©2013 The Acoustical Society of Japan

ACOUSTICAL LETTER

Effects of presentation level on the identification of speech with altered short-term spectra

Hisaaki Tabuchi*

Department of Psychology, Utah State University, Logan, UT 84322-2810, USA

(*Received 26 September 2012, Accepted for publication 28 November 2012*)

Keywords: Presentation level, Phase spectra, Amplitude spectra, Short-term Fourier transform, Overlap addition method

PACS number: 43.66.Nm, 43.72.Ja [doi:10.1250/ast.34.229]

1. Introduction

Since Rhode [1] discovered that the amplitude of basilar membrane (BM) motion is highly compressive at moderate to high presentation levels ($\gtrsim$40–60 dB SPL), the compressive mechanism has been found, especially at tone frequency close to the characteristic frequency (CF) [2]. A consequence of the compressive BM displacement can be viewed as a rate-intensity function that gives the discharge rate for single auditory nerve (AN) fibers as a function of tone level [3,4]. A typical sigmoid function is estimated using free parameters, such as spontaneous and saturation rates, threshold, and slope; these parameters change with tone frequency and CF. The slope generally rises as tone frequency and CF become close, whereas the slope diminishes as tone frequency and CF become further apart. Rose *et al.* [5] equivalently displayed a number of rate-intensity functions for different tone frequencies at a fixed CF as iso-intensity contours, which describe the discharge rate as a function of tone frequency with different tone levels. In general, the iso-intensity contours spread out as tone level increases and the extent of "spreading" reflects the sequence of increasing slopes from rate-intensity functions.

The phenomenon of "bandwidth expansion" allows one to assume that the internal representation of speech with different spectra could change with presentation levels. The utility of manipulating the short-term spectrum of speech was evaluated in the 1960s for the purpose of designing an efficient coding algorithm for the vocoder [6,7]. Schroeder [8] was among the first to report that speech is unintelligible when the phase spectra are randomized (random-phase speech; RPS) in a long 100 s window, whereas it is intelligible when randomized in a short 50 ms window. The raised intelligibility was interpreted by Schroeder as a high sensitivity of phase spectra over a short period of time. Oppenheim and Lim [9] found that "phase-retained" speech with flat amplitude spectra (flat-spectrum speech; FSS) was intelligible when processed with a long window, whereas "amplitude-retained" speech with zero phase spectra (cosine-phase speech) was unintelligible. Oppenheim and Lim demonstrated that the unintelligibility of cosine-phase speech in a long window can be attributed to the degraded amplitude spectra from unprocessed speech.

Three later studies contributed precise measurements. Liu *et al.* [10] systematically manipulated amplitude and phase spectra for German consonants across various window lengths. They showed that the intelligibility of FSS increases as window length increases (from 16 to 512 ms), but the intelligibility of RPS, in contrast to FSS, decreases as window length increases. Paliwal and Alsteris [11] obtained results similar to those of Liu *et al.*'s study by using Australian English consonants; they found, however, that the intelligibility of FSS can vary with the type of window function and overlap length, as applied in the overlap addition method. More recently, Kazama *et al.* [12] swapped the amplitude and phase spectra of Japanese sentences with those of white noise and generated similar stimuli to RPS and FSS, and demonstrated that the correlation of envelope waveforms between unprocessed and spectrum-altered speech can capture the general trend of intelligibility across window lengths. Kazama *et al.* interpreted the effect of window lengths to be evidence that the intelligibility varies with the time-frequency resolution of an envelope waveform.

This study was aimed at measuring the intelligibility of RPS and FSS in a varying window length and obtaining previously unreported changes in intelligibility with level. The effect of the presentation level was expected to vary with the type of spectral alteration and window lengths as the amplitude spectra change in accordance with these two factors. Finally, the spectrograms of RPS and FSS are displayed to discuss a plausible amplitude-spectrum cue.

2. Methods

The listeners were seven native speakers of American English between the ages of 19 and 27. All had thresholds below 20 dB SPL for pure tones at octave frequencies between 0.25 and 4 kHz. The listeners provided informed consent and were paid an hourly wage for their participation.

