Japanese Listeners’ Perception of English Fricatives in AMR-NB Cell Phone Speech

Junichi Fujinuma s1150202

Abstract

From the fact that Japanese listeners have difficulty perceiving English fricative sounds in natural speech, it is natural to expect that such listeners would have even more difficulty perceiving fricatives in degraded speech. This research tests the ability of 72 Japanese university students to identify fricatives in English words presented in cell phone speech simulated with the AMR-NB codec. Participants were presented with 120 stimuli – 40 at CD quality speech (44.1 kHz), 40 downsampled to 8.0 kHz, and 40 resembling cell phone speech. There were 20 words, each with a fricative consonant (/s/, /ʃ/, /θ/, /θ/, /z/, /dʒ/, /v/ or /ð/) in either word-initial or word-final position (e.g., “sick” or “mouse” for /s/). In addition, words were presented in two different contexts: in isolation and embedded in a carrier sentence. Results showed that the ability to correctly identify the fricatives depends not only on signal type, but also on the type of fricative and whether it is located at the beginning or end of the word. We found that voiceless sounds are more often misidentified than voiced sounds, in most contexts and across most signal types. We also found that having the stimulus word appear at the end of a carrier sentence helps perception of the fricative sound in that stimulus. An interesting point is, perceptibility of /θ/ as in “thick” was not significantly better than chance in both types of speech signal (normal and cell phone). It can be presumed that the reason for this is the Japanese language does not contain the sound /θ/. Correlation of perceptibility with duration of voiceless fricatives and perceptibility with mean intensity are also described in this paper to determine what might be affecting the perceptibility rate.

1 Introduction

In 2010, the number of global cell phone subscriptions reached 5 billion [1], compared to only about 1.0 to 1.5 billion fixed phone subscriptions [2]. Many of the cell phone conversations that take place for business, education and research are carried out in English, the de-facto standard language all over the world. Because cell phone signals are compressed to save bandwidth, some information in the speech signal is lost. For people whose native language is not English, this degraded signal could cause problems in English communication. This is especially true for fricative sounds: /s/, /ʃ/, /θ/, /θ/, /z/, /dʒ/, /v/ and /ð/, aperiodic sounds that have important spectral information at higher frequencies – the frequencies that are cut to compress the cell phone speech signal. Although research has been done testing native listeners’ perception of degraded speech signals [3], and testing non-native listeners’ perception of fricatives in normal speech [4], to our knowledge, little research has been done on non-native listeners’ perception of fricatives in a degraded speech signal. It should be noted that all non-native listeners in our study live in their homeland from their birth and have average English proficiency in their country, compared to listeners with very high proficiency in English spending several years in the United States in past research [5] [6].

In this research, we focus on Japanese listeners’ perception of English fricatives in cell phone (degraded) speech. Specifically, the present study was undertaken in order to compare Japanese university students’ ability to identify fricatives in English words presented in three types of signals: speech sampled at 44.1 kHz (“Norm” for normal), speech sampled at 8 kHz (“Down” for downsampled), and speech that was created with a codec to simulate cell phone speech with 8 kHz sampling rate, band-pass filtered (“Cell” for cell phone). In addition, within each signal type (Norm, Down, and Cell), words were presented in two different contexts: in isolation and embedded in a carrier sentence.

Mayo et al. [6] tested how age of acquisition influences perception of second-language speech and claimed, “This study indicates that even when non-native listeners have developed a high level of fluency, their ability to discriminate second-language speech in the presence of noise is influenced by the age during which they acquired the language”. Miller & Nicely [3] studied native listeners’ fricative confusions in noisy speech, and they found that /s/ is confused with /θ/, /θ/ with /θ/, and /v/ with /ð/. The rate of correct identification was 75.4% for /ʃ/, 75.0% for /v/, 74.6% for /s/, 51.5% for /θ/, and 49.1% for /θ/. Of these five English fricative sounds, only /s/ occurs naturally in Japanese (a sound similar to /ʃ/ occurs, but it is produced with the two lips instead of the upper teeth and lower lip). Lamberger et al. [4] studied Japanese university students’ perception of English
fricatives in clear-signal nonsense syllables (i.e., not real words), and they found identification rates of 76.19% for /ʃ/, 71.35% for /ʃ/, 55.20% for /θ/ - very similar rates to Miller & Nicely [3]. The similarity of the identification rates shown in the above two studies show that when the signal is degraded, a native listener performs at about the same rate as a non-native speaker does when the signal is clear. An important question to answer, then, is how a non-native speaker (already at a linguistic disadvantage) performs when the signal is degraded (a further disadvantage), as it is during cell phone speech.

