実践論文

Second Language Learners' Perception of Phonetics Reduction in Natural and Mechanical Speech

中田　ひとみ*1
Hitomi Nakata

Email: hitominakata5@gmail.com

英語の発音及びリスニング指導においては昨今「生きた英語を聴く」ことが奨励され、教育現場でもこれに応じて様々なメディアを通じたいわゆる生教材（BBC ニュース等）を素材としたものが導入されている。しかし、手加減のない英語には様々な減約（脱落、連結、同化現象などの音変化）を含むのが現実で、学習者の聴解力をはばむ原因となっている。本研究では、機械的に速度を圧縮した英語音声と、ほぼ同一の速めの速度で話す英語母語話者の聴取理解を比較検証した。実験を経て、少なくとも減約箇所においては、中級～中上級レベルの学習者であっても減約の少ない機械圧縮による音声の方を自然発話サンプル音声よりも多く聴き取るという結果を得た。英語学習者にとって細部の音現象を聴き取る困難が示される結果となったが、部分的に、学習者が持つ形態素や文法の知識を手掛かりにして聴取の補完を行ったと思われる箇所もみられる。これらの結果を踏まえ、将来的のリスニング教育においては、従来の音声生教材に加えて幅広い教授法導入の検証が必要となることが示唆された。

This study addresses the question of whether intermediate to upper-intermediate university students, aged 19–21 years and majoring in English various courses, demonstrate better listening performance when faced with naturally fast speech rather than with mechanically compressed fast speech. The results show that the EFL learners in Japan were able to hear the mechanically compressed fast speech better than naturally spoken English. The results also suggest that it is difficult to detect reduced sounds in naturally connected speech delivered by native speakers, even for upper-intermediate level learners. It was often observed from their test results, however, that the learners occasionally reconstructed their misheard words by relying on their morphological or grammatical knowledge. Therefore, this paper suggests that acquiring other language skills – such as grammar and vocabulary – as well as conventional tasks would be beneficial for the improvement of EFL learning ability to detect phonetic reduction when listening to a second language.

キーワード：第二言語のリスニング、音声的減約、自然言語、機械圧縮音声
Keywords ：L2 listening, phonetic reduction, connected speech, mechanical speech

*1：獨協大学　外国語学部　英語学科
非常勤講師

1. Introduction

In teaching the second language (hereafter, L2) English listening and pronunciation skills, teachers tend to believe that they should provide learners with some naturally spoken English such as TV commercials, weather reports on the radio, and other authentic-sounding speech. Even a pre-recorded read-aloud speech for textbook is not acoustically processed but reflects a connected speech including assimilations, linking, and/or deletions.

However, perceiving and listening to natural English has been a big challenge for most learners in Japan, as the connected speech always involves some reduced sounds or coarticulation such as elision and linking (or liaisons) (e.g., Gilbert[1]). Shockey and Bond[2] argue that native speakers can compensate for the reduced (or, changed) segments using their phonological knowledge and/or lexical inventory. But this strategy cannot be used by L2 learners (hereafter, L2ers). Non-native speakers have difficulties in decoding the change and take such phonetic events at their face value, which leads to obstacles in improving their listening skills.

This paper suggests that this difficulty arises because the L2ers most likely learn L2 vocabulary in isolation starting from their elementary level studies and this learning habit continues to their intermediate and upper-intermediate stages. This conventional learning style gives them the misleading idea that English words are always pronounced clearly with a pause between them. This study examines this gap between naturally spoken fast speech with speech reductions and mechanically modified fast speech without the reductions.

2. Methods

2.1 Participants

Twenty-five students aged 19–21 years enrolled in an English Listening and Pronunciation class at a Japanese university participated in the tests. The proficiency in English of the participants can be categorized as intermediate to upper-intermediate level, as their TOEIC® (Test of English for International Communication) scores range from 550 to 875, averaging approximately 706.

2.2 Procedures

The participants were instructed to transcribe some words from two sets of recordings by filling blank spaces on an answer sheet. Each set consisted of recordings in two different styles; the first one was naturally spoken by a male native speaker of English from Canada, and the other one consisted of mechanically compressed fast speech that was adopted from an e-learning online resource: *NetAcademy 2 Super Standard Course produced by ALC Education*[3].

