Hearing Research 315 (2014) 80–87

Contents lists available at ScienceDirect

Hearing Research

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

Research paper

T'ain't the way you say it, it's what you say – Perceptual continuity of voice and top–down restoration of speech☆

Jeanne Clarke [a, b, *], Etienne Gaudrain [a, b], Monita Chatterjee [c], Deniz Başkent [a, b]

[a] University of Groningen, University Medical Center Groningen, Department of Otorhinolaryngology/Head and Neck Surgery, Groningen, The Netherlands
[b] University of Groningen, Graduate School of Medical Sciences, Research School of Behavioral and Cognitive Neurosciences, Groningen, The Netherlands
[c] Boys Town National Research Hospital, Omaha, NE 68131, USA

ARTICLE INFO

Article history:
Received 13 December 2013
Received in revised form
25 June 2014
Accepted 2 July 2014
Available online 11 July 2014

ABSTRACT

Phonemic restoration, or top–down repair of speech, is the ability of the brain to perceptually reconstruct missing speech sounds, using remaining speech features, linguistic knowledge and context. This usually occurs in conditions where the interrupted speech is perceived as continuous. The main goal of this study was to investigate whether voice continuity was necessary for phonemic restoration. Restoration benefit was measured by the improvement in intelligibility of meaningful sentences interrupted with periodic silent gaps, after the gaps were filled with noise bursts. A discontinuity was induced on the voice characteristics. The fundamental frequency, the vocal tract length, or both of the original vocal characteristics were changed using STRAIGHT to make a talker sound like a different talker from one speech segment to another. Voice discontinuity reduced the global intelligibility of interrupted sentences, confirming the importance of vocal cues for perceptually constructing a speech stream. However, phonemic restoration benefit persisted through all conditions despite the weaker voice continuity. This finding suggests that participants may have relied more on other cues, such as pitch contours or perhaps even linguistic context, when the vocal continuity was disrupted.

1. Introduction

In real-life communication, while speech often happens in the presence of background masking noise, people are most of the time still able to understand the message intended by the speaker. Perhaps contributing to this (Warren, 1983), under certain circumstances, the brain has the ability to restore missing speech segments. This phenomenon is referred to as *perceptual* or *phonemic restoration* (Warren, 1970).

The phonemic restoration effect can be quantified by measuring the increase in intelligibility of sentences with periodic silent intervals after these intervals are filled with noise bursts (Powers and Wilcox, 1977; Verschuure and Brocaar, 1983). Phonemic restoration was described as a "two-stage process of perceptual synthesis" (Bashford et al., 1992; Bregman, 1990) consisting of: (i) the perceived continuity of speech (described as "continuity illusion" in this context) with simple auditory induction, and (ii) the repair mechanisms of the missing sounds with knowledge-driven processes. First, the interrupted speech is illusorily perceived as continuous when the filler noise acts as a plausible masker for the missing segments of speech and if there is no perceptual evidence against continuity (Miller and Licklider, 1950; Warren, 1970). Second, intelligibility increases with repair mechanisms of top–down restoration, using linguistic knowledge and context (Bashford et al., 1992; Wang and Humes, 2010; Warren and Sherman, 1974). While previous studies showed better restoration in conditions where the perceived continuity of noise-interrupted sentences was stronger, thus indicating a close connection between the two stages

Abbreviations: CI, Cochlear implant; dB HL, decibel hearing loss; dB SPL, decibel sound pressure level; D/A, digital/analogous; F, resynthesized voice with change in Fundamental frequency; F0, Fundamental frequency; FDR, false discovery rate; FV, resynthesized voice with change in Fundamental frequency and Vocal tract length; IR, interruption rate; PR, phonemic restoration; RAU, rationalized arcsine transformed unit; RM ANOVA, repeated measure analysis of variance; RMS, root mean square; s.d., standard deviation; S/PDIF, Sony/Philips Digital Interface Format; SER, Spectral Envelope Ratio; SNR, signal to noise ratio; SU, (re)Synthesized voice with Unmodified vocal characteristics; V, resynthesized voice with change in Vocal tract length; VTL, Vocal tract length

☆ Portions of this work were presented in "Phonemic Restoration: Studying the Effect of Voice Alternation" ARO MidWinter Meeting, Baltimore, Maryland, February 2013.

* Corresponding author. University of Groningen, University Medical Center Groningen, Department of Otorhinolaryngology/Head and Neck Surgery, PO Box 30.001 9700RB, Groningen, The Netherlands. Tel.: +31 50 3611315.

E-mail addresses: j.n.clarke@umcg.nl (J. Clarke), e.p.c.gaudrain@umcg.nl (E. Gaudrain), monita.chatterjee@boystown.org (M. Chatterjee), d.baskent@umcg.nl (D. Başkent).

http://dx.doi.org/10.1016/j.heares.2014.07.002
0378-5955/

(Bashford et al., 1992; Başkent et al., 2009), evidence from imaging experiments showed that the continuity illusion and repair mechanisms are two separate neural mechanisms that seemingly interact (Shahin et al., 2009). Consequently, the extent to which continuity and repair mechanisms are linked is not yet clear.

The fact that the term "continuity illusion" refers to different, albeit similar and likely related, phenomena and paradigms across the literature, may have contributed to this lack of clarity. Continuity illusion, as described by Bregman (1990) in the context of phonemic restoration and auditory scene analysis, is the perception of an interrupted target sound as a single object as if uninterrupted behind a louder masking noise. One of the four prerequisites of continuity illusion is the grouping rule (Bregman, 1990, pp. 345–394). In other words, for the continuity illusion to happen, the successive segments of the target must be grouped into a single, coherent, auditory stream. This sequential grouping between each target's segments strongly depends on their similarity in their spectral content, fundamental frequency, and location in space (Hartmann and Johnson, 1991). Consequently, if these acoustic cues are changing significantly from one segment to the next, the successive segments are less likely to be integrated into a single stream, thereby weakening or removing the continuity illusion effect.