2 Method

2.1 Participants

We gave the perception test to 72 Japanese undergraduate students ranging in age from 18 to 22 (mean = 19.75, standard deviation = 0.78). All the participants were enrolled in the same professor’s sophomore English class, “Listening and Reading”, held at the University of Aizu in Japan, and the experiments were held at the beginning of the semester. All participants had at least six years of English education at the junior high and senior high school levels, basically focused on grammar translation.

2.2 Stimuli

Participants were presented with 120 stimuli, consisting of 20 words with fricative consonant (/ʃ/, /ʃ/, /θ/, /θ/, /z/, /z/, /v/ or /ð/) in word-initial or word-final position (e.g. “sick” or “mouse” for /ʃ/). The stimuli were chosen to be minimal-triplets or -pairs (e.g. “she”, “see”, and “fee”; “fourth” and “force”) where each word had similar frequency of occurrence to the others, based on Collins Cobuild English Dictionary [7]. For each word, two contexts were created: the word in isolation and the word in the carrier sentence “He wrote ___. Then, for each of those stimuli, three types of speech signals were created: 16-bit WAV at 44.1 kHz, 16-bit WAV at 8.0 kHz, and 13-bit AMR-NB at 8.0 kHz band-pass filtered at 200-3400 Hz. In short, 20 words, with/without a carrier sentence, in 3 types of speech signal, makes a total of 120 stimuli. Of the 20 recorded words, 6 had an /t/ sound next to the fricative and these were not included in the analysis for the present study. The remaining 14 words can be seen in Table 1.

2.3 Apparatus

A personal computer (Windows XP Professional ver. 2002 sp 3), a handheld microphone (Sony F-V610) and a preamplifier (Audio-Technica AT-MA2) were used to record the original 40 speech samples – 20 words with and without a carrier sentence, sampled at 44.1 kHz – in Praat (ver. 5.1.43), free acoustic analysis software. Mobile Media Converter, a free audio and video converting software was used to downsample/convert each original speech sample to WAV 8 kHz and AMR 8 kHz. Sony Vegas Movie Studio 8 was used to create an audio file for the experiments. Praat was also used for acoustic analysis of the speech samples with a Praat script written by the first author to facilitate automatic acoustic measurements of the data. An iMac computer (Mac OS X 10.6), headset mic (Sennheiser PC 131) and Praat were used in the listening experiments. Since AMR-NB is used in both 2G (GSM) and 3G (UMTS/3GSM) mobile communication networks, which cover more than 80% of market share [8], we decided to use the AMR-NB filter to create the cell phone effect.

2.4 Procedure

2.4.1 Stimuli Preparation

The original 40 speech samples were recorded at a 44.1 kHz sampling rate. Each sample was downsampled/converted to WAV 8 kHz and AMR-NB 8 kHz. In the case of AMR-NB, the normally variable bitrate was set to a constant 12.2 kbps, the maximum possible. A 13-minute audio file for the experiments was generated from the 120 samples, in randomized order, with 2.5-second pauses between each stimulus giving time to an-

<table>
<thead>
<tr>
<th>Voiceless</th>
<th>/ʃ/</th>
<th>/ʃ/</th>
<th>/θ/</th>
<th>/θ/</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word Initial</td>
<td>sick</td>
<td>she</td>
<td>fee</td>
<td>thick</td>
</tr>
<tr>
<td>Word Final</td>
<td>mouse,</td>
<td>kiss</td>
<td>life</td>
<td>mouth</td>
</tr>
<tr>
<td>Voiced</td>
<td>/z/</td>
<td>/dʒ/</td>
<td>/v/</td>
<td>/ð/</td>
</tr>
<tr>
<td>Word Initial</td>
<td>chains,</td>
<td>zoo</td>
<td>jail</td>
<td>van</td>
</tr>
<tr>
<td>Word Final</td>
<td>kids</td>
<td>change</td>
<td>live</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Stimuli analysed in this study (without /t/ adjacent to the target fricatives)

<table>
<thead>
<tr>
<th>Initial 1</th>
<th>see</th>
<th>she</th>
<th>fee</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial 2</td>
<td>sail</td>
<td>fail</td>
<td>jail</td>
</tr>
<tr>
<td>Initial 3</td>
<td>sick</td>
<td>thick</td>
<td></td>
</tr>
<tr>
<td>Initial 4</td>
<td>shoe</td>
<td>zoo</td>
<td></td>
</tr>
<tr>
<td>Initial 5</td>
<td>fan</td>
<td>van</td>
<td>than</td>
</tr>
<tr>
<td>Final 1</td>
<td>kiss</td>
<td>kids</td>
<td></td>
</tr>
<tr>
<td>Final 2</td>
<td>lice</td>
<td>life</td>
<td>live</td>
</tr>
<tr>
<td>Final 3</td>
<td>mouse</td>
<td>mouth</td>
<td></td>
</tr>
<tr>
<td>Final 4</td>
<td>change</td>
<td>chains</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Nine occurring combinations of answer choices on the answer sheet (not including /t/ words)
swen, and an extra 3-second pause for each page turn. Instructions were created in Japanese and one example stimulus was given at the beginning (a word with no relation to the present study). The final file was exported as a 16-bit WAV file at 44.1 kHz.