The author recorded the fastest-version speech of two units (from the intermediate level of the NetAcademy) with an IC recorder connected to the website, and subsequently asked the native speaker to read the same text aloud in almost the same tempo as the NetAcademy version. For the first unit a news report on African penguins (hereafter, U1), the natural speech was 10.5% faster, but for the second unit, a news report of an event on Martin Luther King, Jr. Day (hereafter, U2), the natural speech was 12.4% slower than the mechanical speech. These differences of speech tempo were further examined for the number of stressed syllables. Dauer[4], referred to in Patel[5], reports that preferred musical beats occur *roughly* every 500–700 ms, and that frequency in the occurrence of a beat is similar to that of stressed syllables in language. In this study, the average durations between stressed syllables are 487 ms for the natural and 544 ms for the mechanical speech recording of U1, and 521 ms for the natural and 463 ms for the mechanical speech recording of U2.

The participants listened to the sets of speech sentence by sentence twice. They wrote down targeted words in blank spaces after listening to a sentence and confirmed what they wrote during the second time. They also filled in a questionnaire sheet to describe their English background including the TOEIC® scores.

3. Materials

Three items of reduced speech (i.e., speech that contains instances of phonetic reduction) were set out as targeted components based on Roach[6]: assimilation (hereafter, Assimilation), linking (hereafter, Linking), and phrasal words where two reductions are mixed (hereafter, Mix) such as 'one of a'. As for types of assimilation, we adopted the following criteria by Jones[7]: 'Assimilation is a process found in all languages which causes speech sounds to be modified in a way which makes them more similar to their neighbours'. U1 speech texts (transcripts titled 'African Penguins') consists of five carrier sentences with 12 items, and U2 speech texts (transcripts titled 'Martin Luther King Jr. Day') consists of six carrier sentences with nine items. Table 1 and Table 2 show how the targeted words, bolded and underlined, were extracted in sentences from U1 and U2, respectively. Each sentence is numbered (1)–(5) for U1 and (1)–(6) for U2. The abbreviation 'A' stands for Assimilation, 'L' for Linking, and 'M' for Mix phrases. Table 3 exhibits a combined list of items in U1 and U2, aligned from the group of A, L and M.

Table 1 Targeted reduction items in U1

Transcripts of speech: (1) *A South African penguin called Peter is back home after swimming 800 kilometers to get there.* (2) *Peter and thousands of other birds were rescued from an oil spill in June. They were taken from Robben and Dassen Islands off the Western Cape to Port Elizabeth.* (3) *After the spill was cleaned up, the birds had to find their way home through shark-filled waters.* (4) *Peter was one of a few birds wearing a special radio device to show his whereabouts.* (5) *South African wildlife officials say they heard his signal Tuesday coming from somewhere on Robben Island.*						
A, L, or M	A1	L1	A2	A3	L2	M1
Target words	called Peter	Peter is	back home	get there	thousands of	from an oil
A, L, or M	L3	A4	A5	A6	M2	A7
Target words	cleaned up	had to	find their	filled waters	one of a	heard his

Table 2 Targeted reduction items in U2

Transcripts of Speech: (1) *The United States is observing the birthday of former civil rights leader, Martin Luther King Jr.* (2) *Reverend King's wife, Coretta, spoke during ceremonies in Atlanta, Georgia.* (3) *She called for an end to hostility in American politics. (4) Reverend Martin Luther King, Jr. was a major leader of the movement to gain civil rights for African-Americans.* (5) *The Nobel Peace Prize winner used nonviolent protest to force social change.* (6) *Reverent King would have been 75 years old this year. He was murdered in 1968.*									
A, L, or M	A8	L4	M3	A9	A10	A11	A12	M4	A13
Target words	of former	ceremonies in	for an end	movement to	rights for	used non-violent	protest to	would have been	this year

Table 3 Targeted reduction items in U1 and U2 grouped in each target

A	A1	A2	A3	A4	A5	A6	A7
	called Peter	back home	get there	had to	find their	filled waters	heard his
	A8	A9	A10	A11	A12	A13	
	of former	movement to	rights for	used non- violent	protest to	this year	

L	L1	L2	L3	L4	M	M1	M2	M3	M4
	Peter is	thousands of	cleaned up	ceremonies in		from an oil	one of a	for an end	would have been

There are 13 Assimilations especially targeting word-final alveolars (e.g., mostly /t/ and /d/) adjacent to word-initial consonants of the following words, in which the alveolar consonant is not likely articulated. Items for Linking are extracted from four places where word-final consonants (/r/, /ʤ/, /d/, and /z/) are linked with the initial vowel of the following preposition as if there is only one word such as seen in 'Peter is' (pronounced as /pi(ː)təɹɪz/). And lastly, four Mix phrases were targeted: 'from an oil', 'one of a', and 'for an end' from U1, and 'would have been' from U2, where most of the reductions take place due to linking effects for all four of these items.

