Phonemic restoration: The brain creates missing speech sounds

Makio Kashino1,2,3,*

1NTT Communication Science Laboratories, NTT Corporation, 3–1, Morinosato Wakamiya, Atsugi, 243–0198 Japan
2ERATO SHIMOJO Implicit Brain Function Project, Japan Science and Technology Agency
3Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology

Abstract: Under certain conditions, sounds actually missing from a speech signal can be synthesized by the brain and clearly heard. This illusory phenomenon, known as the phonemic restoration effect, reveals the sophisticated capability of the brain underlying robust speech perception in noisy situations often encountered in daily life. In this article, basic aspects of the phonemic restoration effect are described with audio demonstrations.

Keywords: Speech, Perception, Illusion

PACS number: 43.71.Es, 43.71.Pc [doi:10.1250/ast.27.318]

This article contains the supplementary media files. Underlined file names in the article correspond to the supplementary files. For more information, see http://www.asj.gr.jp/2006/data/ast2706.html.

Supplementary files
original.wav, mlt050_bb_no.wav, mlt100_bb_no.wav, mlt200_bb_no.wav, mlt050_bb_high.wav, mlt100_bb_high.wav, mlt200_bb_high.wav, mlt050_bb_low.wav, mlt100_bb_low.wav, mlt200_bb_low.wav, sgl100_high.wav, sgl100_no.wav, mlt100_bpf1500.wav, mlt100_bpf1500_no.wav, mlt100_f0375_high.wav, mlt100_f0750_high.wav, mlt100_f1500_high.wav, mlt100_f3000_high.wav, mlt100_f6000_high.wav, rev025.wav, rev050.wav, rev075.wav, rev100.wav

1. BASIC PHENOMENA

Under certain conditions, sounds actually missing from a speech signal can be synthesized by the brain and clearly heard (for an extensive review, see [1]). This illusory phenomenon, known as the phonemic restoration effect, reveals the sophisticated capability of the brain underlying robust speech perception in noisy situations often encountered in daily life. In this article, I will describe some aspects of the phonemic restoration effect with audio demonstrations. All the audio demonstrations were created digitally on a personal computer (sampling rate: 44.1 kHz, quantization: 16 bit, monaural), based on an utterance “Do you understand what I am trying to say?” spoken by a male native American-English speaker.

A typical demonstration of the phonemic restoration effect is as follows [2]: When portions of a recorded utterance (original.wav; Fig. 1a) are replaced by gaps of silence from 2 to 15 times a second, the utterance sounds disrupted and unnatural, and it is very difficult to understand what is being said (mlt050_bb_no.wav (replaced by silence every 50 ms), mlt100_bb_no.wav (every 100 ms), and mlt200_bb_no.wav (every 200 ms; Fig. 1b)). However, when the gaps are filled with broadband noise that is louder than the recorded voice, the utterance sounds more natural and continuous, making it much easier to understand (mlt050_bb_high.wav (replaced by noise burst louder than the speech by 10 dB RMS every 50 ms), mlt100_bb_high.wav (every 100 ms), and mlt200_bb_high.wav (every 200 ms; Fig. 1c)). The noise has a spectral slope of −6 dB/octave. In all demonstrations described in this article, a raised-cosine ramp was applied to every onset and offset of noise and speech to avoid spectral splatter, and noise and speech segments were cross-faded at their transitions. In both cases, exactly the same amount of the speech signal has been deleted; however, the deleted portions are restored perceptually only when the gaps are filled with louder broadband noise. When the noise is fainter than the speech signal, the restoration effect is hardly observed (mlt050_bb_low.wav (replaced by noise burst fainter than the speech by 10 dB RMS every 50 ms), mlt100_bb_low.wav (every 100 ms), and mlt200_bb_low.wav (every 200 ms; Fig. 1d)).

This perceptual restoration effect clearly indicates that the sounds we hear are not copies of physical sounds. The
brain fills the gaps with the sounds that should exist in the portions masked by noise bursts based on the information in the remaining speech signal. What we perceive is the result of such unconscious interpretation.

The phonemic restoration effect is so strong that listeners cannot notice what segment has been replaced by noise. As a consequence, it is extremely hard to identify the temporal position of the noise burst in the sentence when the phonemic restoration occurs [3,4]. This can be experienced by sgl100_high.wav, in which a 100-ms portion of speech signal is deleted and replaced by a louder broadband noise (10 dB RMS louder than the speech signal). On the other hand, it is much easier to tell what segment is missing in sgl100_no.wav, in which the same portion is left silent (the answer is around the second /n/ in the word “understand”). When the phonemic restoration occurs, the speech and noise make different perceptual streams. It is difficult to judge temporal relationship across different streams, but it is not usually required in everyday situations.