The phenomenon described in this definition likely contributes to the phenomenon of perceived continuity in general. It is this general concept of perceived continuity that we investigated here. More specifically, in the present study, we modified the acoustic cues from a male voice into a female voice to induce the perception of two different talkers. Alternating between these two voices in a sentence would break the continuity of the vocal characteristics. We hypothesized that the disrupted voice continuity would hinder the perception of the speech segments as a single stream.

The goal of the present study was to investigate whether voice continuity is necessary for phonemic restoration. If this is the case, we hypothesized that breaking the voice continuity of the speech stream would prevent, or at least reduce, the phonemic restoration benefit. The voice continuity of interrupted speech with filler noise was disrupted with manipulations that were applied at the indexical[1] level. This way, the linguistic content (as this is an important factor for the repair mechanisms) was left intact, while acoustic cues important for perceptual organization in general, and sequential grouping of speech specifically, were manipulated. A two-talker percept was created from the interrupted speech by alternating between two voices on each speech segment. The vocal characteristics we manipulated were the fundamental frequency (F0) and the vocal tract length (VTL) as these are the most important for gender identification (Skuk and Schweinberger, 2013) and can be used for speaker identity manipulation (Gaudrain et al., 2009). The F0 is related to the pitch of the voice, and the VTL to the size of the speaker (Fitch and Giedd, 1999). These give information about the size and the sex of a speaker (Hillenbrand and Clark, 2009; Smith et al., 2007; Titze, 1989), and can also play an important role for the intelligibility of speech in adverse listening scenarios (Darwin et al., 2003; Mackersie et al., 2011). Furthermore, continuity of these vocal characteristics influences speech recognition performance (Best et al., 2008; Kidd et al., 2008; Maddox and Shinn-Cunningham, 2012; Shinn-Cunningham et al., 2013), suggesting that F0 and VTL are used to perceptually construct a speech stream (Gaudrain et al., 2007; Tsuzaki et al., 2007) by linking successive segments of speech over time. Hence, grouping successive segments of speech with different vocal characteristics should be more difficult in comparison with grouping speech segments from the same voice, and if the grouping rule of the continuity percept is a prerequisite for the repair mechanisms of missing speech segments, the voice manipulations that cause a disruption at the indexical level should reduce phonemic restoration benefit. We also manipulated F0 and VTL separately to systematically investigate the importance of each parameter independently on voice continuity and on phonemic restoration. Because F0 varies substantially within the same speaker in natural speech, whereas VTL does not, the effect of breaking the continuity could be different for the two cues.

In this study, three experiments were conducted. In experiment 1, the voice manipulations were assessed to confirm that the target female voice was indeed perceived as a different talker than the original male voice. In experiment 2, the effect of the voice manipulation on perceived continuity was assessed to confirm when the voice continuity was perceived as broken by the voice alternations. In experiment 3, the main experiment of the study, the effect of voice manipulation on phonemic restoration was investigated.

2. General methods

This section describes methods that were common to all three experiments. Note that in order to keep the participants as naïve as possible to both speech stimuli and the experimental paradigm during the main experiment, experiment 3 was run first. The voice assessment experiment, experiment 1, was run after the phonemic restoration experiment. Continuity assessment, experiment 2, was run in another session with different participants.

2.1. Stimuli

Meaningful Dutch sentences, spoken by a male talker and digitized at a 44.1 kHz sampling rate, were used (from Versfeld et al., 2000). Each sentence was grammatically and syntactically correct and contained between four and nine words. The words were no longer than three syllables and had an average duration of 325 ms (s.d. 45 ms). The corpus was divided into 39 homogeneous lists of 13 sentences, where the lists of sentences were equally intelligible. Two lists were excluded: list #39 because its distribution of phonemes did not match the average frequency of phonemes in Dutch (Versfeld et al., 2000); and list #13 because it contained a sentence also present in list #21.

2.2. Signal processing

We manipulated the talker's voice using two independent parameters, the F0 and the VTL, offline, using the STRAIGHT software (v40.006b) implemented in Matlab (Kawahara et al., 1999). The speech signal was first decomposed into a spectral fine structure reflecting the F0 contour, and a spectral envelope at each time sample. The F0 was then manipulated by multiplying all values of the F0 contour by a factor, thus changing the average F0 but preserving the relative fluctuations around the average. The VTL was manipulated by expanding the extracted spectral envelope towards the high frequencies, which produced shorter VTLs. The two modified parts of the sound were then recombined using a pitch synchronous overlap-add resynthesis method. Note that all stimuli were resynthesized with STRAIGHT, even when the F0 and the VTL were both left unchanged (the baseline male voice condition), to control for any perceptual effects of resynthesis.

To ensure discontinuity of the vocal characteristics, speech segments were designed to alternate between voices of a man and a

[1] Indexical cues refer to the voice characteristics specific to a talker (e.g. Helfer and Freyman, 2009; Mclennan and Luce, 2005), in opposition with the lexical (or linguistic) cues, which can be learnt and depend on the language.

woman. In studies investigating the effect of F0 and envelope shifting on gender identification, it has been shown that when both F0 and formants were shifted up from male toward typical values of female, or down from female toward typical values of male voices, the speaker was identified as from the opposite sex (thus a different speaker; Fuller et al., in revision; Hillenbrand and Clark, 2009; Skuk and Schweinberger, 2013). In order to change the original voice of the male into that of a female, the F0 was multiplied by two and the spectral envelope was expanded by a ratio of 1.26 (i.e., all formant frequencies were shifted upwards by a third octave). The calculation of the length of the vocal tract of the original male talker was based on work by Fitch and Giedd (1999). We estimated a VTL of 15.4 cm for the male talker (corresponding to the VTL for a Dutch man of average height = 180 cm), and used this as a reference (spectral envelope ratio = 1), as was done by Ives et al. (2005). This resulted in an apparent VTL of 12.2 cm for the female voice. The sentences used in experiment 1 were processed under 16 voice conditions, with different F0 and/or VTL modification ratios. The sentences used in experiments 2 and 3 were processed under four voice conditions: (SU) (re)synthesis with unmodified vocal characteristics, (F) one octave F0 shift only, (V) shorter VTL only, and (FV) both F0 and VTL modified.

