INVITED REVIEW

Perception of noise-vocoded speech sounds: Sentences, words, accents and melodies

Hiroshi Riquimaroux*

*Department of Knowledge Engineering and Computer Sciences, Doshisha University,
Kyotanabe, 610-0321 Japan*

Abstract: Recent works on perception of noise-vocoded speech sound (NVSS) have revealed that amplitude envelope information is very important for speech perception when spectral information is not sufficiently available. Basically, the fundamental frequency information is not available and formant peaks cannot be identified in NVSS. However, we can even recognize accent and distinguish male voice from female voice in NVSS. More, melody can be created from lyrics once lyrics are intelligible. In the present study, findings from fMRI measurement are introduced to show neural activities in the central nervous system during listening to NVSS. The present data indicate that various sites in the brain, which are not ordinarily used for speech recognition, participate in making NVSS intelligible. Applications of the present work include an innovative speech processor and a training system for hearing impaired people.

Keywords: Speech perception, Noise-vocoded speech sounds, Brain plasticity, Amplitude envelope, Functional MRI

PACS number: 43.71.Ft, 43.71.Ky, 43.71.Pc, 43.72.Ja [doi:10.1250/ast.27.325]

This article contains the supplementary media files (see Table 1). Underlined file names in the article correspond to the supplementary files. For more information, see <http://www.asj.gr.jp/2006/data/ast2706.html>.

1. INTRODUCTION

It has been believed that frequency information is essential for speech perception, e.g., formants and the fundamental frequency. However, recent studies adopting processed speech made of noise bands, called the noise-vocoded speech sound (NVSS), have revealed that amplitude envelope information may also play an important role in speech perception [1,2]. NVSS is a synthesized speech sound where the speech signal is replaced by several bands of noise while the amplitude envelope is preserved (Figs. 1 and 2). NVSS is hardly intelligible when one listens to it for the first time without any training. However, after a short period of training, NVSS consisted of more than two bands of noise becomes very intelligible. Correctly perceived mora rate exceeds 80% with NVSS made of more than three bands of noise (Fig. 3). Japanese accents are known to be carried by the fundamental frequency, which can be also identified with NVSS even without the fundamental frequency. In Japanese, a difference in accent makes a total difference in the meaning of a word. So, these

findings have suggested that amplitude envelope may carry important information for speech perception usually carried by the fundamental frequency. The present paper strongly shows that speech perception may be executed depending upon the amplitude envelope information rather than frequency or spectral information when frequency information is not sufficient.

2. SYNTHESIS OF NVSS

We fixed frequency ranges of NVSS. For one band NVSS, the range was between 0 and 4,000 Hz. For two band NVSS, the first range was between 0 and 600 Hz and the second range was between 600 and 4,000 Hz. For three band NVSS, the first range was between 0 and 600 Hz, the second was between 600 and 1,500 Hz, and the third was between 1,500 and 4,000 Hz. For four band NVSS, the first range was between 0 and 600 Hz, the second was between 600 and 1,500 Hz, the third was 1,500 and 2,100 Hz, and the fourth was between 2,100 and 4,000 Hz [2–4]. Here, the process to synthesize the four band NVSS is representatively explained. First, the original speech sound was bandpass-filtered to be divided into four frequency bands (0–600, 600–1,500, 1,500–2,100 and 2,100–4,000 Hz). The outputs of each frequency band were lowpass-filtered (< 16 Hz) to obtain amplitude envelopes of each frequency band [1]. Then, the amplitude envelopes of each frequency band was multiplied by a noise with the same frequency

*e-mail: hrikimar@mail.doshisha.ac.jp

Table 1 Noise-vocoded speech sound (NVSS) files used.

Sentence	
Bokuwa kirikabuni koshikaketa: I sat on a stump	
Original:	bokuwa.wav
1-band NVSS:	bokuwa_nvs1.wav
2-band NVSS:	bokuwa_nvs2.wav
3-band NVSS:	bokuwa_nvs3.wav
4-band NVSS:	bokuwa_nvs4.wav
Word accent	
Hashi: chopsticks vs. bridge	
F_0 fall, Ampl. fall:	hashi_f-a-.wav
F_0 fall, Ampl. const.:	hashi_f-ae.wav
F_0 fall, Ampl. rise:	hashi_f-a+.wav
F_0 rise, Ampl. fall:	hashi_f+a-.wav
F_0 rise, Ampl. const.:	hashi_f+ae.wav
F_0 rise, Ampl. rise:	hashi_f+a+.wav
F_0 fall, Ampl. fall, NVSS:	nv_hashi_f-a-.wav
F_0 fall, Ampl. const., NVSS:	nv_hashi_f-ae.wav
F_0 fall, Ampl. rise, NVSS:	nv_hashi_f-a+.wav
F_0 rise, Ampl. fall, NVSS:	nv_hashi_f+a-.wav
F_0 rise, Ampl. const., NVSS:	nv_hashi_f+ae.wav
F_0 rise, Ampl. rise, NVSS:	nv_hashi_f+a+.wav
Songs	
Kirakiraboshi and Kaerunouta	
Kirakiraboshi:	kirakira-orig.wav
Kaerunouta:	kaeru-orig.wav
Kirakiraboshi lyrics with Kaerunouta melody:	kira-lyric.kaeru-melody.wav
Kirakiraboshi NVSS:	nv_kirakira-orig.wav
Kaerunouta NVSS:	nv_kaeru-orig.wav
Kirakiraboshi lyrics with Kaerunouta melody NVSS:	nv_kira-lyric.kaeru-melody.wav

