ARCHIVES OF ACOUSTICS
32, 4 (Supplement), 325–330 (2007)

PERCEPTUAL CONSEQUENCES OF CHANGE IN VOCODED SPEECH PARAMETERS FOR VARIOUS REVERBERATION CONDITIONS

Szymon DRGAS, Magdalena A. BŁASZAK

Adam Mickiewicz University
Institute of Acoustics
Umultowska 85, 64-614 Poznań, Poland
e-mail: Szymon.Drgas@amu.edu.pl

(*received July 15, 2007; accepted November 14, 2007*)

The perceptual consequence of change in parameters of vocoded speech for various reverberation conditions has been examined. The three controlled variables were: number of bands, instantaneous frequency changes rate and reverberation conditions. The effects were quantified in terms of: (a) non-words' recognition scores for young normal-hearing listeners, (b) "ease of listening" based on the time of reaction (response delay) and (c) the subjective measure of difficulty (ten-degree scale). The results have shown that the fine structure information is a relevant cue in speech perception in reverberation conditions. It has also been observed that only the slow variations of instantaneous frequency are critical in perception. A good correlation between all subjective measures considered was found in this study.

Keywords: vocoder, speech perception, reverberation conditions.

1. Introduction

In order to transmit speech signal effectively, compression procedures have to be employed. Lossy compression methods give high efficiency. However, degradation of speech signal can affect its robustness to noise and reverberation. Decoded speech sound can be difficult to understand in adverse environment. For example, in videoconference speech signal transmitted via Internet can be played in a room and then reflections distort played sound.

In order to find optimal strategy for speech signal compression, critical properties of speech signal have to be determined. To assess which features of speech signal are critical in speech perception, in given acoustical environment, one can degrade speech and measure speech intelligibility. If a substantial decrease in intelligibility is observed after reducing some feature in the speech signal, the critical significance is assigned to this feature.

There are several ways to degrade the speech signal. SHANNON *et al*. [3] showed that the slow modulations are the most important characteristics of the speech signal. These slow modulations carry phonetic information. They showed that four bands of noise modulated by temporal envelope extracted from the original signal can give a high speech intelligibility. However, there is a high decrease in vocoded speech intelligibility in the presence of additive noise [6]. Vocoded speech has much lower intelligibility when the fine structure in respective bands is not transmitted. SMITH *et al*. [5], in experiments with auditory chimeras have shown that fine structure carries the information needed for pitch perception. It is well known that pitch information is significant in auditory streaming. STICKNEY *et al*. [5] have proved that the fine structure reduced to slow frequency modulations give higher speech intelligibility in the presence of additive noise, especially of concurrent speakers. The signal model can be expressed by the following equation [2]:

$$s(t) \approx \sum_{k=1}^{N} x_k(t) = \sum_{k=1}^{N} A_k(t) \cos\left[2\pi f_{ck} t + 2\pi \int_0^t g_k(\tau)\,\mathrm{d}\tau + \theta_k\right], \qquad (1)$$

where $x_k(t)$ denotes a signal in the k-frequency band, $A_k(t)$ is the temporal envelope in the k-th band, f_{ck} is the central frequency and g_k is the frequency modulating signal in the k-th band.

ZENG *et al*. [6] have shown that a slow frequency modulation in every carrier frequency can raise speech intelligibility in the presence of additive noise. The vocoded voice should be adapted not only to environmental noise but also to different acoustic situations, especially different reverberation conditions. Thus, the situation-dependent adaptation of vocoded speech is the main topic of this research thus an influence of parameters of vocoded speech additionally degraded by convolutive noise on itelligibility and "ease of listening" is investigated. The main aim of the study is to determine which features of the speech signal are critical to ensure the robustness of the speech signal to reverberation.