The resynthesized sentences were interrupted using square wave modulation with an interruption rate (IR) of 2.2 Hz and a 50% duty cycle, producing speech and silent/noise intervals of 227 ms duration (which is close to the average syllabic rate of the speaker in this corpus). The IR was chosen in a pilot study, which showed that from IRs of 1.5, 2.2, 3 and 5 Hz, 2.2 Hz produced the most robust phonemic restoration effect, with levels of speech intelligibility far from floor or ceiling. A 5-ms raised-cosine ramp was applied to the onsets and offsets of the square wave to smooth the alternations between speech segments and interruptions and to reduce spectral splatter. Vocal characteristics (F0 and VTL) of successive speech segments alternated from original values to other values (see Fig. 1b). The interruptions were either left silent or filled with speech-shaped noise (signal to noise ratio [SNR] of −5 dB). For each voice condition, all resynthesized sentences from the SU voice and the alternate voice were concatenated for the computation of the long-term spectrum. A single speech-shaped noise file was generated from white noise modulated by the long-term average spectrum in each voice condition.

2.3. Apparatus

The processed digital stimuli were sent through the S/PDIF output of an AudioFire 4 soundcard (Echo Digital Audio Corporation). After conversion to an analog signal via a DA10 D/A converter (Lavry Engineering Inc.), the stimuli were played back diotically through HD600 headphones (Sennheiser Electronic Corporation). The speech segments for all voice conditions were set to a RMS level of 65 dB SPL. The calibration of the stimuli was performed on the first 20 sentences from the corpus with a Sound & Vibration Analyser (Svan 979 from Svantek) connected to a Kemar head (G.R.A.S.). The participants were seated in a sound-attenuated booth facing a computer monitor. Their verbal response was recorded on a PalmTrack digital voice recorder (ALESIS).

2.4. Procedure

Participants came for a single session, that included providing the instructions, obtaining written informed consents, conducting the audiometric test, the training, the experiment(s), the debriefing and occasional breaks. A single session for experiments 1 and 3 lasted 1.5–2 h. A session for experiment 2 lasted 30 min.

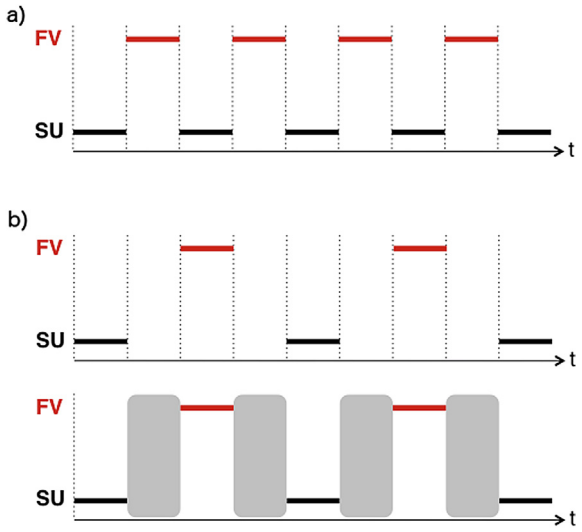

Fig. 1. a) Schematic of a stimulus for the SU-FV condition used in experiments 1, showing alternating voices (SU voice in black and FV voice in red) in sentence segments without interruptions. The first segment was always from the SU voice, which was the resynthesized unprocessed voice. The alternating voice was the same throughout one sentence, and was one of the 16 resynthesized voices. b) Schematic of stimuli for the SU-FV condition in experiment 2 and 3, showing the alternating voices (SU voice in black and FV voice in red) in successive sentence segments with silent interruptions (upper panel) and filler noise bursts (lower panel). The first segment was always from the SU voice, and the alternating voice could be SU (SU-SU voice condition), F (SU-F voice condition), V (SU-V voice condition) or FV (SU-FV voice condition). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3. Experiment 1: voice assessment

3.1. Material and methods

3.1.1. Participants

Sixteen normal-hearing native Dutch speakers, with no history of hearing problems and aged 20–29 years (mean = 23.3, s.d. = 2.7), participated in the study. Their pure-tone thresholds were 20 dB HL or less at audiometric frequencies between 250 and 6000 Hz in both ears. The study was approved by the Medical Ethical Review Committee (Medisch Etische Toetsingscommissie) of the University Medical Center Groningen, and written informed consent was collected from each participant. An hourly fee was paid.

3.1.2. Procedure

This voice assessment experiment was conducted to confirm that the chosen voice manipulation did lead to the perception of different speakers. The voice was modified in 16 steps between the original male voice and the target female voice as shown in Fig. 2. Alternations between the original parameters and the other voice conditions (see Fig. 1a) were applied by modulating the original sentence with the same square wave as described in the general method (see Section 2.2), and the sentence from the alternated voice with the inverse square wave. In this case, no speech segments were missing but they alternated between the original male voice (SU) and one of the 16 voice conditions described in Fig. 2. Seven sentences per condition were played to the participants whose task was to report whether they heard one talker or two different talkers for each sentence. The sentences were taken from lists the participants had not heard during experiment 3.