band. The noise was generated by normally distributed random numbers. Finally, the amplitude modulated noises of each frequency band were added to create NVSS. (Fig. 1).

3. CHARACTERISTICS OF NVSS

Without any listening trainings NVSS is hardly intelligible. However, an easy training with short sentences (5 sentences/session), just like an example shown here, for a short period (about ten sessions are enough) would improve intelligibility dramatically (correct mora perception reaches greater than 80%, Fig. 3). In training sessions correct sentences were visually given after each training sentence. However, in test sessions no feedbacks were given. Any sentences once used were never used again for the same subject in training or test sessions. The intelligibility increases as a function of the number of frequency bands. The one-band NVSS is little intelligible, while the three- and four-band NVSS are very intelligible [1]. We preserve the amplitude envelope, while speech signals are

replaced by noise in each band. Therefore, frequency information is greatly reduced while the amplitude envelope remains intact. Lowpass filtering of 16Hz hardly alters the amplitude envelope [1]. Perception of NVSS, which mainly depends on information about temporal change in the amplitude envelope, has been studied [1,2]. Here, examples made from one Japanese sentence, “Bokuwa kirikabuni koshikaketa, (I sat on a stump)” are given. There are five examples, one band NVSS ([bokuwa_nvs1.wav](#)), two band NVSS ([bokuwa_nvs2.wav](#)), three band NVSS ([bokuwa_nvs3.wav](#)), four band NVSS ([bokuwa_nvs4.wav](#)) and the original speech sound ([bokuwa.wav](#)) (Fig. 4). I would suggest readers to start with one band NVSS and increase the number of bands, then finally listen to the original speech. Please do not try in the opposite direction.

A series of studies on Japanese NVSS have revealed the following facts. Correct response rate obtained from single words stays very low and that from meaningless words becomes even lower (Murata, 2000 [5]). While individual talker identification or talker identification within the same gender appears to be difficult, discrimination between male and female talkers are feasible (Riquimaroux and Katayama, 2003 [6]). Interestingly, we may perceive accent (change in pitch) of Japanese words in NVSS, which does not contain the fundamental frequency (F_0). Accent of a word is normally carried by a temporal change in F_0 . In Japanese, a change in F_0 is usually accompanied by a change in the amplitude envelope.

How can we perceive accents in NVSS of Japanese, which are carried by temporal change in F_0 ? A possible hypothesis is that the central nervous system would create perception of F_0 change by translating changes in the amplitude envelope into changes in F_0 . Accents cannot be carried by the amplitude envelope no matter how it would change in time but carried by temporal change in F_0 when speech signals are not replaced by bands of noise. Therefore, it is likely that only after getting rid of F_0 by transferring speech sounds into NVSS, accents are produced within the central nervous system by utilizing the amplitude envelope information. In other words, as long as F_0 information is available, accents are carried by F_0 not by the amplitude envelope (Fig. 5, [3]).

Here, examples of a pair of words, hAshi (chopsticks) and hashI (bridge), made by four-band NVSS are shown in Fig. 5. There are two types of F_0 patterns. F_0 falls in hAshi (top row), while F_0 rises in hashI (bottom row). There are three types of amplitude envelopes, descending (left), constant (middle) and ascending (right). So, we have six types of word not transferred into NVSS and six types of word transferred into NVSS. For the words not transferred into NVSS, accents are determined by F_0 change in all cases, while for NVSS words, accents are determined by

Fig. 1 Four-band speech processor. This figure schematically shows how to produce 4-band noise-vocoded speech sounds (NVSS).

Fig. 2 Comparison between an original speech sound “hashi” (chopsticks) and a synthesized noise-vocoded speech sound from the original.

Fig. 3 Time course of training of 5 subjects. Percent correct increased and easily exceeded 80% during short training sessions.