2. Experiment

2.1. Subjects and apparatus

Six untrained young normal-hearing listeners (students at the age 20–25, with the hearing threshold $<$20 dB HL at octave intervals from 125–8000 Hz) participated in the experiment. None of the subjects had a history of hearing difficulties. A PC-compatible computer with a signal processor (TDT System 3) generated a stimulus through a 24-bit D/A converter (RP2) at the 24.414 kHz rate, recorded the listener's responses and excuted the experimental procedure. The signals were presented binaurally via the Sennheiser HD 580 headphones. All equipment was located outside of a double-walled, sound-attenuated booth, where the listeners were seated .

2.2. Stimulus and experimental scenery

The speech material consisted of Polish logatoms [1]. These words were processed with vocoder [2]. The speech signal was filtered by means of bandpass filters. Cutoff frequencies were set uniformly according to the Greenwood frequency map. Intelligibility was measured for 6 and 12 bands. The amplitude envelope and instantaneous frequency in every band were extracted. First, the amplitude envelope was filtered with a lowpass filter (cutoff frequency 500 Hz). The waveform of instantaneous frequency was clipped to the limit frequency range of 500 Hz. In the next stage this clipped waveform was lowpass filtered with cutoff frequencies of 50 and 400 Hz. The condition with no frequency modulation was also considered (0 Hz). These extracted and processed parameters were used synthesize the speech signal. The sinusoidal signals with frequencies equal to the frequencies of the analyzing filters were amplitude and frequency modulated by extracted modulating signals. These signals were filtered through the filters with the same parameters as those of the analyzing signals.

The synthesized signals were played back and recorded in three enclosures. The experiments were conducted in three types of sound fields: direct sound only (an anechoic chamber) direct sound with reverberation (room 1 and room 2), see Fig. 1. Data were gathered on disk and prepared for listening tests. The listeners were asked to (a) repeat the logatom to a microphone (response delay measurements), (b) write it correctly (intelligibility measurements) and (c) evaluate on a 10-degree scale a subjective listening difficulty.

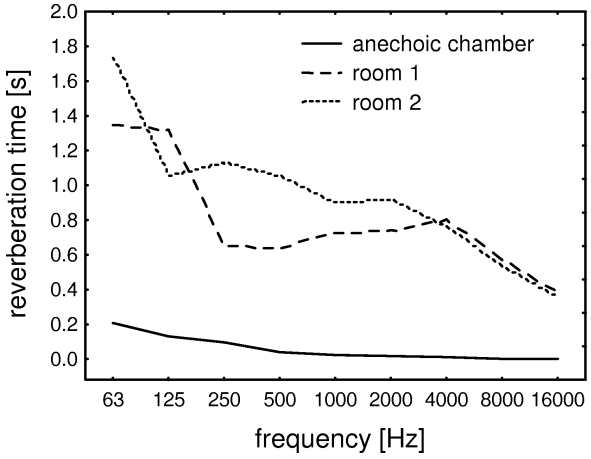

Fig. 1. The reverberation time in rooms versus frequency.

3. Results

The data gathered in the experiment were subjected to a 3-way analysis of variance (ANOVA) with respect to the number of bands, the FM cutoff frequency and the enclo-

sure (reverberation time) was tested. The analysis has shown that all of the factors and almost all the interactions tested turned out to be statistically significant (see Table 1).

Table 1. The analysis of variance (ANOVA) for intelligibility, subjective rating of difficulty and reaction time.

	Intelligibility		Subjective rating of difficulty	
Enclosure	$F(2, 10763) = 22.56$	$p < 0.05$	$F(2, 10763) = 249.52$	$p < 0.05$
Number of bands	$F(1, 10763) = 215.38$	$p < 0.05$	$F(1, 10763) = 1429.81$	$p < 0.05$
FMC	$F(2, 10763) = 61$	$p < 0.05$	$F(2, 10763) = 1750.39$	$p < 0.05$
Enclosure$*$Number of bands	$F(2, 10763) = 0.53$	$p < 0.05$	$F(2, 10763) = 8.55$	$p < 0.05$
Enclosure$*$FMC	$F(4, 10763) = 2.79$	$p < 0.05$	$F(4, 10763) = 18.50$	$p < 0.05$
Number of bands$*$FMC	$F(2, 10763) = 2.61$	$p < 0.05$	$F(2, 10763) = 16.54$	$p < 0.05$
Number of bands$*$Enclosure$*$FMC	$F(4, 10763) = 1.55$	$p < 0.05$	$F(4, 10763) = 3.26$	$p < 0.05$