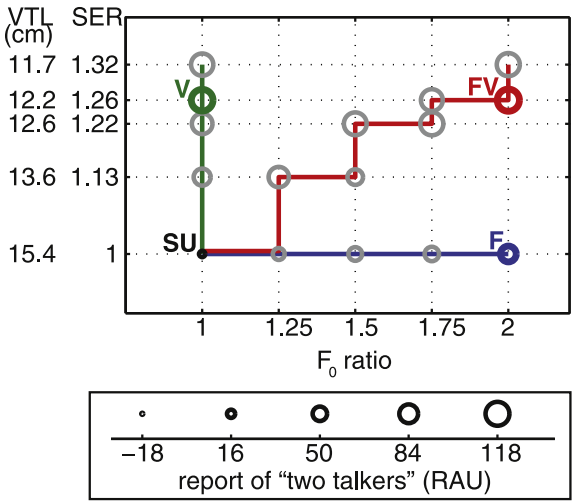

Fig. 2. Voice conditions represented in the F0-VTL plane. The *x*-axis shows the modification ratio of F0, with a ratio of 1 for the original male voice. The *y*-axis shows the VTL apparent size (in cm) corresponding to the spectral envelope ratio (SER), with a ratio of 1 for the original male voice. The four voice manipulations used for experiments 2 and 3 are represented by the colored circles. SU is the baseline resynthesis of the original male voice with unmodified vocal characteristics. F is the resynthesis of the male voice with F0 shifted up by an octave. V is the resynthesis of the male voice with a shorter VTL. FV is the resynthesis of the male voice with both parameters modified. The 16 voices resynthesized for experiment 1 are represented by the gray circles. The blue horizontal line shows the change in F0 only; the green vertical line shows the change in VTL only; and the red zigzag line shows the combined change in the two dimensions. The area of the circles for each voice condition represents the RAU scores for trials where the participants reported hearing two different talkers (results of experiment 1). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The average percent of report of two talkers were converted to RAU (rationalized arcsine transform units, Studebaker, 1985). RAU are used when the range of scores is finite, to temper homoscedasticity problems in order to fulfill one of ANOVA's homogeneity of variances assumptions.

3.2. Results

The RAU scores for reports of hearing two different talkers are shown in Fig. 2, as the area of the circle for each voice condition. A repeated-measure (RM) ANOVA with voice condition (16 levels) as factor showed a significant effect [$F(15,225) = 23.14$, $p < 0.001$]. Post hoc comparisons (with false discovery rate (FDR) correction) showed that the four conditions used in experiment 3 were significantly different from each other [$p < 0.01$ for SU-SU vs. SU-F, SU-F vs. SU-V, SU-F vs. SU-FV, and $p < 0.001$ for SU-SU vs. SU-V, SU-SU vs. SU-FV] except for SU-V vs. SU-VF [$p = 0.74$]. These results showed that (i) the original but resynthesized condition (SU-SU) produced a one-talker percept, as expected, (ii) manipulating F0 only (SU-F) did not produce a strong two-talker percept, with only 43 RAU rated as two talkers, even for a difference in F0 as big as one octave, (iii) manipulating VTL only (SU-V) or both F0 and VTL (SU-FV) produced a strong two-talker percept, with 94 RAU and 99 RAU, respectively, rated as two talkers.

To ensure that the voice resynthesis alone did not change speech intelligibility, three additional participants selected with the same inclusion criteria, were tested for intelligibility of uninterrupted sentences with the voice manipulations (SU, F, V and FV). All scores were at ceiling, confirming that the voice manipulations did not introduce artifacts or unnaturalness that significantly reduced speech intelligibility.

4. Experiment 2: continuity assessment

4.1. Methods

This experiment was conducted to confirm that voice continuity was really perceived as broken, by making sure that the male and the female voices could *not* be identified as a single talker (rather than testing if two different voices can be identified like in experiment 1). Sixteen participants aged 19–30 years (mean = 22.7, s.d. = 2.8), included upon the same criteria as experiment 1, but who only participated in this experiment, were asked to judge if they thought it possible that a single talker uttered the whole sentence. The participants were tested in the same conditions of experiment 3 (noise and silent interruptions, and 4 voice conditions: SU-SU, SU-F, SU-V, and SU-FV).

4.2. Results

The RAU scores for percentage of trials where participants reported hearing a single talker are shown in Fig. 3. The RM ANOVA with voice condition (4 levels) and interruption filler (2 levels) as factors showed a significant effect of voice condition on the judgment of single talker [$F(3,45) = 48.21$, $p < 0.001$], a significant effect of interruption filler on the judgment of single talker [$F(1,15) = 6.47$, $p = 0.023$], and no interaction between the two factors [$F(3,45) = 0.96$, $p = 0.42$]. Posthoc pairwise comparisons were computed with False Discovery Rate (FDR) control for multiple comparisons. First, the results showed that, SU-SU and SU-FV conditions were significantly different from each other, both with silent interruptions (104 RAU and −6 RAU, respectively, for reports of hearing a single talker) and with noise filler (109 RAU and −2

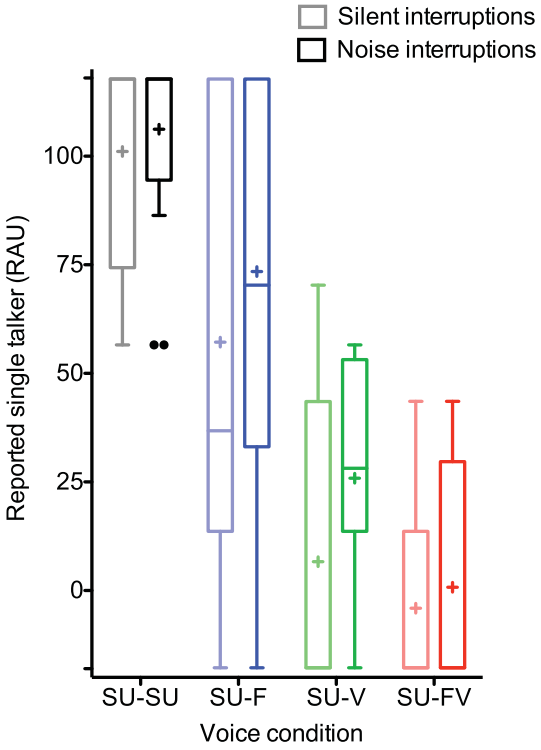

Fig. 3. Results of experiment 2. Boxplot of RAU scores for trials where the participants reported hearing a single talker, shown as a function of voice condition (SU-SU, SU-F, SU-V, and SU-FV), with silent gaps (light boxes) and with gaps filled with noise bursts (dark boxes). The bar indicates the median, the box indicates the 25th and 75th quartiles, and the whiskers indicate the 1.5 IRQ. The mean is displayed with a cross and dots indicate outliers.