Altogether 21 tokens were extracted from 11 sentences. The participants gained one point if they wrote the correct word including the assimilated consonant (e.g., 'called'), but obtained 0.5 points if they wrote down only one of two consecutive linked words. Mix phrases all consist of three words, so the participants gained 0.3 points if they wrote only one word, 0.6 points for two consecutive words, and 1.0 for all three.

Figure 1 and Figure 2 show the acoustic differences of an M1 example: 'from an oil spill in June' in U1 between the natural speech and the mechanical speech, respectively.

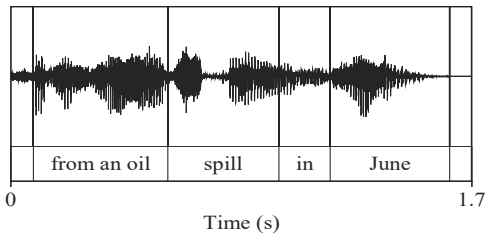

Figure 1. Sound waveforms extracted from M1 example: 'from an oil' (spill in June) of the natural speech

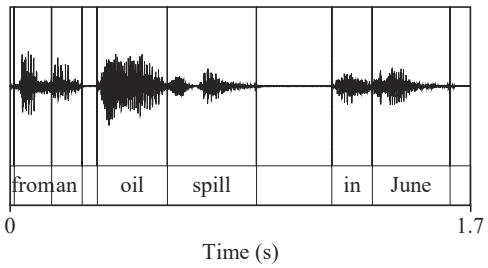

Figure 2. Sound waveforms extracted from M1 example: 'from an oil' (spill in June) of the mechanical speech

As can be seen from above, there is no obvious pause between words in the natural speech, whereas in the mechanical speech, there is a clear boundary between 'an' and 'oil' with a silent part. This distinction can be further observed between 'spill' and 'in' between two speech styles; the phonation continues in the natural speech but there are no such sound traits in the mechanical speech.

4. Results

4.1 Performance in All Items

Table 4 exhibits the sum scores of all the participants (×25) for each item. The first line of each group refers to the scores for the natural speech version, the second line for the mechanical speech version, followed by the difference ('Natural' minus 'Mechanical', marked as 'N-M"). The asterisk at the bottom line indicates items of where 'Natural' obtained higher scores than 'Mechanical', marked as N > M.

Table 4 Scores for all items in comparison between natural and mechanical speech tasks

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13
Natural	19	20	23	2	9	9	15	22	22	14	19	16	24
Mechanical	22	24	24	13	12	8	20	22	19	18	17	21	22
N – M	-3	-4	-1	-11	-3	1	-5	0	3	-4	2	-5	2
N > M					.	*			*		*		*

	L1	L2	L3	L4	M1	M2	M3	M4
Natural	11	12	22	4.5	15	18	7.2	6.2
Mechanical	23	22	20	6	21	22	2.7	6.2
N – M	-12	-10	2	-1.5	-6	-4	4.5	0
N > M			*				*	

As shown in the table, a number of items in 'Mechanical' speech' exceeded that in 'Natural speech'. There are only six items that accumulated higher scores in 'Natural' than 'Mechanical'; they are L3 ('cleaned up'), A6 ('filled water'), A9 ('movement to'), and A13 ('this year'). A two tailed test confirmed that the variances between the two different types of speech were statistically significant (F [0, 46] = -2.47, p = .002 < .05). The next section will present a detailed breakdown of the items for each type of reduction.

4.2 Performances in Reduction Types

Figure 3 exhibits the results of the learners' performance on dictation, namely, of how well they could catch the assimilated and reduced sounds. The figures indicate the number of participants who wrote the correct word (total number = 25), and A1–A13 stand for item code numbers.

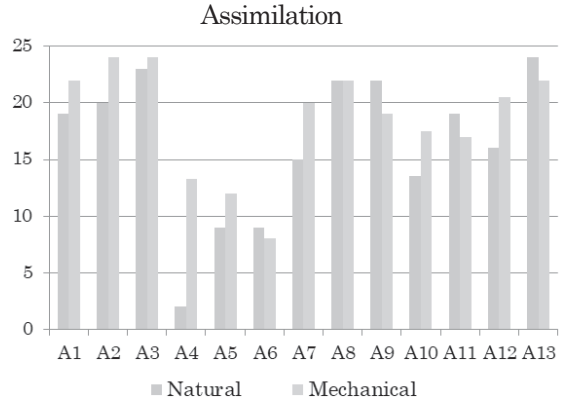

Figure 3. The number of participants (nos. = 25) who wrote Assimilation items correctly both in the natural and mechanical speech tasks

As can be seen from the data above, the learners could detect assimilated items better in the mechanical speech than in the natural speech. Exceptions are found in four items: A6 'filled water',

A9 'movement to', A11 'used nonviolent, and A13 'this year'. A big difference between the natural speech and the mechanical speech tasks is observed for A4 'had to' in which the rate of the detection of the alveolar consonant /d/ of 'had' is extremely low in the natural speech task (8%), but in the mechanical speech task the ratio raises up to 53%. There were only two participants who obtained the score for this part in the natural speech task, but 12 participants revised their answer to the correct one 'had' in the mechanical speech task. In the mechanical speech, word-final consonants (mostly alveolar stops /t/ and /d/) were *not* likely assimilated but each word was pronounced in isolation, so word-final alveolars were always followed by a pause.

