Effect of Harmonic Structure of Noises on Noisy Vowel Perception

Kentaro Ishizuka and Kiyoaki Aikawa

NTT Communication Science Laboratories
Morinosato-Wakamiya 3-1,
Atsugi City, Kanagawa, 243-0198, Japan.
ishiuzuka@atom.brl.ntt.co.jp, aik@idea.brl.ntt.co.jp

ABSTRACT

This paper reports new findings on noisy vowel perception experiments designed to obtain a new feature parameter for noise-robust automatic speech recognition. To obtain this new parameter, we analyze the human auditory mechanism. We conducted two experiments to examine the way in which listeners perceive natural vowels under very noisy environmental conditions, namely a signal to noise ratio (SNR) of around -2 dB. First, we used eight types of noise; white noise and seven types of harmonic structured noise each with the same flat spectral envelope and energy. Spectral envelopes have been widely used as the feature parameter for automatic speech recognition. However, our experimental results showed that perceptual identification scores differ significantly depending on the detailed spectral shape of the noise. The result suggests that the human auditory system uses sound features that are more detailed than the spectral envelopes to perceive vowels in noisy environments. The difference between noises suggests that the even harmonic components of a vowel contribute less to noisy vowel perception than the odd harmonic components. Furthermore, the result implies that the human auditory system changes dynamically in its use of time/frequency features corresponding to waveform and spectral structure. Second, we used five types of harmonic structured noise each with a fundamental frequency close to that of the vowels. The result suggests that vowel perception is affected depending on each fundamental frequency of the noise and each SNR.

1. INTRODUCTION

Conventional automatic speech recognition is insufficiently robust with regard to environmental noise. A goal of our research is to analyze the human auditory mechanism to obtain a new robust feature parameter.

The spectral envelope has been widely used as the feature parameter for automatic speech recognition [1]. Therefore, when the spectral envelope of the input speech in a very noisy environment differs greatly from that of clean speech, it is difficult to recognize the noisy speech with conventional automatic speech recognition methods. Specifically, the automatic speech recognition results are the same for all input sounds when the noise power spectrum exceeds that of the speech signals. This is because all input sounds have similar spectral envelopes regardless of the speech included in the sounds.

However, the human auditory system can identify target sounds in such noisy environments depending on the kind of noise [2]. It is reasonable to suppose that the human auditory system extracts other acoustical features using certain sound cues.

Noisy speech perception has long been studied as part of research undertaken in such fields as speech intelligibility, masking [3][4], phonemic restoration [5], and auditory filters [6]. In addition, tone perception has been studied in the presence of noise with a harmonic structure [7][8][9]. However, little is known about speech perception in the presence of noise, with a harmonic structure.

In this paper, we report on vowel perception in very noisy environments. We conducted two experiments to examine how listeners perceive natural vowels under such conditions at a signal to noise ratio (SNR) of around -2 dB. First, we used eight types of noise to examine the effect of the harmonic structure of noises on vowel perception. The eight types we used were white noise and seven types of harmonic structured noise with the same flat spectral envelope and energy. We decided the harmonic structure of the noises based on the mean fundamental frequency of the vowels. Second, we used five types of harmonic structured noise, each with its fundamental frequency close to that of the vowels, to examine the effect of the interval between the fundamental frequency of the noises and vowels. Because we could not examine this effect of such slight difference between the fundamental frequencies in the first experiment. Hereafter, the first experiment is described as "experiment 1", and the second experiment as "experiment 2."

2. EXPERIMENT 1

2.1 Stimuli

Stimuli were produced by adding one of the noises to one of the vowels at various SNRs.

The vowel samples were five Japanese vowels /a/, /i/, /u/, /e/, and /o/ in the ATR Japanese speech database. These vowels were spoken solely by a female Japanese speaker. Their mean powers were adjusted so that they were the same. Since the original data sampling rate was 12 kHz, we re-sampled them at 11.025 kHz. Their fundamental frequencies were from 254.3 to 266.3 Hz; which we calculated by the autocorrelation method. The mean fundamental frequency was 259.7 Hz.

Eight types of noise were produced related to the mean fundamental frequency of the vowels. These noises were composed of a pure tone and its harmonics each with the same amplitude. The maximum frequencies of the harmonics were the maximum number less than the Nyquist frequency. These noises are described below, and Fig.1 shows their spectra.

(A) white noise
(B) a pure tone of 260 Hz and its harmonics
(C) a pure tone of 520 Hz and its harmonics
(D) a pure tone of 260 Hz and its odd harmonics
(E) a pure tone of 130 Hz and its harmonics
(F) a pure tone of 130 Hz and its harmonics except for a pure tone of 260 Hz and its odd harmonics
(G) a pure tone of 130 Hz and its harmonics except for a
pure tone of 520 Hz and its harmonics
(H) a pure tone of 130 Hz and its odd harmonics

Stimuli were produced by adding one of these noises to one of the vowels at SNRs of -1.58, -2.28, and -2.92 dB. The powers of the vowels were constant. The total number of stimuli was 120. The stimuli included a 10 ms taper and the duration was around 0.3 of a second. The sampling frequency of the stimuli was 11.025 kHz.