RAU, respectively). Furthermore, for the SU-FV condition, the participants clearly judged the two voices as not possibly being from a single talker. This confirms that our manipulation of the male voice (SU) toward a female target voice (FV) indeed induced a two-talker percept. Thus, for this SU-FV condition, participants cannot rely on vocal characteristics of the talkers to integrate the speech segments into a single stream.

5. Experiment 3: phonemic restoration

5.1. Methods

5.1.1. Participants and stimuli

The same participants were involved in experiment 1 and 3. The stimuli from the same speech corpus were used for experiment 1 and 3, although participants never heard the same sentence twice.

5.1.2. Procedure

The participants listened to one stimulus at a time. A short beep preceded the stimulus to alert the participant. They were asked to verbally repeat what they could understand from each sentence, and were encouraged to guess as much as possible (Başkent, 2012). The spoken responses were recorded for offline scoring. A native Dutch speaking student assistant, who was unaware of the experimental conditions, listened to the recordings, and calculated the percent-correct scores as the ratio of correctly identified words to the total number of words presented to the listener. For familiarization with the procedure and the stimuli, training was provided before data collection. The first four lists of sentences were used for training, with four conditions taken randomly from the eight conditions used in the experiment. The task was similar to that for the main experiment, except that feedback was provided after each response by playing the full sentence in one of the resynthesized voices (original or manipulated), and by playing the interrupted sentence once more, as well as displaying its text on the screen. This form of training was previously shown to be effective with similarly interrupted sentence materials (Benard and Başkent, 2013). The main experiment consisted of 8 conditions [4 voice conditions (SU-SU, SU-F, SU-V, SU-FV) × 2 interruption conditions (silent intervals and with noise filler)]. The orders of the sentence lists and of the conditions were randomized.

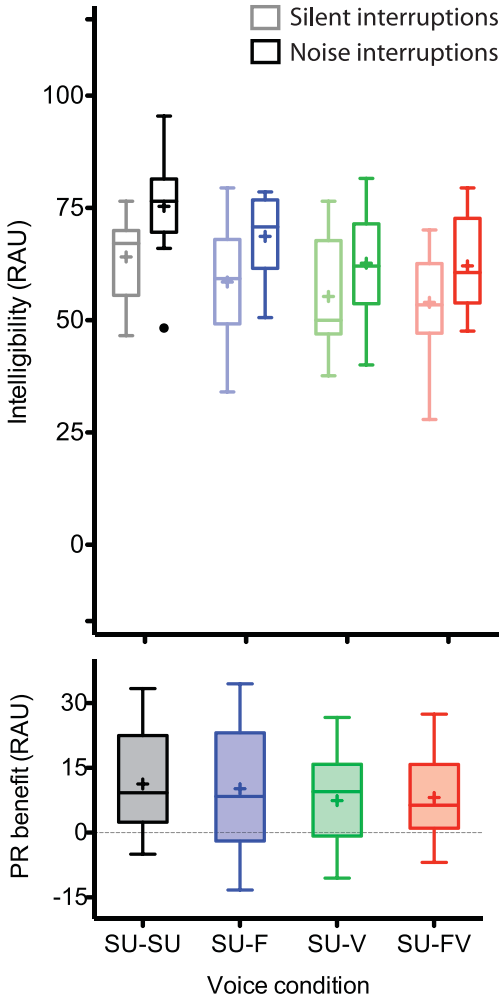

Fig. 4. Results of experiment 3. The upper panel displays the mean intelligibility scores in RAU for each voice condition with silent gaps (light boxes) and with gaps filled with noise bursts (dark boxes). For each voice condition, the difference of scores between the light and the dark boxes indicates the phonemic restoration effect, shown as filled boxplot in the lower panel.

5.2. Results

The intelligibility of the interrupted sentences is shown in Fig. 4 (upper panel) for the four voice conditions tested for sentences with silent gaps (light boxes) and with filler noise (dark boxes). The phonemic restoration effect is shown by better scores with filler noise (see also Fig. 4, lower panel). As the alternating voices became more different, there was a decrease in intelligibility without or with the filler noise. However, the phonemic restoration effect was present for all conditions. A RM two-way ANOVA on the RAU scores, with the within-subject factors of voice condition (four levels; SU-SU, SU-F, SU-V, and SU-FV) and of interruption condition (two levels; silence, noise) showed significant main effects of voice condition [$F(3,45) = 8.48$, $p < 0.001$], and of interruption condition [$F(1,15) = 29.49$, $p < 0.001$]. Critically, there was no significant interaction between the two factors [$F(3,45) = 0.40$, $p = 0.76$], suggesting that the voice manipulation did not affect the phonemic restoration effect. Post-hoc pairwise comparisons were computed with FDR correction for multiple comparisons. First, the intelligibility in the SU-F condition did not significantly differ from that of the SU-SU condition, both for silent interruptions and interruptions filled with noise. This indicates that the manipulation of F0 alone does not significantly disturb the listeners' intelligibility. Second, the results showed a significant decrease in intelligibility scores from the SU-SU to the SU-V conditions, both for silent and noise interruptions, and an equivalent decrease from the SU-SU to the SU-FV conditions. This indicates that the manipulation of VTL alone could be responsible for the decrease when both vocal parameters are modified. Finally, there was a significant phonemic restoration effect in all voice conditions.

The lower panel of Fig. 4 shows the phonemic restoration effect directly for each voice condition, calculated by subtracting the scores obtained without filler noise from those obtained with filler noise. The figure shows that, as suggested by the significant main effect of interrupted condition (silence vs. noise), in all voice conditions there was a significant phonemic restoration effect (between 7.41 RAU and 11.26 RAU). As the lack of interaction in the RM-ANOVA implied, the voice condition had no significant effect on the size of the phonemic restoration benefit. In addition, comparison of the phonemic restoration effect scores to zero confirms that the phonemic restoration benefit was present for all voice conditions [SU-SU: $t(15) = 3.79$, $p = 0.0018$; SU-F: $t(15) = 2.84$, $p = 0.012$; SU-V: $t(15) = 2.60$, $p = 0.020$; SU-FV: $t(15) = 3.38$, $p = 0.0041$]. In short, phonemic restoration was unaffected by the voice manipulations.

