Brandewie and Zahorik, 2010; Srinivasan and Zahorik, 2013). Changes in both room acoustics and speech rate are a familiar part of our listening experience. What is less clear is whether compensatory transformations operate in the face of novel forms of speech. Azadpour and Balaban (2015) trained listeners to recognise a particularly challenging form of distortion, spectrally rotated speech, but found no evidence to support compensation via an inverse transformation. For some of the distortions tested in the current study, in particular, for those in which substantial portions of the spectrum are removed, there is no obvious inverse transformation or compensatory process.

In contrast, explanations based on mapping involve learning the association between distorted and more typical forms of speech. While, in principle, this form of mapping could occur at lower levels (for instance, mapping sine-wave to natural formants), most studies have involved perceptual learning of phonological representations. In a typical perceptual learning paradigm, listeners are presented with multiple examples of one or more deviant sounds in a lexical context during an exposure phase. Learning has been demonstrated after exposure to as few as ten clear examples

(Poellmann *et al*., 2011). Applying this rate of adaptation to the scenario of this study where the entire sentence is "deviant" is not straightforward. Here, listeners were able to identify only 1.7 keywords per sentence at the outset. Under the assumption that lexically guided perceptual learning requires individual words to be correctly perceived, it is hard to see how this form of learning could explain the rapid initial jump in performance that we observe, although the more gradual phase of improvement could conceivably involve lexically guided perceptual learning.

Adaptation based on selection involves identifying those features that are least degraded by distorted speech and subsequently selecting or reweighting these cues. Azadpour and Balaban (2015) favoured a cue-weighting explanation in their study of spectrally rotated speech. In contrast, Green *et al*. (2013), also using spectrally rotated speech, argued that improved intelligibility involved adaptation to altered acoustical properties (in their case, spectral shape and dynamics) and found no evidence for weighting information that was relatively unaffected by the distortion (viz., intonational contrasts).

It is plausible that individual mechanisms of the types described above could play a part in explaining adaptation to specific forms of disortion in the current study. For instance, time-compressed speech might make use of compensatory mechanisms acquired during exposure to naturally fast speech, and glimpsed/sculpted forms of distortion could conceivably be identified via selective use of relatively undistorted time-frequency regions. However, this study shows that the listeners are capable of extracting meaning from deviant forms of speech that collectively show extreme diversity, raising the question of whether there is a unified adaptation mechanism that can account for distorted speech in general. The issue of single or multiple mechanisms is discussed by Sohoglu and Davis (2016), who distinguish very fast ("immediate") adaptation based on prior knowledge and more incremental learning, and show, using computational simulations, that a single mechanism based on minimizing prediction error can account for both time scales of adaptation. In their case, rapid learning was bootstrapped by providing matching written text immediately prior to degraded speech, giving listeners a reliable target for training. However, here, no such feedback was provided at any point, yet, the listeners were still capable of very rapid improvements.

VI. CONCLUSIONS

Listeners are capable of adapting to multiple forms of distorted sentences in the absence of feedback. Adaptation is rapid, with half of the eventual gain in intelligibility taking place over the first few sentences, followed by a more gradual improvement. The mechanism or mechanisms that enable adaptation in the face of a variety of distorted speech forms are presently unclear. Although listeners are unlikely to encounter speech degraded by artificial distortion in everyday scenarios, their ability to extract useful information from severely modified forms of speech highlights a perceptual flexibility that may well be important in handling natural forms of variability in communicative settings.

ACKNOWLEDGMENTS

The authors thank Dr. Stephen Hedger and Dr. Matt Davis for useful discussions. This research was supported by the University of the Basque Country Research Grant PES LISTA No. 15/06 to M.C. and the Deutsche Forschungsgemeinschaft (Cluster of Excellence 1077/1 Hearing4all) to B.T.M.

[1]See https://archive.org/details/100ClassicalMusicMasterpieces (Last viewed April 6, 2022).
[2]See https://doi.org/10.5281/zenodo.5849725 (Last viewed April 6, 2022).