2. THE MASKING POTENTIAL RULE

The phonemic restoration effect involves two aspects: apparent continuity and increased intelligibility. The former aspect is not restricted to speech. The apparent continuity of a sound disrupted by an extraneous noise can be observed for various types of sounds, including music [4], environmental sounds, and pure tones. The phonemic restoration effect can be thought of as a special case of more general auditory continuity illusion [1]. The necessary condition for the auditory continuity illusion is that the acoustic (spectral, temporal, and spatial) characteristics of the interrupting sound must be sufficient to mask the interrupted sound if the two sounds were presented.
simultaneously [1,5]. This condition is known as the masking potential rule.

A way to demonstrate the masking potential rule in the frequency domain is to use band-limited speech and noise [6]. First, the original speech signal is filtered by a bandpass filter having a 1/3-octave band centered at 1,500 Hz (mlt100_bpf1500.wav). In spite of the extreme band limitation, this is highly intelligible. Next, every 100 ms portion of the signal is deleted (mlt100_bpf1500_no.wav). The intelligibility drops remarkably. Now, the silent intervals are replaced with noise bursts with a 1/3-octave band centered at various frequencies (mlt100_f0375_high.wav (noise center frequency is 375 Hz), mlt100_f0750_high.wav (750 Hz), mlt100_f1500_high.wav (1,500 Hz; Fig. 2a), mlt100_f3000_high.wav (3,000 Hz), and mlt100_f6000_high.wav (6,000 Hz; Fig. 2b); noise level was 10 dB RMS higher than the speech signal.). The speech signal seems most continuous when the center frequencies of the speech and noise are matched (at 1,500 Hz).

The masking potential rule is quite reasonable considering everyday situations. Perceptual restoration of a segment is appropriate when that segment is actually present but masked by an extraneous sound. If the segment could not have been masked by the extraneous sound, then synthesizing the segment would be inappropriate. The selectivity of perceptual restoration reduces the possibility of an inappropriate perception of a signal fragment.

3. CUES EXPLOITED FOR THE RESTORATION OF PHONETIC SEGMENTS

As described in the previous section, the apparent continuity in the phonemic restoration is determined by the masking potential rule, which holds for the auditory continuity illusion in general. Then, how about increased intelligibility? Apparently, the brain exploits at least two types of redundancy specific to speech in the reconstruction of missing segments.

The first cue is a phenomenon called coarticulation. Speech signal is produced by the movement of articulatory organs such as lips, a tongue, and a jaw, all of which cannot move freely and abruptly. Instead, these organs move smoothly in a highly cooperative manner, resulting in the articulation of adjacent phonetic segments interacting with each other. This is coarticulation. As a consequence of coarticulation, information for a phonetic segment distributes over time in the range of a few hundred milliseconds, overlapping with that for adjacent segments. The brain can exploit such acoustic redundancy of speech signal in the phonemic restoration [7].

Another cue is semantic context provided by sentences. In some cases, semantic information that appears after the disrupted portion can affect the perceptual restoration of the disrupted portion in an apparently retroactive fashion [8]. Semantic context seems to work when the masking potential rule is met [9].

4. RELATED PHENOMENA

Finally, I mention two types of perceptual phenomena related to the phonemic restoration effect. The first one is the perceptual restoration of locally time-reversed speech [10]. First, the original sentence is divided into segments of fixed duration. Then, every segment is time-reversed (rev025.wav (segment duration 25 ms), rev050.wav (segment duration 50 ms), rev075.wav (segment duration...
Intelligibility is quite high up to the segment duration of 50 ms. This manipulation destroys fine spectro-temporal structure, but the global structure larger than the segment duration is preserved. Phonetic perception is possible based on slow coarticulatory information distributed over time.

Speech signal is redundant not only in the time domain but also in the frequency domain. Perceptual restoration of missing information can occur also in the frequency domain [11]. For example, a speech stimulus consisted of two widely separated narrow bands of speech is not very intelligible. However, when noise was introduced in the spectral gap separating the two speech bands, intelligibility increases significantly. This spectral restoration is analogous to the phonemic restoration in time domain: both represent mechanisms for minimizing interference when extraneous sounds replace portions of speech.

These phenomena, as well as the phonemic restoration effect, demonstrate the sophisticated capabilities of the brain exploiting various types of redundancy in speech signal to realize stable recognition of linguistic messages even under noisy situations in the real world.

ACKNOWLEDGMENT

The author thanks Hieu M. Nguyen.

REFERENCES