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.