	Reaction time (response delay)	
Enclosure	$F(2, 4848) = 30.72$	$p < 0.05$
Number of bands	$F(1, 4848) = 61.17$	$p < 0.05$
FMC	$F(2, 4848) = 24.12$	$p < 0.05$
Enclosure$*$Number of bands	$F(2, 4848) = 1.133$	$p > 0.05$
Enclosure$*$FMC	$F(4, 4848) = 0.38$	$p > 0.05$
Number of bands$*$FMC	$F(2, 4848) = 5.34$	$p < 0.05$
Number of bands$*$Enclosure$*$FMC	$F(4, 4848) = 0.68$	$p > 0.05$

Figure 2 presents the speech intelligibility results. In the left panel (a) the results gathered for six channel vocoder are presented, while in the right one (b) the twelve-channel vocoder data are shown. The speech intelligibility has been plotted as a function of frequency modulation cutoff frequency (fmc). The parameter of the presented curves is the enclosure, in which the vocoded speech was recorded.

As can be seen from Fig. 2, for the six-channel vocoder, the speech intelligibility increases from about 12% to 60% along the increase in fmc, while in the case of twelve-channel vocoder changes from 40% to 80%. Thus, for the higher number of bands there is about 30% increase in the speech intelligibility. It can be noted that in the case of anechoic chamber there is no significant speech intelligibility improvement resulting from the FM cutoff frequency change from 50 to 400 Hz, while in the reverberating conditions the improvement seems to play important role. This result suggests that high frequency FM modulations significant influence the speech intelligibility in reverberation conditions.

Figure 3a presents the results of reaction time measurements (response delay) [s]. The differences caused by the reverberation time of the room can be also noted – the

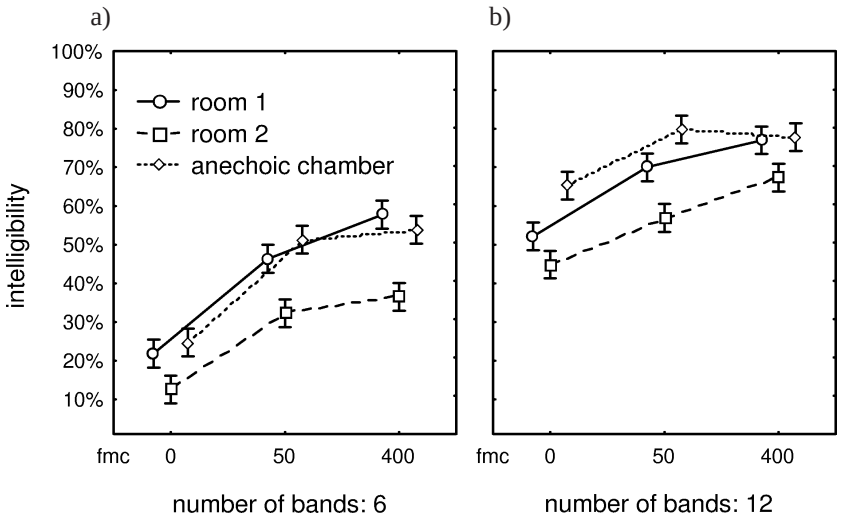

Fig. 2. Speech intelligibility scores as a function of the number of bands, FM cutoff frequency (fmc). The parameter is the reverberation condition (room1, room2, anechoic chamber).

smallest response delay was found for the signals recorded in the anechoic chamber, while the highest in room 2 (with the greatest reverberation). The increase in the effect of high frequency FM modulation with reverberation time of the room can be observed both for the six-channel and the twelve-channel vocoder, however the reaction time is markedly higher when the six band vocoder is used. The subjective rating difficulty results presented in Fig. 3b show that the greatest decrease in the difficulty with the FM cutoff frequency can be noted for the range of 0–50 Hz rather than for 50–400 Hz, which is in line with the intelligibility measurements.