Fig. 4 Amplitude and spectrum changes in an example sentence made of noise-vocoded speech sounds (NVSS). Top left: original speech sound, *bokuwa.wav*. Top right: 2-band NVSS, *bokuwa_nvs2.wav*. Bottom left: 3-band NVSS, *bokuwa_nvs3.wav*. Bottom right: 4-band NVSS, *bokuwa_nvs4.wav*. The sentence is *Bokuwa kirikabuni koshikaketa* (I sat on a stump). One-band NVSS, *bokuwa_nvs1.wav*, is not illustrated here.

Fig. 5 An example of accent perception. Two-mora words, *hashi*, with different accent were used; “*hAshi*” (chopsticks) in top row and “*hashI*” (bridge) in bottom row. Changes in amplitude envelope are illustrated, high-low (left), high-high (middle), and low-high (right). When the speech signal was not transformed into bands of noise, the accent was always determined by a change in F_0 but not by a change in amplitude envelope. However, the accent was ruled by a change in amplitude envelope once the speech signal was replaced by noise bands.

the change in the amplitude envelope. The word *hAshi* with only amplitude envelope modification but without transferring into NVSS (*hashi_f-a+.wav*, top right Fig. 5) gives the identical accent to the word with the original amplitude envelope (top left). However, once the word is transferred into NVSS (*nv_hashi_f-a+.wav*), the accent becomes the same as *hashI* (bottom right Fig. 5). In the similar way, the word *hashI* with only amplitude envelope modification but without transferring into NVSS (*hashi_f+a-.wav*) would not change the accent of the word. The accent remains identical to the word with the original amplitude envelope (bottom right Fig. 5). However, when the word is transferred into NVSS (*nv_hashi_f+a-.wav*), the perceived accent becomes the same as *hAshi* (top left). So, these findings have suggested that amplitude envelope may carry important information for supra-segmental speech perception.

Previous studies have indicated that melodies of songs with well known lyrics can be recognized to some extent [7]. Melody of songs is usually determined by subtle changes in F_0 . However, F_0 of songs transferred into NVSS cannot be detectable because F_0 moves within the first frequency range of NVSS (between 0 and 600 Hz). So, conventional information to determine melody cannot be obtained in NVSS. Therefore, the melody of NVSS must be created by the central nervous system based on recognized lyrics [7]. In other words, when lyrics are not intelligible, melody of the songs cannot be created.

For example, if one knows the first chorus but does not

know the second, the lyrics of the first chorus can generate a melody but a melody may not be created when the second chorus is presented [8]. Here, two songs with an identical rhythm are presented. They are “Kirakiraboshi (Twinkle star)” (original: *kirakira-orig.wav*, NVSS: *nv_kirakira-orig.wav*, Fig. 6a) and “Kaerunouta (Frog song)” (original: *kaeru-orig.wav*, NVSS: *nv_kaeru-orig.wav*, Fig. 6b). Further, the melody of Kaerunouta with lyrics of Kirakiraboshi is given (original: *kira-lyric_kaeru-melody.wav*, NVSS: *nv_kira-lyric_kaeru-melody.wav*, Fig. 6c). Please check if you recognize a melody of each song.

4. PROCESSING OF NVSS BY CENTRAL NERVOUS SYSTEM

Trainings are usually essential for NVSS to be effective. Because the central nervous system processes natural speech sounds by mainly using frequency information, the existing neural circuit for speech recognition is designed to extract speech signals from frequency information encoded in the nervous system. Without frequency information it is difficult for the system to extract speech signals. However, speech signals are also encoded in amplitude envelope information, which has not been utilized but existing in the nervous system. What we need is to establish a network that can translate NVSS into speech sounds by using amplitude envelope information. So, we need plasticity in the central nervous system to process NVSS. Areas, which are not involved in processing

Fig. 6 Spectral and amplitude patterns of original songs and synthesized songs with NVSS. Three examples with the identical rhythm pattern are shown here. **a:** *Kirakiraboshi* (Twinkle star), **b:** *Kaeruno-uta* (Frog song), and **c:** *Kirakiraboshi* lyrics with *Kaeruno-uta* melody.

of natural speech sounds, may be expected to participate in processing of NVSS. Recent fMRI studies have shown that frontal and parietal cortices are involved in processing of NVSS in addition to ordinary temporal cortex used to process natural speech sounds [9]. Thus, perception of NVSS is an integrated outcome of memory, learning and speech production.