2.2 Procedure

The stimuli were presented diotically through headphones (STAX SR-1A Signature) in a soundproof room. The vowels were presented at 60 dB SPL. The stimuli were presented at intervals of six seconds. The subjects were asked to identify the vowel they heard for each stimulus and indicate the vowel on a mark sheet. The same stimulus was presented five times to each subject, making a total of 600 tests. We randomized the order in which the stimuli were presented to each subject. We employed a computer to produce the stimuli and to control the experiment.

We used an introductory session with five clean vowels, five noises, and eight example stimuli to familiarize the subjects with the procedure and stimuli.

**Table 1:** The mean 5-vowel identification score and standard deviation for each kind of noise in experiment 1. The mean identification score for each SNR is also shown.

<table>
<thead>
<tr>
<th>Noise</th>
<th>Mean</th>
<th>StdDev</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A)</td>
<td>33.73%</td>
<td>8.39%</td>
</tr>
<tr>
<td>(B)</td>
<td>59.93%</td>
<td>7.11%</td>
</tr>
<tr>
<td>(C)</td>
<td>66.80%</td>
<td>7.24%</td>
</tr>
<tr>
<td>(D)</td>
<td>60.27%</td>
<td>7.54%</td>
</tr>
<tr>
<td>(E)</td>
<td>67.27%</td>
<td>8.33%</td>
</tr>
<tr>
<td>(F)</td>
<td>59.47%</td>
<td>6.41%</td>
</tr>
<tr>
<td>(G)</td>
<td>59.93%</td>
<td>8.08%</td>
</tr>
<tr>
<td>(H)</td>
<td>58.80%</td>
<td>8.95%</td>
</tr>
</tbody>
</table>

**SNR**

<table>
<thead>
<tr>
<th>SNR</th>
<th>(A)</th>
<th>(B)</th>
<th>(C)</th>
<th>(D)</th>
<th>(E)</th>
<th>(F)</th>
<th>(G)</th>
<th>(H)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.58dB</td>
<td>52.20%</td>
<td>63.60%</td>
<td>66.60%</td>
<td>55.80%</td>
<td>64.60%</td>
<td>66.60%</td>
<td>54.80%</td>
<td>45.20%</td>
</tr>
<tr>
<td>-2.28dB</td>
<td>26.80%</td>
<td>63.60%</td>
<td>66.60%</td>
<td>55.80%</td>
<td>64.60%</td>
<td>66.60%</td>
<td>54.80%</td>
<td>45.20%</td>
</tr>
<tr>
<td>-2.92dB</td>
<td>22.20%</td>
<td>63.60%</td>
<td>66.60%</td>
<td>55.80%</td>
<td>64.60%</td>
<td>66.60%</td>
<td>54.80%</td>
<td>45.20%</td>
</tr>
</tbody>
</table>

2.3 Subjects

The 20 subjects were 10 male and 10 female Japanese who were from 19 to 65 years old. All subjects were paid. None had a history of hearing difficulties.

2.4 Results

The mean vowel identification score for each kind of noise is presented in Fig. 2 and Table 1. The figure shows the standard deviation of the mean identification score and the identification score for each SNR. There was a 20% chance of guessing the answer correctly.

The identification score was the highest with noise (E), and the lowest with noise (A). Of the stimuli with noises (B)-(H), the mean identification score with noise (H) was the lowest. Regardless of the types of noise, the mean identification scores became lower as the SNR became lower. The identification score with noise (A), which had an SNR of -2.92 dB, was as low as the chance level, but the scores with the other noises were all higher than 40%.

An ANOVA confirmed that the mean identification scores were significantly different for the different types of noise [F=35.834, p<0.001]. The Turkey-Kramer multiple comparison test confirmed that there was a significant difference between the mean identification scores with noise (A) and the other noises (alpha=0.01), between the identification scores with noise (E) and noise (F)(H) [alpha=0.05], and between the scores with noise (C) and noise (H) [alpha=0.05].
2.5 Discussion

While each noise has the same flat spectral envelope, the identification score with noise (A) is significantly lower than that of the other noises. In contrast, the identification scores with the other noises do not fall below 40%. This result suggests that the human auditory system uses more detailed features of sounds rather than the spectral envelopes such as used in conventional speech recognition to perceive vowels in noisy environments.

One reason for the highest identification score with noise (E) is that the vowel features in the frequency region are perceived more easily than the other noises. Noise (E) is composed of a 130 Hz pure tone and its 41 harmonics each with the same amplitude. All of the harmonics were with the same phase at the time origin. Therefore, the power spectral peak of noise (E) is the lowest among noises (B)-(H), because noise (E) has the most components. The result also suggests that the vowel features in the time scale region are perceived more easily than the other noises. The waveform of noise (E) decays rapidly at intervals of 1/130 seconds (about 7.6 ms), because noise (E) has many tone components. Therefore, it becomes easier to perceive the vowel features from the waveform between the noise pulses. (See also Fig. 3).

