COMPUTER-AIDED HIGH VARIABILITY PHONETIC TRAINING TO IMPROVE ROBUSTNESS OF LEARNERS' LISTENING COMPREHENSION

Haoyu Zhang[†], Yusuke Inoue[†], Daisuke Saito[†], Nobuaki Minematsu[†], Yutaka Yamauchi[‡]

[†]School of Engineering, The University of Tokyo, [‡]School of Education, Soka University
{haoyuzhang,inoue0124,dsk_saito,mine}@gavo.t.u-tokyo.ac.jp, yutaka@soka.ac.jp

ABSTRACT

Speech acoustics are easily degraded due to extra-linguistic and environmental factors. This degradation is often troublesome to non-native listeners. To improve their robustness in listening, high variability phonetic training (HVPT) is widely used. In this paper, technically-enhanced HVPT is tested and new types of acoustic variability, namely, vocal tract length and channel distortion, are examined. Listening tests using the degraded speech materials are carried out with Japanese learners of English and it is found that their listening abilities are particularly impaired when radio communication distortion is added. An 18-day listening drill for speech samples with different levels of radio distortion is imposed on the English learners. Post-tests show that robustness of listening comprehension is significantly improved in the case of advanced learners, and that improvement is transferred effectively to listening to materials with other distortions.

Keywords: L2 learning, HVPT, speech modification, robustness, listening comprehension

1. INTRODUCTION

Acoustic variability in speech draws attention of L2 researchers, as High Variability Phonetic Training (HVPT) has been shown to be effective for improving listening ability [7, 15, 4]. They claim that stimulus variability is the core of HVPT. Acoustically degraded speech materials were used in L2 studies to compare listening ability between learners and natives [12, 6, 1, 14]. In [2], learners' listening ability was tested with speech samples with babble noise, and in [5], it was also examined between the two cases where the babble noise was L1 and L2. The performance of phonological perception in the presence of reverberation and background noise was investigated in [8, 9, 10, 11]. As claimed in [15, 4], stimulus variability in HVPT is important but these studies prepared stimulus variability manually or with simple signal processing techniques. In this paper, more technically-enhanced HVPT is introduced to enlarge the acoustic variability.

In animation movies, to prepare the voices of very tall or small characters like giants or fairies, voice changers are sometimes applied to the original voices of actors or actresses. Here, frequency-axis warping techniques are used for voice morphing. In this paper, as one kind of new acoustic variability, a vocal tract length (VTL) changer is introduced as a novel kind of acoustic variability.

Learners often claim that they can listen well in a face-to-face situation but have some troubles when talking on the telephone. This means that channel-based distortions easily cause communication troubles. High distortions are found on radio communication channels such as air traffic control, police radio, taxi radio, etc. For safety reasons, learners should acquire good listening skills for voices on a radio communication channel, e.g., emergency alerts. In this study, as another kind of new acoustic variability, radio distortions are introduced.

In this paper, listening comprehension tests with the above distortions clearly show learners' weakness and native speakers' robustness with respect to radio distortions. After that, an 18-day HVPT-based listening drill is imposed on the learners and it is shown that robustness is improved effectively and transferred to listening to speech samples with other kinds of distortions and original speech samples.

2. TECHNICAL IMPLEMENTATIONS OF THE NEW ACOUSTIC VARIABILITIES

2.1. Vocal tract length (VTL) manipulation

VTL manipulation can be represented in the form of linear transform in the cepstrum domain [13]. In this paper, generation of giant voices or fairy voices is realized by applying those linear transformations to the original voices. Further, pitch manipulation is also done automatically according to the resulting VTL. Examples are embedded here in the PDF file.

2.2. Simulation of radio communication distortion

After applying low-pass filtering with 8kHz cut-off, the amplitude of each sample is multiplied by a

fixed value, and the samples above a fixed threshold are clipped. The processed speech signals are frequency-modulated at 100 kHz. After adding two kinds of noises to the modulated speech signals, which are speech signals modulated at 110 kHz and noise signals modulated at 100 kHz, the resulting signals are demodulated at 100 kHz. Examples are embedded here in the PDF file.