5. CLINICAL APPLICATION OF NVSS

It is conceivable that a creation of a system to recognize NVSS may result from a creation of a pathway, which may extract speech signals from NVSS by using the amplitude envelope information rather than the spectral information. Therefore, we expect that unimpaired areas in the central nervous system might process natural speech sounds by using their amplitude envelope information after the training to perceive NVSS [9]. NVSS has been used for the hearing impaired and aphasic people [10–12]. A male patient with congenital hearing loss has shown high correct recognition ratio for NVSS without trainings. This may be the case where the patient trained himself in his life by utilizing the amplitude envelope [12]. A real-time processor to generate NVSS in a hearing-aid case has already been developed for versatile trainings for impaired people [13].

6. SUMMARY

Readers must have experienced how easy for our central nervous system to create intelligible speech sounds from NVSS. The system must have employed the plasticity in the central nervous system to establish a new pathway to recognize NVSS. We may expect that NVSS could be utilized for foreign language learning, prevention of dementia and the fields related to leaning and memory.

ACKNOWLEDGEMENTS

The present series of work has been partly supported by Special Coordination Funds from MEXT, by Grant to RCAST at Doshisha University from MEXT, by Grant-in-Aid for Scientific Research on Priority Area (B) “Prosody and Speech Processing” from MEXT and by the Innovative Cluster Creation Project promoted by MEXT.

REFERENCES

- [1] R. V. Shannon, F. G. Zeng, V. Kamath, J. Wygonski and M. Ekelid, “Speech recognition with primarily temporal cues,” *Science*, **270**, 303–305 (1995).
- [2] Y. Obata and H. Riquimaroux, “Speech perception based on temporal amplitude change with spectrally degraded synthetic sound,” *Tech. Rep. Psychol. Physiol. Acoust. Acoust. Soc. Jpn.*,

- H-99-6, pp. 1–8 (1999).
- [3] Y. Obata and H. Riquimaroux, “Intelligibility of synthesized Japanese speech sound made of band noise—Preliminary study for a speech recognition processor utilizing central auditory function—,” *Tech. Rep. Psychol. Physiol. Acoust. Acoust. Soc. Jpn.*, H-2000-3, pp. 1–8 (2000).
- [4] T. Ishizaki, “Perception of speech sounds where frequency information is replaced by bands of noise: Optimal boundary frequencies between noise bands,” Unpublished Bachelor’s Thesis, Doshisha University (2001).
- [5] T. Murata, “Perception of Japanese noise-vocoded speech sounds: Sentences and words,” Unpublished Bachelor’s Thesis, Doshisha University (2000).
- [6] H. Riquimaroux and T. Katayama, “To what extent the degraded band-noise processed speech is intelligible? Speaker discrimination,” *Tech. Rep. Psychol. Physiol. Acoust. Acoust. Soc. Jpn.*, **33**, 25–30 (2003).
- [7] H. Riquimaroux, “Can we perceive song melodies without the fundamental frequency: Degraded band-noise processed speech,” *Tech. Rep. Psychol. Physiol. Acoust. Acoust. Soc. Jpn.*, **32**, 77–84 (2002).
- [8] H. Riquimaroux, “Perception of noise-vocoded children’s songs: Effects of familiarity of lyrics and melody and tempo,” *Tech. Rep. Psychol. Physiol. Acoust. Acoust. Soc. Jpn.*, **33**, 647–652 (2003).
- [9] R. Tachibana and H. Riquimaroux, “Cortical activities related to noise-vocoded speech intelligibility measured by functional magnetic resonance imaging,” Abstract CD, 752.13, The 34th Annual Meeting, Society for Neuroscience, San Diego (2004).
- [10] K. Mitamura, M. Sudo, T. Kanda, T. Adachi, T. Nakaichi, S. Sakamoto and H. Riquimaroux, “Perception of noise-vocoded speech sounds: Data from patients with impaired vascular system in the brain,” *Tech. Rep. Psychol. Physiol. Acoust. Acoust. Soc. Jpn.*, **34**, 177–182 (2004).
- [11] T. Nakaichi, K. Watanuki, S. Sakamoto, M. Harada and H. Riquimaroux, “Perception of degraded band-noise processed speech by subjects with mild to moderate hearing loss,” *Tech. Rep. Psychol. Physiol. Acoust. Acoust. Soc. Jpn.*, **33**, 141–146 (2003).
- [12] T. Nakaichi, S. Sakamoto and H. Riquimaroux, “Perception of noise-vocoded speech sound by congenital hearing-impaired subjects,” *Tech. Rep. Psychol. Physiol. Acoust. Acoust. Soc. Jpn.*, **34**, 523–528 (2004).
- [13] T. Nakaichi, S. Sakamoto and H. Riquimaroux, “Development of a real-time processor for generating noise-vocoded speech sounds,” *Tech. Rep. Psychol. Physiol. Acoust. Acoust. Soc. Jpn.*, **35**, 243–248 (2005).