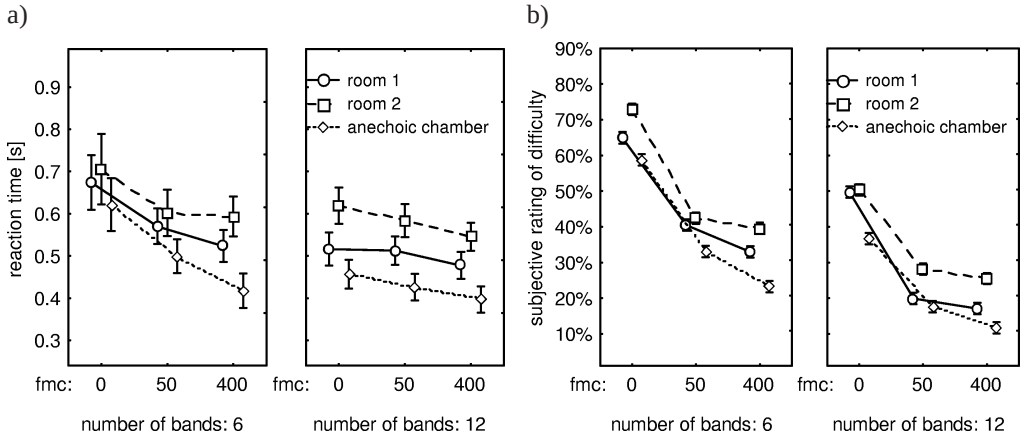

Fig. 3. Response delay values (a) and subjective rating of difficulty (b) as a function of the number of bands, FM cutoff frequency (fmc) and reverberation conditions (room1, room2, anechoic chamber).

4. Conclusions

The paper reports the results of the vocoded speech intelligibility and "listening difficulty" (in the binaural presentation) measured in different acoustical conditions. The results have shown that the fine structure information is a relevant cue in speech perception in reverberation conditions. However, in the anechoic chamber there is no significant speech intelligibility improvement when the FM cutoff frequency was changed from 50 to 400 Hz. This result suggest that high frequency variations of instantaneous frequency have significant role in speech recognition in reverberation conditions. Nevertheless, it should be emphasized that the improvement is observed in the listening difficulty rating, although the speech intelligibility does not increase, the comfort of listening does. The results obtained for the different numbers of bands, FM cutoff frequencies and the reverberation conditions have shown that all these parameters are important in the perception. However, only the slow variations of the instantaneous frequency (< 50 Hz) seem to be critical for the speech intelligibility in anechoic conditions while in reverberant rooms fast fluctuations of instantaneous frequency are also significant.

References

[1] BRACHMAŃSKI S., STARONIEWICZ P., *Phonetic structure of a test material used in subjective measurements of speech quality* [in Polish], Speech and Language Technology, **3**, 71–80 (1999).

[2] NIE K. B., STICKNEY G. S., ZENG F.-G., *Encoding frequency modulation to improve cochlear implant performance in noise*, IEEE Transactions on Biomedical Engineering, **52**, 1, 64–73 (2005).

[3] SHANNON R. V., ZENG F.-G., KAMATH V., WYGONSKI J., EKELID M., *Speech recognition with primarily temporal cues*, Science, **270**, 303–304 (1995).

[4] STICKNEY G. S., NIE K.,ZENG F.-G., *Contribution of frequency modulation to speech recognition in noise*, J. Acoust. Soc. Am., **118**, 2412–2420 (2005).

[5] SMITH Z. M., DELGUTTE B., OXENHAM A. O., *Chimaeric sounds reveal dichotomies in auditory perception*, Nature, **416**, 87–90 (2002).

[6] ZENG F.-G., NIE K. B., STICKNEY G. S., KONG Y.-Y., VONGPHOE M., BHARGAVE A., WEI C. G., CAO K., *Speech recognition with amplitude and frequency modulations*, Proceedings of the National Academy of Science, **102**, 2293–2298 (2005).