2.3. Combination of various types of distortions

Combination of different distortions with quantitative control is straightforward technically. We built a tool for acoustic morphing or degradation of input speech, which combines four factors: a) waveform addition (background noise), b) convolution (reverberation), c) VTL manipulation, and d) communication channel distortion. For each factor, the degree of degradation can be controlled quantitatively. Examples are embedded here in the PDF file.

3. LISTENING COMPREHENSION TESTS USING DISTORTED SPEECH

Listening tests using distorted speech were carried out with Japanese college learners. After the initial tests, special listening drills were imposed to them, after which they took the same listening tests again.

In the listening tests, a listening comprehension task was adopted, which is different from the task adopted in [7, 15, 4, 8, 9, 10, 11]. In Japan, there is a well-known English proficiency test, called EIKEN [3], of which Grade 2 tests are designed for high school students. Listening questions from EIKEN G2 were used as original questions in this paper. The listening part of EIKEN G2 has 2 subparts, questions using dialogues (Part A) and those using monologues (Part B). Test-takers listen to a short dialogue between a male speaker and a female speaker (Part A) or a short monologue (Part B). Then, a four-choice question is given. The chance level is 25% and Part B is more difficult than Part A.

Our difficult listening test was designed using EIKEN G2 listening tests. A test was composed of 16 dialogue questions (A) and 16 monologue questions (B). To prepare distorted dialogues and monologues, three kinds of distortions were applied.
1) Male and female voices were converted to giant and fairy voices, respectively, referred to as GF.
2) Any original voice was converted to the voice in air traffic control, referred to as ATC henceforth.
3) Combination of 1) and 2). Male/female voices were converted to those of a giant/fairy pilot, called as GF+ATC. An example of GF+ATC, which was used in our test, is embedded here in the PDF file.

In addition to the above distorted voices, the original voices were also used. In total, four types of voices were prepared. In the 16 dialogue questions, four questions were assigned to each type of voice. The monologue questions were prepared similarly. The order of presentation was random and a session of all the 32 questions took about half an hour.

4. THE PRE-TEST (JULY 2017)

4.1. Participants and experimental setup

125 Japanese college learners of English and 2 native speakers participated in the test. The learners took the test in a classroom, where questions were presented with two loud speakers. The native speakers took this test in their private rooms using the built-in speakers of their laptop PCs.

A test was composed of 16 dialogue-based questions and 16 monologue-based questions, and four tests were prepared. The native speakers took all the four tests but each learner took only two tests and each test was taken by approximately 30 learners. Since all the questions are from official and old EIKEN tests, there is no or ignorable difference in the difficulty level among the four tests.

4.2. Accuracy rates of learners and natives

The accuracy rates are compared between the learners and the native speakers, shown in Table 1. The rates are calculated for each part (A or B) and each type of voices. ANOVA shows that, in the learners' rates, between any two types, a significant difference is observed ($p<0.001$). The learners' performances are decreased in the order of Original, GF, ATC, and GF+ATC. By comparing the learners' rates of GF and ATC, the latter distortion makes utterances much more difficult to understand. Furthermore, the rates for GF+ATC are almost chance-level. It is very surprising that the native speakers can answer in ATC completely and in GF+ATC rather correctly. The results in ATC and GF+ATC exhibit a striking contrast of learners' weakness and native speakers' robustness. In the phoneme identification task under adverse conditions of background noise and reverberation [8, 9, 10, 11], such huge contrasts were not observed. This is probably due to differences of the task and the type of distortion adopted.