Ahmadi, M., Gross, V. L., and Sinex, D. G. (**2013**). "Perceptual learning for speech in noise after application of binary time-frequency masks," J. Acoust. Soc. Am. **133**, 1687–1692.
Aubanel, V., García Lecumberri, M. L., and Cooke, M. (**2014**). "The Sharvard Corpus: A phonemically-balanced Spanish sentence resource for audiology," Int. J. Audiol. **53**, 633–638.
Azadpour, M., and Balaban, E. (**2015**). "A proposed mechanism for rapid adaptation to spectrally distorted speech," J. Acoust. Soc. Am. **138**, 44–57.
Banai, K., and Lavner, Y. (**2014**). "The effects of training length on the perceptual learning of time-compressed speech and its generalization," J. Acoust. Soc. Am. **136**, 1908–1917.
Bates, D., Mächler, M., Bolker, B., and Walker, S. (**2015**). "Fitting linear mixed-effects models using lme4," J. Stat. Soft. **67**(1), 1–48.
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**, 2660–2669.
Bent, T., Loebach, J. L., Phillips, L., and Pisoni, D. B. (**2011**). "Perceptual adaptation to sinewave-vocoded speech across languages," J. Exp. Psychol. Hum. Percept. Perform. **37**, 1607–1616.
Bieber, R. E., and Gordon-Salant, S. (**2021**). "Improving older adults' understanding of challenging speech: Auditory training, rapid adaptation and perceptual learning," Hear. Res. **402**, 108054.
Boersma, P. (**2001**). "Praat, a system for doing phonetics by computer," Glot Internat. **5**, 341–435.
Bradlow, A. R., and Bent, T. (**2008**). "Perceptual adaptation to non-native speech," Cognition **106**, 707–729.
Brandewie, E., and Zahorik, P. (**2010**). "Prior listening in rooms improves speech intelligibility," J. Acoust. Soc. Am. **128**, 291–299.
Carrell, T. D., and Opie, J. M. (**1992**). "The effect of amplitude comodulation on auditory object formation in sentence perception," Percep. Psychophys. **52**, 437–445.
Casaponsa, A., Sohoglu, E., Moore, D. R., Füllgrabe, C., Molloy, K., and Amitay, S. (**2019**). "Does training with amplitude modulated tones affect tone-vocoded speech perception?," PLoS One **14**, e0226288.
Clarke, C. M., and Garrett, M. F. (**2004**). "Rapid adaptation to foreign-accented English," J. Acoust. Soc. Am. **116**, 3647–3658.
Cooke, M., and García Lecumberri, M. L. (**2020**). "Sculpting speech from noise, music, and other sources," J. Acoust. Soc. Am. **148**, EL20–EL26.
Davis, M. H., Johnsrude, I. S., Hervias-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. Psych: General **134**, 222–241.
Dorman, M. F., Loizou, P. C., and Rainey, D. (**1997**). "Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs," J. Acoust. Soc. Am. **102**, 2403–2411.
Dupoux, E., and Green, K. (**1997**). "Perceptual adjustment to highly compressed speech: Effects of talker and rate changes," J. Exp. Psych. Human Percept. Perform. **23**, 914–927.