2.4.2 Data Collection
The experiment was carried out twice, each time with a different group of participants, but in exactly the same environment. The listeners were told to identify the word they heard in each stimulus from given choices that consisted of minimal-pairs, -triplets, or -quadruplets (see Table 2), printed on an answer sheet. The audio file was played by the professor through the participants’ headphones at exactly the same time, so the experiment started/ended at exactly the same moment for every participant. It should be pointed out that in order to carry out this timing, the question order could not be changed between participants. However, the one order that was used for all participants was created by randomizing the order of the questions.

3 Results
The first result of the perception experiment can be seen in Figure 1.

When one compares the identification rate for the three different types of signals, it is apparent (and not surprising) that participants perform best when listening to the 44.1 kHz signal (“Norm”), regardless of the position of the fricative sound in the word (word-initial versus word-final). When considering the cell phone speech (“Cell”) though, it is interesting that it does not always act as a detriment to listeners’ fricative perception — it depends on the position of the fricative in the word. When the fricative is word-final, cell phone speech is the worst type of signal for identification, but when the fricative is word-initial, the effect of cell phone speech on the non-native listener is no different from the effect of simply downsampling speech (“Down”).

Figure 2 breaks down the results of the previous figure by showing results for each fricative sound. It is apparent that the ability to correctly identify the fricatives depends not only on the signal type (cell phone vs. downsampled vs. normal), but also on the type of fricative and where it is located in the word. Specifically, looking at the average across fricative types (the two rightmost bars in each graph), having the word in a carrier sentence (even though it is always the same unrelated sentence) helps identification of the fricative — especially for word-final fricatives. In addition, we found that voiceless sounds are more often misidentified than voiced sounds, in most contexts and across most signal types.

In the case of word-initial sounds (to compare with previous studies), the average rate of correct identification for normal speech was 66.67% for /s/, 55.56% for /ʃ/, and 50.00% for /θ/ (see Table 3). For cell phone speech, it was 52.08% for /s/, 32.64% for /ʃ/, and 54.86% for /θ/.

To determine what might be affecting the rate of perceptibility, we made measurements of the waveform of the acoustic signal. Specifically, we measured the duration and intensity of each fricative. Figure 3 shows four scatterplots: two for perceptibility versus duration, and two for perceptibility versus intensity. Word-final fricatives are plotted separately from word-initial fricatives, because the two tend to differ greatly in both duration and intensity. Using a one-tailed test of the Pearson Product-Moment Correlation Coefficient, results show a significant correlation ($\rho<0.05$) for word-initial perceptibility versus duration, and for word-final perceptibility versus intensity.

4 Discussion
An interesting point to be noticed is that the result shown in Figure 1 is contrary to what Takata & Nábělek [5] reported; they reported Japanese (and American) subjects made more errors in the word-final position than in the word-initial position; however, these results cannot be compared fairly because consonants used in the study by Takata & Nábělek [5] include not only fricatives but also others such as stops or plosives, which are understandable to be worse perceived in the word-final position because of the nature of pronunciation. The overall difference that was observed in Figure 1, between word-
initial and word-final position – the fact that word-final fricatives are more easily identified – is possibly due to the fact that the duration of the vowel that precedes the word-final fricative is often affected by the fricative, providing a cue for distinguishing word-final sounds. However, the difference between the results for different signal types within the word-final context is interesting. In word-final position, sounds are generally softer (less intense), and this can be confirmed by looking at the two lower scatterplots in Figure 3. So, it is possible that with the reduction in intensity in the word-final context, the further degraded cell phone signal was much different from the other signal types. In the word-initial context, however, the signal was loud enough that the perception of cell phone speech did not differ significantly from perception of downsampled speech.