The identification score with noise (C) is high. Since noise (C) is composed of tones close to the even harmonics of the vowels, this result suggests that the even harmonic components of the vowel contribute less to noisy vowel perception than the odd harmonic components. Therefore, this result also suggests that fundamental frequency estimation is important as regards perceiving vowels in noisy environments. One reason for the identification score with noise (H) being the lowest of the scores for noises (B)-(H) is that the vowel harmonics are even harmonic components of the stimulus with noise (H).

3. EXPERIMENT 2

3.1 Stimuli

Stimuli were produced by the same method used for experiment 1. The vowels were the same as those used for experiment 1 and their mean fundamental frequency was 259.7 Hz.

Five types of noise were produced related to the mean fundamental frequency of the vowels. These noises were composed of a pure tone and its harmonics each with the same amplitude. The maximum frequencies of the harmonics were the maximum number less than the Nyquist frequency. These noises are described below. Noise (K) is the same as noise (B) in experiment 1.

(I) a pure tone of 220 Hz and its harmonics
(J) a pure tone of 240 Hz and its harmonics
(K) a pure tone of 260 Hz and its harmonics
(L) a pure tone of 280 Hz and its harmonics
(M) a pure tone of 300 Hz and its harmonics

Stimuli were produced by adding one of the above noises to one of the vowels at SNRs of -1.58, -2.28, and -2.92 dB. The powers of the vowels were constant. The total number of stimuli was 75. The stimuli included a 10 ms taper and the duration was 0.3 of a second. The sampling frequency of the stimuli was 11.025 kHz.

3.2 Procedure

The stimuli were again presented dichotically through headphones (STAX SR-A Signature) in a soundproof room. The vowels were presented at 60 dB SPL. The stimuli were presented at intervals of six seconds. The subjects were asked to identify the vowel they heard for each stimulus and enter their answer on a mark sheet. The same stimulus was presented to each subject five times giving a total of 375 tests. The stimuli were randomly presented to each subject. We used a computer
to produce the stimuli and to control the experiment.
This experiment was conducted 15 minutes after experiment 1. There was no introductory session because the subjects were the same for both experiments.

3.3 Subjects
All the subjects were the same as those who participated in experiment 1.

3.4 Result
The mean vowel identification score for each kind of noise is shown in Fig.4 and Table 2. The figure includes the standard deviation of the mean identification score and the identification score for each SNR. There was a 20% chance of guessing the answer correctly.

The mean identification score with noise (I) was the highest, and that under noise (K) was the lowest. Regardless of the type of noise, the identification scores became lower as the SNR became lower. All the identification scores were higher than 40%.

An ANOVA confirmed that the mean identification scores were significantly different for the different types of noise [F=3.205, p<0.05]. The Turkey-Kramer multiple comparison test confirmed that there was a significant difference between the identification scores with noise (K) and noise (I) [alpha=0.05]. It should be noted that the standard deviations of the identification scores were high. However, the mean identification scores for each subject had an approximately normal distribution.

However, it should be noted that the vowel identification score with noise (I) was not the lowest at an SNR of -1.58 dB. In addition, the identification score with noise (I) was not always the highest at each SNR.

3.5 Discussion
The mean identification score with noise (K) is the lowest. In addition, noise (K) has the most similar harmonics structure to the vowel. Therefore, we may say that vowel perception is affected more when the fundamental frequency of the noise is closer to that of the vowel.

However, this is not true in all cases. At an SNR of -1.58 dB, the identification score with noise (J) is the lowest. It is difficult to determine which factor affected the vowel perception from the results of experiment 2. We can speculate that such factors as masking, beating, harmonic interval or SNR may be involved but we need more experimental data and analysis before we can be certain.

4. CONCLUSION
In this paper, we described two experiments designed to examine the effect of the harmonic structure of noises on noisy vowel perception. Experimental results showed that perceptual identification scores differ significantly depending on the detailed spectral shape of the noise. In addition, we reported some new findings on the vowel perception in very noisy environments.

The result of experiment 1 revealed that the vowel perception is affected more by non-harmonic structured noise with a flat spectrum. The effects of the harmonic structure of noises differ greatly depending on the number of harmonic components and power spectral shape of the components.

The results suggest that the human auditory system uses more detailed features of sounds than just the spectral envelopes to perceive vowels in noisy environments, and that fundamental frequency estimation is important to the perception of vowels in noisy environments. Furthermore, a reasonable explanation is that the human auditory system changes dynamically in its use of time/frequency features corresponding to waveform and spectral structure. In addition, the results suggest that vowel perception is affected in various ways depending on each fundamental frequency of the noise and each SNR when the harmonic frequencies of the noise are close to those of vowels.

ACKNOWLEDGEMENT
The authors thank Dr. Norihiro Hagita, Head of Media Information Laboratory, NTT Communication Science Laboratories. The authors also thank members of Dialogue Understanding Research Group for stimulating discussions.

REFERENCES