Sentences from the Coordinate Response Measure (CRM) corpus [13] were processed and presented for identification. Each sentence includes a call sign, a color ("blue," "red," "white," or "green") and a number (1 to 8), in a carrier of the form "Ready Baron go to blue one now." A subset of CRM sentences all spoken by the same male speaker using one call sign (Baron) was used. Each of these sentences was resampled from 44.1 kHz to 22.05 kHz and padded with equal numbers of preceding and trailing zeros to create a set of 32 sentences, each 4096 ms in length. The sentences were then processed in

[*] e-mail: hitab@shinshu-u.ac.jp. Present address: School of Education, Shinshu University, 6-ro, Nishinagano, Nagano, 380–8544 Japan

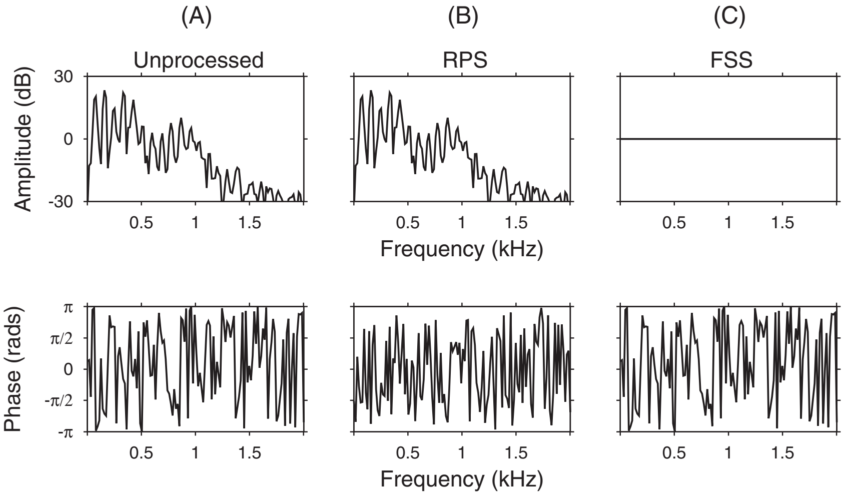

Fig. 1 Short-term Fourier spectra before and after spectral alterations. A: Spectral estimates obtained for a sample of unprocessed speech. Amplitude (upper panel) and phase (lower panel) spectra were calculated for a single 64 ms window. B: Spectral estimates obtained after processing the signal shown in Panel A to produce RPS. The estimate of the amplitude spectrum was unchanged from Panel A to Panel B, but a different phase spectrum was obtained. C: Spectral estimates obtained after processing the signal shown in Panel A to produce FSS. The estimate of the phase spectrum was unchanged from Panel A to Panel C, but the amplitude spectrum within the 64 ms window was perfectly flat.

one of two ways. In each case, the waveform was divided into fixed-length analysis windows. The window length varied by a factor of two from 1 to 1,024 ms (with minor rounding due to the sample period), and successive windows always overlapped by 50 percent of the window length. Within each analysis window, a Hanning window was applied and the short-term Fourier transform was calculated.

Amplitude and phase spectra calculated for a representative 64 ms window from one sentence are shown in Fig. 1(A). To generate what will be referred to as "random-phase speech" (RPS), the complex coefficients in each short-term Fourier spectrum were altered in a way that left the amplitude spectrum unchanged, but replaced the phase spectrum with a sequence of random values in the range from $-\pi$ to $+\pi$ radians. Figure 1(B) shows the results of that alteration; the amplitude spectrum is identical to the one shown in Fig. 1(A), but the phase spectrum has changed. To generate what will be referred to as "flat-spectrum speech" (FSS), the complex coefficients in each short-term Fourier spectrum were altered in a way that set each amplitude coefficient to be constant while leaving the phase spectrum unchanged. The results of that alteration are shown in Fig. 1(C); the amplitude spectrum is flat, but the phase spectrum matches the original phase spectrum shown in Fig. 1(A).

The waveforms presented for identification were reconstructed by summing the processed waveforms across windows. For data collection, the reconstructed waveforms were presented at three nominal levels, 40, 55, and 70 dBA. The actual level on individual trials was varied within a range of ±3 dB with 1 dB steps. The original unprocessed sentences were also presented at the same overall levels.