<table>
<thead>
<tr>
<th>Stimuli</th>
<th>/s/</th>
<th>/f/</th>
<th>/θ/</th>
</tr>
</thead>
<tbody>
<tr>
<td>Miller &amp; Nicely</td>
<td>74.6%</td>
<td>75.4%</td>
<td>46.1%</td>
</tr>
<tr>
<td>Lambacher et al.</td>
<td>71.4%</td>
<td>76.2%</td>
<td>55.2%</td>
</tr>
<tr>
<td>Our Study</td>
<td>66.7%</td>
<td>55.4%</td>
<td>50.0%</td>
</tr>
</tbody>
</table>

Table 3: Comparison of perceptibility rates with past research

We found that having the stimulus word appear at the end of a carrier sentence helps perception of the fricative sound in that stimulus. This is probably explained by the fact that participants (especially non-native listeners) most likely need some warning that a word is coming up soon. When a target word for listening appears in isolation, it is possible that the time between the start of speech and the start of the fricative is too short, not giving the listener enough time to properly process the signal.
English voiceless fricatives (top) and perceptibility with word-initial position (left) and word-final position (right) for words in a sentence.

Figure 2 shows that, as expected, the average perceptibility rate of our non-native listeners was higher for normal 44.1 kHz speech than for cell phone speech, for word-initial /s/ as in “see” and /f/ as in “fee”. On the other hand, /θ/ as in “thick” (versus “sick”) was perceived equally poorly in both speech signals. In fact, the perceptibility of /θ/ was not significantly better than chance (50% — there were only two choices in the case of /θ/) in both types of speech signal. Perhaps the reason for this is that the Japanese language does not contain the sound /θ/, but it does contain /s/ and it contains a sound similar to /θ/ (namely, /ð/).

A comparison of the perceptibility rates in our experiment to those of Miller & Nicely [3] and Lambacher et al. [4] can be seen in Table 3.

For normal speech, which is what was studied in Lambacher et al. [4], our non-native listeners performed about 5% lower for /s/ and /θ/, but about 20% lower for /f/. The difference in results between the Lambacher et al. [4] study and ours shows that students can be fooled by hearing the target sounds in real words. Students have learned some words already in English and so they have a sound pattern memorized for those words. What our results here suggest is that the memorized pattern is often the wrong one (e.g., students have memorized the sounds for “mouse” as representing the sounds for “mouth”). It’s possible that if they just heard “ouse” versus “outh”, then they would do better at the perception test (because the words would not be tricking them).

In the case of cell phone speech perception, our non-native listeners only performed at chance for /f/ and /θ/, and about 20% lower than listeners in Lambacher et al. [4] for /s/. The results in Figure 3 are interesting in that they show that duration of the fricative sound is important for perceptibility in word-initial cases, but that intensity of the fricative sound is important for perceptibility in word-final cases. It seems that for fricatives in word-initial position, the intensity is consistently higher than for fricatives in word-final position, and so duration becomes the important point in word-initial position. On the other hand, in word-final position, duration is extremely variable and does not affect perceptibility. Instead, in word-final position, intensity is what matters.

One reason that duration is important factor for perceiving fricatives in word-initial position could be because listeners’ ears and brain are not ready to perceive sounds. Another possibility for the reason that intensity is important factor for perceiving fricatives in word-final position is because since the sound will get softer when it comes closer to the end of their word, listeners cannot perceive if the intensity is too low no matter how long the duration is.

In this case, we found that the first possibility was unsupported. The result in Figure 4 and Figure 5 break down the result in Figure 3 into two based on whether the word is said in isolation or in a sentence. By compar-
ing the graphs at top-left (perceptability versus duration, word-initial) in each figure, we can see that duration is more important when the word is said in a sentence. Notice that when the word is said in a sentence, listeners are already perceiving sounds before target sound reaches to their ears, which may make perception easier - even the target sound is in word-initial position.

5 Conclusions

In conclusion, we have shown that cell phone speech is often, but not always, a problem for non-native listeners’ perception of English fricatives. Factors that must be considered include the type of fricative (i.e., the actual phoneme), its position in the word, and whether or not the word is said in isolation or in a sentence.

With a clear understanding of the problems that non-native listeners face in the perception of cell phone speech, we hope that it may someday be possible to design cell phones to automatically detect fricatives and enhance the signal (through changes in intensity, duration, etc.) so that fricatives and other misperceived sounds may become more intelligible by non-native listeners. In the meantime, we have outlined the areas in which more listening practice is necessary to compensate for the degraded signal available when using cell phones.

The results of this speech perception study also have implications for speech production: Japanese learners of English pronunciation should focus on fricatives in word-initial position (especially /f/ and /θ/), and should practice saying them with longer duration and greater intensity.

Acknowledgements

First of all, I should like to express my deepest gratitude to Professor Ian Wilson, my supervisor, for his constant encouragement and guidance throughout my research work.

I am extremely grateful to 72 students at University of Aizu who willingly participated in the study and provided invaluable data. Without their contribution, this research would never have been accomplished.

References