Ellis, D. (**2022**). "A phase vocoder in MATLAB," [software], available at https://www.ee.columbia.edu/~dpwe/resources/matlab/pvoc (Last viewed April 6, 2022).
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**, 2154–2164.
Erb, J., Henry, M. J., Eisner, F., and Obleser, J. (**2013**). "The brain dynamics of rapid perceptual adaptation to adverse listening conditions," J. Neurosci. **33**, 10688–10697.
García Lecumberri, M. L., Cooke, M., and Bryant, C. (**2015**). "Accent evaluation from extemporaneous child speech," Poznan Stud. Contemp. Linguist. **51**, 227–246.
Green, T., Rosen, S., Faulkner, A., and Paterson, R. (**2013**). "Adaptation to spectrally-rotated speech," J. Acoust. Soc. Am. **134**, 1369–1377.
Hervais-Adelman, A. G., Davis, M. H., Johnsrude, I. S., Taylor, K. J., and Carlyon, R. P. (**2011**). "Generalization of perceptual learning of vocoded speech," J. Exp. Psychol. Hum. Percept. Perform. **37**, 283–295.
Jekel, C. F., and Venter, G. (**2019**). "pwlf: A Python Library for fitting 1D continuous piecewise linear functions," available at https://github.com/cjekel/piecewise_linear_fit_py (Last viewed April 6, 2022).
Kakehi, K. (**1992**). "Adaptability to differences between talkers in Japanese monosyllabic perception," in *Speech Perception, Speech Production and Linguistic Structure*, edited by Y. Tohkura, E. Vatikiotis-Bateson, and Y. Sagisaka (OHM, Tokyo), pp. 135–142.
Kato, K., and Kakehi, K. (**1988**). "Listener adaptability to individual speaker differences in monosyllabic speech perception," J. Acoust. Soc. Jpn. **44**, 180–186.
Ladefoged, P., and Broadbent, D. E. (**1957**). "Information conveyed by vowels," J. Acoust. Soc. Am. **29**, 98–104.
Lehet, M. I., Fenn, K. M., and Nusbaum, H. C. (**2020**). "Shaping perceptual learning of synthetic speech through feedback," Psychon. Bull. Rev. **27**, 1043–1051.
Melguy, Y. V., and Johnson, K. (**2021**). "General adaptation to accented English: Speech intelligibility unaffected by perceived source of non-native accent," J. Acoust. Soc. Am. **149**, 2602–2614.
Miller, J. L., and Liberman, A. M. (**1979**). "Some effects of later-occurring information on the perception of stop consonant and semivowel," Percept. Psychophys. **25**, 457–465.
Norris, D., McQueen, J. M., and Cutler, A. (**2003**). "Perceptual learning in speech," Cognit. Psychol. **47**, 204–238.
Ortiz, J. A., and Wright, B. A. (**2009**). "Contributions of procedure and stimulus learning to early, rapid perceptual improvements," J. Exp. Psychol. Hum. Percept. Perform. **35**, 188–194.
Poellmann, K., McQueen, J. M., and Mitterer, H. (**2011**). "The time course of perceptual learning," in *17th International Congress of Phonetic Science*, Hong Kong, pp. 1618–1621.
R Core Team. (**2021**). "R: A language and environment for statistical computing" (*R* Foundation for Statistical Computing, Vienna, Austria), available at https://www.R-project.org/ (Last viewed April 6, 2022).
Remez, R., Rubin, P. E., Pisoni, D. B., and Carrell, T. D. (**1981**). "Speech perception without traditional speech cues," Science **212**, 947–949.
Robinson, K., and Summerfield, A. Q. (**2006**). "Adult auditory learning and training," Ear Hear. **17**, 51–65.
Rosen, S., Faulkner, A., and Wilkinson, L. (**1999**). "Adaptation by normal listeners to upward spectral shifts of speech: Implications for cochlear implants," J. Acoust. Soc. Am. **106**, 3629–3636.
Rosen, S., Wise, R. J. S., Chadha, S., Conway, E.-J., and Scott, S. K. (**2011**). "Hemispheric asymmetries in speech perception: Sense, nonsense and modulations," PLoS One **6**, e24672.
Rothauser, E. H., Chapman, W. D., Guttman, N., Silbiger, H. R., Hecker, M. H. L., Urbanek, G. E., Nordby, K. S., and Weinstock, M. (**1969**). "IEEE Recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoust. **17**, 225–246.
Rotman, T., Lavie, L., and Banai, K. (**2020**). "Rapid perceptual learning: A potential source of individual differences in speech perception under adverse conditions?," Trends Hear. **24**, 2331216520930541.
Saberi, K., and Perrott, D. R. (**1999**). "Cognitive restoration of reversed speech," Nature **398**, 760.
Samuel, A. G., and Kraljic, T. (**2009**). "Perceptual learning for speech," Attent., Percept. Psychophys. **71**, 1207–1218.
Simantiraki, O., and Cooke, M. (**2021**). "SpeechAdjuster: A tool for investigating listener preferences and speech intelligibility," in *Proceedings of Interspeech 2021*, pp. 1718–1722.
Sohoglu, E., and Davis, M. H. (**2016**). "Perceptual learning of degraded speech by minimizing prediction error," Proc. Natl. Acad. Sci. U.S.A. **113**(12), E1747–E1756.
Srinivasan, N. K., and Zahorik, P. (**2013**). "Prior listening exposure to a reverberant room improves open-set intelligibility of high-variability sentences," J. Acoust. Soc. Am. **133**, EL33–EL39.
Van Hedger, S. C., Heald, S. L. M., Nusbaum, H. C., Batterink, L. J., Davis, M. H., and Johnsrude, I. S. (**2019**). "Learning different forms of degraded speech as a cognitive skill," in *Annual Meeting of the Psychonomic Society*, Montreal, Canada.
Warren, R. M., Riener, K. R., Bashford, J. A., and Brubaker, B. S. (**1995**). "Spectral redundancy: Intelligibility of sentences heard through narrow spectral slits," Percept. Psychophys. **57**, 175–182.
Watkins, A. J. (**2005**). "Perceptual compensation for effects of reverberation in speech identification," J. Acoust. Soc. Am. **118**, 249–262.
Zhang, P., Zhao, Y., Dosher, B. A., and Lu, Z. L. (**2019**). "Assessing the detailed time course of perceptual sensitivity change in perceptual learning," J. Vision **19**, 9.