Measurements were made in a single-walled, sound-attenuating booth (IAC). Sounds were delivered from a high-quality sound card (Echo Gina 3G) and presented diotically over headphones (Sennheiser HD280 Pro). The listener responded by clicking on one of 32 colored numbers presented in a 4×8 grid on a computer screen. Feedback was provided after every trial and block of trials. Intelligibility (probability of correct response) for both color and number was shown for the results. At the beginning of the experiment, each listener completed three blocks of practice trials with unprocessed speech at 55 dBA to minimize learning effects. The order of conditions was randomized across listeners.

In summary, a three-factorial ($2 \times 11 \times 3$) experiment was conducted with the following independent variables: two spectra alternations (RPS and FSS), eleven window lengths (between 1 and 1,024 ms), and three presentation levels (40, 55, and 70 dBA). In addition, the intelligibility for unprocessed speech was measured at these three levels as control conditions. Each sentence was repeated twice under each condition.

3. Results

Figure 2 shows intelligibility for RPS and FSS as a function of window length. The identification of RPS (Fig. 2(A)) was essentially perfect for window lengths from 4–128 ms, but scores decreased rapidly for window lengths outside that range. The pattern was nearly independent of presentation level. A repeated-measures ANOVA indicated that the effect of level for RPS was not significant [$F(2, 12) = 1.17, p = 0.34$].

The pattern for FSS shown in Fig. 2(B) was more complicated. Essentially, perfect intelligibility was obtained in two separate regions, one with window lengths from 1 to 4 ms, and another including window lengths from 128–1,024 ms. For intermediate window lengths between 8–64 ms, intelligibility was lower but increased with presentation level. For these window lengths, error-free identification was never observed, even at the highest level tested, 70 dBA. A

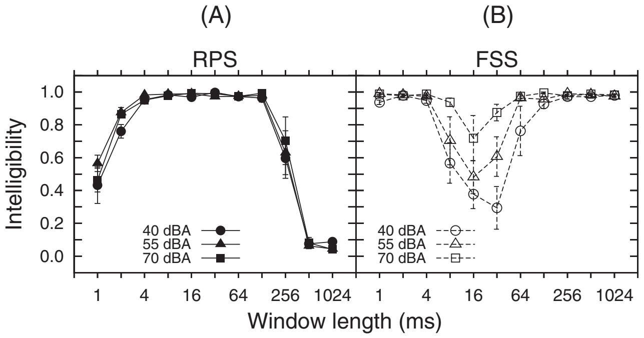

Fig. 2 Intelligibility for RPS (Panel A) and FSS (Panel B) as a function of window length and presentation level. Each curve represents intelligibility at one presentation level. Each symbol is the mean across seven listeners; the error bars are standard errors. The mean intelligibility for unprocessed sentences was 0.97 at 40 dBA, and 0.99 at 55 and 70 dBA (not shown).

repeated-measures ANOVA indicated that the effect of level for FSS was highly significant [$F(2,12) = 33.76, p < 0.001$].

Figure 2 illustrates the complementary pattern of identification accuracy. At the window lengths for which intelligibility for RPS was high, intelligibility for FSS tended to be low. For conditions under which intelligibility for RPS was low, intelligibility for FSS at the same window lengths tended to be high. The ANOVA also indicated that the interaction between the type of spectrum alteration and window length was highly significant [$F(10,60) = 62.71, p < 0.001$].

4. Discussion

The effect of altering the short-term spectrum on speech intelligibility had previously been examined by several groups of investigators [9–12]. In the present study, we used slightly different processing methods and different speech stimuli, but the pattern of results is generally consistent in all the reports. In each case, an interaction between window length and spectral processing condition was observed. The previous investigations did not consider stimulus level, so the effects of presentation level reported here had not been observed before. The identification of RPS was not affected by presentation level in any meaningful way. In contrast, presentation level had a strong effect on the identification of FSS, for an intermediate range of window lengths at which performance was less than perfect.