6. Discussion

The aim of this study was to investigate what role voice continuity plays in phonemic restoration. Continuity was manipulated at an indexical level, changing the vocal characteristics (fundamental frequency and vocal-tract length), leaving the linguistic content intact. Because voice characteristics play an important role in perceptual organization of speech, and particularly for grouping and segregation of speech streams, disrupting acoustical voice cues hinders the formation of speech streams and the linkage of speech segments over time. Across time linkage is required to perceive a sequence of speech segments as a single, continuous speech stream. We thus hypothesized that, if this continuity perception plays a major role in phonemic restoration, the top–down repair of interrupted speech would be reduced (the addition of noise would provide less intelligibility benefit) with disrupted voice cues because the sentences would be perceived as less continuous.

Unexpectedly, the phonemic restoration effect persisted for all voice conditions despite the fact that the voice difference used in the alternating voice patterns (SU-F, SU-V and SU-FV) was large enough for two distinct talkers to be heard (as shown in experiment 2), as well as to disrupt absolute intelligibility significantly. This finding does not support our hypothesis and seems to contradict the idea that the voice continuity is necessary for phonemic restoration. It might imply that the mechanisms involved in phonemic restoration are somewhat different from our initial supposition. In partial support of this idea, a recent study with sentences involving cochlear implant listeners (Bhargava et al., 2014), extending the study of Miller and Licklider (1950), showed that in some cases strong continuity illusion could be observed without a phonemic restoration effect and that in other cases better phonemic restoration benefit could be observed with lesser continuity illusion. To explain our persistent phonemic restoration effect, we propose that the participants were able to focus on the message, and that the high linguistic context of the sentences enabled participants to overcome the voice discontinuity to create a higher-level reconstructed representation (also supported by Billig et al., 2013; Warren and Sherman, 1974). This implies that, at this slow rate of interruptions, participants would rely on the linguistic context to achieve phonemic restoration and would not be disturbed by the inconsistent indexical cues.

Although disrupting the continuity of the vocal characteristics had no effect on the top–down repair of speech (i.e., the phonemic restoration effect did not disappear), the manipulation of the voice had an effect on the global intelligibility. This effect varied depending on the specific voice manipulation, implying that different voice cues may play different roles in understanding interrupted speech. Intelligibility decreased when the alternating voices became more different. This supports the significance of voice continuity for understanding interrupted speech, likely because voice cues are important for grouping speech segments (Darwin et al., 2003; Mackersie et al., 2011) and/or because adaptation to vocal characteristics can influence phonetic processing (e.g. Ladefoged and Broadbent, 1957). The decrease in intelligibility was independent of whether or not the gaps were filled with noise. The effect on absolute intelligibility could also be related to speech artifacts introduced with the resynthesis in STRAIGHT. However, previous studies have evaluated the good quality of voice manipulation (for vowels: Assmann and Katz, 2005; Liu and Kewley-Port, 2004). For the original voice and when F0 only was manipulated (SU-SU and SU-F conditions), we observed similar intelligibility (in experiment 3) while the voice manipulation led to broken perceived continuity half of the time (in experiment 2). This suggests that participants could adapt to sudden, large, changes in F0. However, the F0 manipulation only changed the average F0 value, leaving the F0 contours unchanged. It is possible that the listeners took advantage of intonation and word accentuation cues in the high context sentences. To investigate the importance of prosody and intonation for speech repair, resynthesized voices with manipulated F0 contours will be used in future research. The similarity of results when VTL only and when both F0 and VTL were manipulated (SU-V and SU-FV conditions in all three experiments) suggests that the VTL component alone explains the difference in intelligibility between the male-female voice alternations (SU-FV) and the control condition (SU-SU). Moreover, it suggests that participants could not adapt as well to sudden changes in VTL as to changes in F0. In short, this experiment showed that, for the present voice manipulations, the continuity of VTL was more important for intelligibility of interrupted speech than that of F0.

Even though phonemic restoration was possible when speech segments were perceived as different voices, other consequences may not be captured in the present study. For example, top–down restoration of interrupted speech with inconsistent voice cues could be more effortful. As intelligibility decreases, listening effort might become more important (Mackersie and Cones, 2011; Wild et al., 2012). Moreover, discontinuity of F0 and VTL may have negative effects on higher-level cognitive functions, such as selective attention, that are needed for robust speech recognition in complex listening environments (Best et al., 2008; Larson and Lee, 2013). Hence, the effects of disruptions of voice cues could be greater in real-life listening than shown in the present study.

In summary, indexical cues are important for understanding interrupted speech and VTL seems to be a more important factor than F0. However, despite the reduction in overall intelligibility as a result of the voice disruption, top–down perceptual restoration still occurred with meaningful sentences. While the acoustic cues in remaining speech segments are very important for the restoration of missing parts (Cooper et al., 1985), we propose that when these acoustic cues are not consistent, listeners can overlook them and make use of the linguistic context and rules for phonemic restoration, which is consistent with other reports (Billig et al., 2013; Warren and Sherman, 1974).