Two figures below compare the waveforms of this item 'had to' in different speech styles. Both waveforms show the assimilated parts of /d/ of 'had' and /t/ of 'to' in the connected (natural) speech. For annotations, the tier above refers to the phonetic segments and the tier below shows the words: 'had' and 'to'.

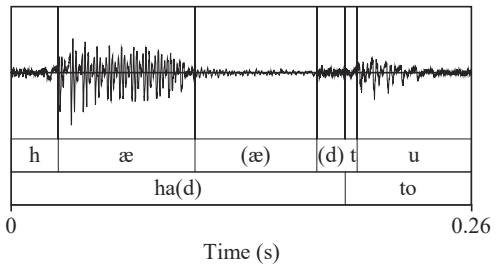

Figure 4. Sound waveforms of 'had to' in the natural speech.

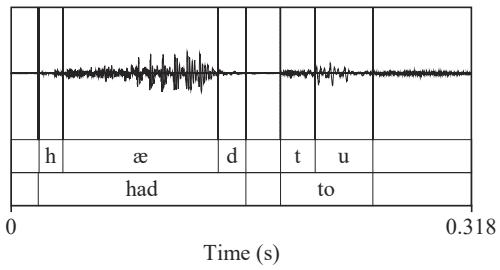

Figure 5. Sound waveforms of 'had to' in the mechanical speech

As can be seen in Figure 4, the boundary between two consonants /d/ and /t/ in the connected speech seems ambiguous. In Figure 5 of the mechanical speech, however, there is a clear boundary and even a pause between the consonants. Even in that artificial-sounding speech, the learners showed their ability to detect L2 words.

Similarly, Figure 6 and Figure 7 refer to results of Linking items and Mix items, respectively.

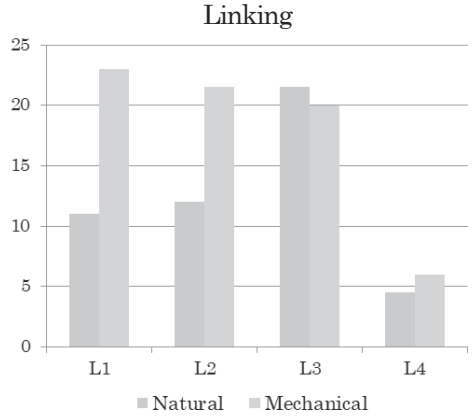

Figure 6. The number of participants who wrote Linking items correctly

The correct number of linking parts is shown in Figure 6, where, again, the number of participants writing the correct words was higher in the mechanical speech task than in the natural one, except for the case of L3 'cleaned up'. Compared to the other three (L1–3), the score for L4 'ceremonies in' was quite low: 18% in the natural speech and 24% in the mechanical speech, whereas it was 44% – 93% in both speeches in L1–3.

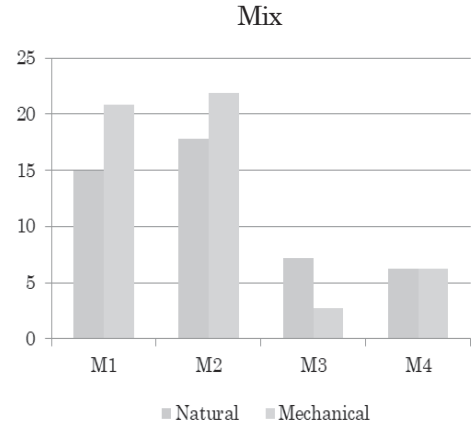

Figure 7. The number of participants who wrote Mix items correctly

Figure 7 likewise exhibits the number of Mix items. It is immediately noticeable that M1 'from an oil' and M2 'one of a' attained much higher scores (especially in the mechanical speech) than M3 'for an end' and M4 'would have been'. It is not clear if the differences are caused by a contextual reason, because there is no specific (and linguistic) ground to explain the particularities among the four items. The only difference is the speech style. Namely, the speaker for the natural speech is the same, but for the mechanical speech, two different speakers were employed. The announcer for the U1 has a quite standard pitch range for a male, but the voice of the other announcer, in U2, is quite low. Actually, the score rates of L4 in U2 were also quite low regardless

of the speech type.