The finding in the present study and in those reported previously [9–12] that speech described as having a flat spectrum may be perfectly intelligible may seem counterintuitive. It is important to emphasize that the signal described as FSS has a flat spectrum only in spectral estimates calculated in the same window for which the original processing was done, that is, for estimates like those shown in Fig. 1(C). Figure 3(B) shows the spectrograms for unprocessed speech and FSS-1,024 ms at the overall speech level, 55 dB SPL. The nonflat amplitude distribution of FSS-1,024 ms negates the initial intent of making the spectrum flat. Spectral estimates obtained from the reconstructed waveforms that were presented for identification and/or made

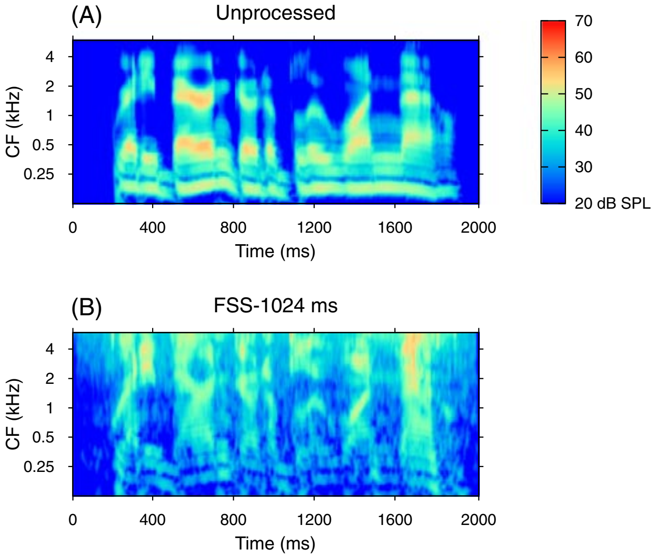

Fig. 3 Amplitude spectrograms at overall level of 55 dB SPL obtained for a sample sentence. Panel (A): Unprocessed speech. (B): FSS-1,024 ms. The spectrograms were obtained through the 4th-order band-pass Butterworth filter from 125 to ~6 kHz with 16 bands per octave, shifted by 10 ms frames with 5 ms steps. The color scale represents pointwise levels per time-and-frequency between 20 and 70 dB SPL. CF indicates the logarithmic center frequency.

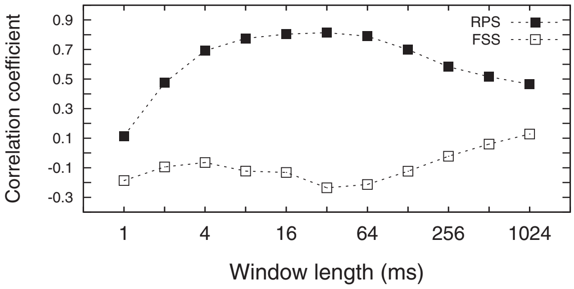

Fig. 4 Amplitude correlation for RPS and FSS across window lengths at the overall presentation level 55 dB SPL. Each point consists of the average of amplitude correlations calculated from all the 32 CRM sentences used for the experiment. Each standard deviation was smaller than 0.02. The correlation was calculated from levels per time-and-frequency higher than 20 dB SPL.

with different analysis windows revealed that some amplitude cues from the original sentences are retained in FSS as similarly observed by Paliwal and Alsteris [11].

In order to evaluate spectrum similarity, a spectrum correlation between unprocessed and modified speech was calculated, that can be compared to intelligibility. As shown in Fig. 4, the correlation of RPS is high for intermediate window lengths (4–64 ms), whereas it decreases outside of the intermediate range. Thus, high intelligibility of RPS for intermediate window lengths can be explained as being due to the highly correlated, preserved amplitude spectra from unprocessed speech, whereas low intelligibility for long (> 128 ms) and short window lengths (< 4 ms) reflects degraded amplitude spectra. On the other hand, the amplitude correlation of FSS is the lowest for intermediate window lengths, but it increases for long windows. These results of amplitude correlation suggest that the perceptual effects of the amplitude spectrum vary with window length as the amplitude spectra accordingly change with window length. This amplitude correlation, however, does not help explain the level-dependent effects of FSS.