These findings have theoretical implications for perceptual organization and practical implications for users of cochlear implants (CIs). The findings imply that perceptual organization is a flexible system that adjusts itself based on what cues are most reliable, be they indexical or linguistic in nature. This contrasts with *prägnanz* law, an assumed general principle of perceptual organization, according to which simplicity governs object formation (Wagemans et al., 2012). In other words, the brain is expected to favor the simplest perceptual organization possible. In vision, Beck (1982) showed that perceptual organization favored simpler properties (such as color) over more complex ones (such as shape) for object formation, indicating a hierarchy in rules. In the auditory system, simpler cues, such as the general spectral profile or harmonic structure of the stimulus, are also believed to be the primary determinants of perceptual organization (Bregman, 1990, pp. 529–594; Darwin and Carlyon, 1995). However, in the current study, these may have been superseded by other factors, such as the linguistic content. Practically, the findings of the present study may have implications for CI users, who show different use of voice information (F0 and VTL) than normal hearing listeners (Fuller et al., in revision). VTL, especially, is not well utilized for gender identification, and considering that the discontinuity of this voice cue reduced intelligibility of interrupted speech, this is perhaps an important factor in the difficulties CI users demonstrate in understanding speech in adverse situations (Nelson et al., 2003; Stickney et al., 2007, 2004). CI users also have difficulty understanding interrupted speech (Bhargava et al., 2014; Nelson and Jin, 2004). On the other hand, and on a positive note, our results indicate that

perceptual restoration could be robust to discontinuous or inconsistent voice cues, which suggests that linguistic information itself can influence perceptual organization (at least in the specific conditions of this study, in the case of high context sentences). Hence, if implant users can similarly utilize linguistic context, they may be able to compensate speech perception difficulties using mechanisms of top–down restoration, and perhaps this can be implemented in special training programs (Benard and Başkent, 2013).

7. Conclusion

We have shown in this study that:

- Voice alternations between consecutive segments of interrupted speech reduced overall intelligibility, confirming the voice as an important perceptual cue for grouping and segregating speech segments.
- The continuity of VTL was more important for the intelligibility of interrupted speech than that of F0, at least for the present voice manipulations.
- None of the disruptions in voice continuity had an effect on the phonemic restoration benefit, even when the two alternating voices were consistently reported to be from different talkers.
- Voice continuity does not seem to be a prerequisite for top–down repair of interrupted speech.

Acknowledgments

The authors would like to thank Kelly Fitz for his assistance in technical aspects of the work, Marije Sleurink for transcribing participant responses, and the participants. The authors would also like to thank the associate editor and the two anonymous reviewers for their valuable comments to improve the quality of this paper. This study was supported by a VIDI grant from the Netherlands Organization for Scientific Research, NWO (grant no. 016.096.397; from Netherlands Organization for Health Research and Development, ZonMw). Further support came from a Rosalind Franklin Fellowship from the University of Groningen, University Medical Center Groningen, and funds from the Heinsius Houbolt Foundation. The study is part of the research program of the Otorhinolaryngology Department of the University Medical Center Groningen: Healthy Aging and Communication.