5. Discussion

Based on these observations, we can suggest the following generalizations: 1) L2ers of English, even at the upper-intermediate level, still have difficulties compensating for the reductions that occur in natural speech (i.e., they take the sounds they hear at face value and could perceive words and phrases from compressed machine-talk in which reduction is limited (Roach[6])); 2) learners probably use non-phonological strategies to decipher what they hear; for instance, they attempt to recover auditory information using grammatical and/or other linguistic knowledge in accordance with the context; 3) it is possible that a morphological difference between English and Japanese may affect the perception of reduced words or phrases in fast speech. The ratio of detecting the linking sounds /ζIv/ at the end of 'ceremonies in' was quite low possibly due to the scarcity of using the plural form of "ceremony" in Japanese. Japanese speakers would say /σερεμOνιι/ for both singular and plural meanings. Therefore, we speculate that the learners noticed the word "ceremony" but failed to perceive the word-final /ζ/ sound in the English plural form. This application of L1 structure was also reported in Nakata and Shockey[8], in which Japanese learners of English often inserted an epenthetic /O/ between /O:στ(≅)/ and /ρεIλφ≅/ of /O:στ≅ρεIλφ≅/ ('Australia') in their pronunciation experiments.

6. Conclusion

This study addressed the question of whether L2ers of English can detect words from phonetically reduced and non-reduced speech. We employed two types of speech: naturally spoken fast speech and artificially modified fast speech. Detection tasks were assigned to intermediate and upper-intermediate learners. They performed better when listening to non-natural speech with limited reductions than naturally spoken speech with several reduced items.

The results lend support to Shockey and Bond's[2] claim that native speakers have access to phonological and lexical knowledge, which allows them to compensate for the missing acoustical information; in contrast, L2ers have only limited access to this domain. The results also suggest that even upper-intermediate level learners still rely on segmental cues to detect missing information, no matter how unnatural the speech is. This can be a dilemma for L2 listening classes in which most materials represent connected and natural speech, and suggest the necessity of providing learners with schematic instruction of phonetic reduction.

In addition, we obtained evidence that the participants probably used their grammatical and/or other knowledge to produce the correct words, in an effort to match the context. Therefore, it would be possible to further improve learners' listening skills using materials that provide information from other linguistic areas such as grammar and morphology; for instance, learning the different roles of content words and function words would be effective.

In future studies, it would be beneficial to vary the acoustic properties of speech materials. It would be interesting to conduct tasks with speech input similar to the one used in this study, but with some modifications, such as era
sing the reduced sounds completely. Furthermore, the pitch range of the speaker's voice would be a factor that affects listeners' performance. The effect of the speaker of U2 in the mechanical speech in this study suggested that the scores tended to be lower if the input was in an unfamiliar voice. Further research is needed to analyze this dimension.

References

(1) Gilbert, J. B., "*Clear Speech: Pronunciation and Listening Comprehension in North American English – Teacher's Resource and Assessment Book (4th ed.)*", New York: CUP (2014)

(2) Shockey, L., Bond, Z. S., "Slips of the ear demonstrate phonology in action", *Proceedings of 16th ICPhS* Saarbrücken, 1385–1388 (2007)

(3) *NetAcademy 2 Super Standard Course* (a website provided by ALC Education). http://www.alc-education.co.jp/academic/net/course-e/super.html?aid=course (2015)

(4) Dauer, R., "Stress timing and syllable timing reanalyzed", *Journal of Phonetics*., Vol..11, 51–62 (1983)

(5) Patel, A. D., "*Music, Language, and the Brain*", New York: Oxford University Press (2008)

(6) Roach, P., "*English Phonetics and Phonology – A practical course (2nd ed.)*", Cambridge University Press (1991)

(7) Jones, D., "*English Pronouncing Dictionary (17th ed.)*", Roach, P., Hartman, P., Setter, J. (eds), Cambridge University Press (2006)

(8) Nakata, H., Shockey, L., "*The effect of Singing on Improving Syllabic Pronunciation – Vowel Epenthesis in Japanese*", Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS). 1442-1445 (2011)

[1] An earlier version of this study was presented at the 55th National Conference of The Japan Association for Language Education and Technology, which was held in Osaka, Japan on August 5, 2015. The proceedings presented the methods and results of this study, although the analysis and discussions were still in development. This paper includes some additions, such as figures of acoustic waveforms to visualize sound tokens and more supportive examples in the discussion section.