The level effects of FSS could be inferred from the increasing bandwidth with noise level, which was revealed by the human psychoacoustical results of tone detection in notched-noise masking [14]. Several studies showed that the bandwidth tends to expand as the masker level and tone frequency increase [15]; bandwidth especially expands toward the lower frequency regions from the tone frequency [16–18]. The level-dependent intelligibility of FSS-16 ms may similarly reflect the expanded bandwidths to lower frequency regions in response to the amplitude spectra at high-frequency regions. Additional experiments and analyses are being carried out in an attempt to understand the acoustic basis for the psychophysical results.

Acknowledgments

This work was supported by the DC010615 funding to Donal G. Sinex. The author thanks Dr. Nandini Iyer for providing CRM recordings.

References

[1] W. S. Rhode, "Observations of the vibration of the basilar membrane in squirrel monkeys using the Mössbauer technique," *J. Acoust. Soc. Am.*, **49**, 1218–1231 (1971).

[2] L. Robles, M. A. Ruggero and N. C. Rich, "Basilar membrane mechanics at the base of the chinchilla cochlea. I. Input-output functions, tuning curves, and response phases," *J. Acoust. Soc. Am.*, **80**, 1364–1374 (1986).

[3] M. B. Sachs and P. J. Abbas, "Rate versus level functions for auditory nerve fibers in cats: tone-burst stimuli," *J. Acoust. Soc. Am.*, **81**, 680–691 (1974).

[4] G. K. Yates, I. Winter and D. Robertson, "Basilar membrane nonlinearity determines auditory nerve rate-intensity functions and cochlear dynamic range," *Hear. Res.*, **45**, 203–219 (1990).

[5] J. E. Rose, J. E. Hind, D. J. Anderson and J. F. Brugge, "Some effects of stimulus intensity on response of auditory nerve fibers in the squirrel monkey," *J. Neurophysiol.*, **34**, 685–699 (1971).

[6] M. R. Schroeder, "Vocoders: Analysis and synthesis of speech," *Proc. IEEE*, **54**, 720–734 (1966).

[7] J. L. Flanagan and R. M. Golden, "Phase Vocoder," *Bell Syst. Tech. J.*, **45**, 1493–1509 (1966).

[8] M. R. Schroeder, "Models of hearing," *Proc. IEEE*, **63**, 1332–1350 (1975).

[9] V. A. Oppenheim and S. J. Lim, "The importance of phase in signals," *Proc. IEEE*, **69**, 529–541 (1981).

[10] L. Liu, J. He and G. Palm, "Effects of phase on the perception of intervocalic stop consonants," *Speech. Commun.*, **22**, 403–417 (1997).

[11] K. K. Paliwal and D. L. Alsteris, "On the usefulness of STFT phase spectrum in human listening tests," *Speech. Commun.*, **45**, 153–170 (2005).

[12] M. Kazama, S. Gotoh, M. Tohyama and T. Houtgast, "On the significance of phase in the short term Fourier spectrum for speech intelligibility," *J. Acoust. Soc. Am.*, **127**, 1432–1439 (2010).

[13] R. S. Bolia, W. T. Nelson, M. A. Ericson and B. D. Simpson, "A speech corpus for multitalker communications research," *J. Acoust. Soc. Am.*, **107**, 1065–1066 (2000).

[14] D. L. Weber, "Growth of masking and the auditory filter," *J. Acoust. Soc. Am.*, **62**, 424–429 (1977).

[15] S. Rosen and D. Stock, "Auditory filter bandwidths as a function of level at low frequencies (125-1 kHz)," *J. Acoust. Soc. Am.*, **92**, 773–781 (1992).

[16] R. A. Lutfi and R. Patterson, "On the growth of masking asymmetry with stimulus intensity," *J. Acoust. Soc. Am.*, **76**, 739–745 (1984).

[17] M. L. Hicks and S. P. Bacon, "Psychophysical measures of auditory nonlinearities as a function of frequency in individuals with normal hearing," *J. Acoust. Soc. Am.*, **105**, 326–338 (1999).

[18] B. R. Glasberg and B. C. J. Moore, "Frequency selectivity as a function of level and frequency measured with uniformly exciting notched noise," *J. Acoust. Soc. Am.*, **108**, 2318–2328 (2000).