References

Assmann, P.F., Katz, W.F., 2005. Synthesis fidelity and time-varying spectral change in vowels. J. Acoust. Soc. Am. 117, 886–895.
Bashford, J.A., Riener, K.R., Warren, R.M., 1992. Increasing the intelligibility of speech through multiple phonemic restorations. Percept. Psychophys. 51, 211–217.
Başkent, D., 2012. Effect of speech degradation on top-down repair: phonemic restoration with simulations of cochlear implants and combined electric–acoustic stimulation. J. Assoc. Res. Otolaryngol. 13, 683–692.
Başkent, D., Eiler, C., Edwards, B., 2009. Effects of envelope discontinuities on perceptual restoration of amplitude-compressed speech. J. Acoust. Soc. Am. 125, 3995–4005.
Beck, J., 1982. Organization and Representation in Perception. Erlbaum, Universtity of Michigan.
Benard, M.R., Başkent, D., 2013. Perceptual learning of interrupted speech. PLoS One 8, e58149.
Best, V., Ozmeral, E.J., Kopčo, N., Shinn-Cunningham, B.G., 2008. Object continuity enhances selective auditory attention. Proc. Natl. Acad. Sci. U. S. A. 105, 13174–13178.
Bhargava, P., Gaudrain, E., Başkent, D., 2014. Top-down restoration of speech in cochlear-implant users. Hear. Res. 309, 113–123. http://dx.doi.org/10.1016/j.heares.2013.12.003.
Billig, A.J., Davis, M.H., Deeks, J.M., Monstrey, J., Carlyon, R.P., 2013. Lexical influences on auditory streaming. Curr. Biol. 23, 1585–1589.
Bregman, A.S., 1990. Auditory Scene Analysis: the Perceptual Organization of Sound. The MIT Press.
Cooper, W.E., Tye-Murray, N., Eady, S.J., 1985. Acoustical cues to the reconstruction of missing words in speech perception. Percept. Psychophys. 38, 30–40.
Darwin, C., Carlyon, R., 1995. Auditory grouping. In: Moore, B.C.J. (Ed.), Hearing, Handbook of Perception and Cognition. Academic Press, London, UK, pp. 387–424.
Darwin, C.J., Brungart, D.S., Simpson, B.D., 2003. Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers. J. Acoust. Soc. Am. 114, 2913–2922.
Fitch, W.T., Giedd, J., 1999. Morphology and development of the human vocal tract: a study using magnetic resonance imaging. J. Acoust. Soc. Am. 106, 1511–1522.
Fuller, C., Gaudrain, E., Clarke, J., Galvin, J.J.I., Fu, Q.-J., Free, R., Başkent, D., 2014. Effect of Fundamental Frequency and Vocal-tract Length on Talker Gender Categorization in Cochlear-implant and Normal-hearing Listeners (in revision).
Gaudrain, E., Grimault, N., Healy, E.W., Béra, J.-C., 2007. Effect of spectral smearing on the perceptual segregation of vowel sequences. Hear. Res. 231, 32–41.
Gaudrain, E., Li, S., Ban, V., Patterson, R., 2009. The role of glottal pulse rate and vocal tract length in the perception of speaker identity. Interspeech 1–5, 152–155.
Hartmann, W.M., Johnson, D., 1991. Stream segregation and peripheral channeling. Music Percept. 9, 155–183.
Helfer, K.S., Freyman, R.L., 2009. Lexical and indexical cues in masking by competing speech. J. Acoust. Soc. Am. 125, 447–456.
Hillenbrand, J., Clark, M., 2009. The role of F0 and formant frequencies in distinguishing the voices of men and women. Atten. Percept. Psychophys. 71, 1150–1166.
Ives, D.T., Smith, D.R.R., Patterson, R.D., 2005. Discrimination of speaker size from syllable phrases. J. Acoust. Soc. Am. 118, 3816–3822.
Kawahara, H., Masuda-Katsuse, I., de Cheveigné, A., 1999. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds. Speech Commun. 27, 187–207.
Kidd, G.J., Best, V., Mason, C.R., 2008. Listening to every other word: examining the strength of linkage variables in forming streams of speech. J. Acoust. Soc. Am. 124, 3793–3802.
Ladefoged, P., Broadbent, D.E., 1957. Information conveyed by vowels. J. Acoust. Soc. Am. 29, 98–104.
Larson, E., Lee, A.K.C., 2013. Influence of preparation time and pitch separation in switching of auditory attention between streams. J. Acoust. Soc. Am. 134, EL165–EL171.
Liu, C., Kewley-Port, D., 2004. STRAIGHT: a new speech synthesizer for vowel formant discrimination. Acoust. Res. Lett. Online 5, 31–36.
Mackersie, C.L., Cones, H., 2011. Subjective and psychophysiological indices of listening effort in a competing-talker task. J. Am. Acad. Audiol. 22, 113–122.
Mackersie, C.L., Dewey, J., Guthrie, L.A., 2011. Effects of fundamental frequency and vocal-tract length cues on sentence segregation by listeners with hearing loss. J. Acoust. Soc. Am. 130, 1006–1019.
Maddox, R.K., Shinn-Cunningham, B.G., 2012. Influence of task-relevant and task-irrelevant feature continuity on selective auditory attention. J. Assoc. Res. Otolaryngol. 13, 119–129.
Mclennan, C.T., Luce, P.A., 2005. Examining the time course of indexical specificity effects in spoken word recognition. J. Exp. Psychol. Learn. Mem. Cogn. 31, 306–321.
Miller, G.A., Licklider, J.C.R., 1950. The intelligibility of interrupted speech. J. Acoust. Soc. Am. 22, 167–173.
Nelson, P.B., Jin, S.-H., 2004. Factors affecting speech understanding in gated interference: cochlear implant users and normal-hearing listeners. J. Acoust. Soc. Am. 115, 2286–2294.
Nelson, P.B., Jin, S.-H., Carney, A.E., Nelson, D.A., 2003. Understanding speech in modulated interference: cochlear implant users and normal-hearing listeners. J. Acoust. Soc. Am. 113, 961–968.
Powers, G.L., Wilcox, J.C., 1977. Intelligibility of temporally interrupted speech with and without intervening noise. J. Acoust. Soc. Am. 61, 195–199.
Shahin, A.J., Bishop, C.W., Miller, L.M., 2009. Neural mechanisms for illusory filling-in of degraded speech. Neuroimage 44, 1133–1143.
Shinn-Cunningham, B., Mehraei, G., Bressler, S., Masud, S., 2013. Influences of perceptual continuity on everyday listening. In: POMA. Presented at the 21st International Congress on Acoustics (ICA 2013), Montreal, Canada. ASA, Montreal, Canada, p. 010026.
Skuk, V.G., Schweinberger, S.R., 2013. Influences of fundamental frequency, formant frequencies, aperiodicity and spectrum level on the perception of voice gender. J. Speech Lang. Hear. Res. 57, 285–296. http://dx.doi.org/10.1044/1092-4388(2013/12-0314).
Smith, D.R.R., Walters, T.C., Patterson, R.D., 2007. Discrimination of speaker sex and size when glottal-pulse rate and vocal-tract length are controlled. J. Acoust. Soc. Am. 122, 3628–3639.
Stickney, G.S., Assmann, P.F., Chang, J., Zeng, F.-G., 2007. Effects of cochlear implant processing and fundamental frequency on the intelligibility of competing sentences. J. Acoust. Soc. Am. 122, 1069–1078.
Stickney, G.S., Zeng, F.-G., Litovsky, R., Assmann, P., 2004. Cochlear implant speech recognition with speech maskers. J. Acoust. Soc. Am. 116, 1081–1091.
Studebaker, G.A., 1985. A "rationalized" arcsine transform. J. Speech Hear. Res. 28, 455–462.
Titze, I.R., 1989. Physiologic and acoustic differences between male and female voices. J. Acoust. Soc. Am. 85, 1699–1707.
Tsuzaki, M., Takeshima, C., Irino, T., Patterson, R.D., 2007. Auditory stream segregation based on speaker size, and identification of size-modulated vowel sequences. In: Kollmeier, B., Klump, G., Hohmann, V., Langemann, U.,

Mauermann, M., Uppenkamp, S., Verhey, J. (Eds.), Hearing – from Sensory Processing to Perception. Springer, p. 285.
Verschuure, J., Brocaar, M.P., 1983. Intelligibility of interrupted meaningful and nonsense speech with and without intervening noise. Percept. Psychophys. 33, 232–240.
Versfeld, N.J., Daalder, L., Festen, J.M., Houtgast, T., 2000. Method for the selection of sentence materials for efficient measurement of the speech reception threshold. J. Acoust. Soc. Am. 107, 1671–1684.
Wagemans, J., Feldman, J., Gepshtein, S., Kimchi, R., Pomerantz, J.R., van der Helm, P.A., van Leeuwen, C., 2012. A century of Gestalt psychology in visual perception: II. Conceptual and theoretical foundations. Psychol. Bull. 138, 1218–1252.
Wang, X., Humes, L.E., 2010. Factors influencing recognition of interrupted speech. J. Acoust. Soc. Am. 128, 2100–2111.
Warren, R.M., 1970. Perceptual restoration of missing speech sounds. Science 167, 392–393.
Warren, R.M., 1983. Auditory illusions and their relation to mechanisms normally enhancing accuracy of perception. J. Audio Eng. Soc. 31, 623–629.
Warren, R.M., Sherman, G.L., 1974. Phonemic restorations based on subsequent context. Percept. Psychophys. 16, 150–156.
Wild, C.J., Yusuf, A., Wilson, D.E., Peelle, J.E., Davis, M.H., Johnsrude, I.S., 2012. Effortful listening: the processing of degraded speech depends critically on attention. J. Neurosci. 32, 14010–14021.