4.3. Analysis based on the learners' TOEIC scores

Many of the students had taken the TOEIC test prior to our tests and they are divided into three groups according to their scores, 400-600, 600-800, and 800-

Table 1: Accuracy rates of learners and natives for the four types of listening tests [%]

Part	Subject	N	Orig.	GF	ATC	GF+ATC
A	learners	125	68.9	59.2	34.2	25.9
	native	2	100	100	100	93.8
B	learners	125	55.6	46.1	31.4	25.4
	native	2	100	100	100	87.5

Table 2: Accuracy rates of each TOEIC-based group of learners for the four tests [%]

Part	TOEIC	N	Orig.	GF	ATC	GF+ATC
A	400–600	28	58.3	50.0	30.6	32.8
	600–800	52	78.2	62.0	35.1	23.4
	800–990	15	81.5	79.6	45.4	25.0
	native	2	100	100	100	93.8
B	400–600	28	42.2	41.1	23.9	25.0
	600–800	52	63.0	48.9	31.7	25.8
	800–990	15	74.1	67.6	41.7	24.1
	native	2	100	100	100	87.5

990, roughly corresponding to beginners' level, intermediate level, and advanced level. The accuracy rates are shown again in Table 2 but separately for each level. Even for advanced learners, a by far bigger drop is found at ATC compared to the drop at GF. Their performance in GF+ATC is chance-level.

Radio communication is used in such situations where extremely high reliability is required, e.g., air traffic control and police radio. From this fact as well as the above experimental facts, the authors can speculate easily that this type of acoustic distortion causes almost no trouble to native speakers but that it causes extraordinary troubles even to advanced learners. A critical difference in listening strategies might be found between native speakers and learners, and the authors can hypothesize reasonably that radio communication distortion may be a good tool for learners to acquire a new strategy of listening.

To verify this hypothesis, we planned to give the same listening test again after some training with distorted materials, but we had to wait for such a long time that the learners would not remember what were asked in the pre-test. If they remembered, they would select the same answers, which were often incorrect. The pre-test was held in July 2017 and we waited for four and a half months, during which we did not inform anything on the post-test. After an 18-day listening drill was conducted for the learners, the same test was carried out again to them.

5. DESIGN OF THE LISTENING DRILL

5.1. An 18-day Listening drill

Five dialogue questions and five monologue questions were prepared per day. A total of 180 questions were prepared for 18 days. As in Section 3, these questions were from the official EIKEN G2 tests but different from the questions used in Section 4.

To reduce the difficulty level of materials with ATC distortion, we prepared 4 levels of ATC distortion. Level 0 is telephone-quality speech with no ATC distortion. The distortions were added to level 0 samples to produce samples of levels 1, 2, and 3. The degree of distortion is increased in this order. The listening questions with ATC distortion in our difficult listening test correspond to level 3.

To each of the 180 original oral questions, four kinds of ATC distortions were added to generate 720 listening materials in total. These listening materials were provided on a web, where the correct answers and the transcriptions of all the materials were also made available. It should be noted that, in this listening drill, distortions of GF and GF+ATC were not used at all, but used in the post-test.

5.2. Procedure of listening in the drill

Without any prior instruction, the learners would use the prepared listening materials in their own ways. To avoid this and control their learning behaviors, the authors gave the following instructions.

You have five new questions everyday. You should start with level 3 of question 1. If you do not understand what is said, repeat listening to level 3 of question 1 up to three times. If you still do not understand, use level 2 of question 1 and listen to it up to three times. After that, you may use levels 1 and 0. This is the end of question 1 and go to question 2.

6. THE POST-TEST (DECEMBER 2017)

6.1. Participants of the post-test

Out of the 125 students who participated in the pre-test, 63 students underwent the same test again in the same environment (post-test). Due to time constraints imposed by the college curriculum, they took only a half number of questions of the pre-test. There were 16 dialogue questions and 16 monologue questions in the post-test.

6.2. Effectiveness of ATC-based HVPT

55 out of the 63 students had taken the TOEIC test and the results of these 55 students in the pre-test are shown separately for each proficiency level in Table 3. The score distribution in Table 2 is similar to that of Table 3. Table 4 shows the 55 students' results of the post-test, which was held a week after the 18-day listening drill. Differences between Table 3

Table 3: Results of the first test of the 55 learners

Part	TOEIC	N	Orig.	GF	ATC	GF+ATC
A	400–600	15	66.7	48.3	25.0	41.7
	600–800	32	77.3	65.6	38.3	25.8
	800–990	8	84.4	84.4	43.8	21.9
B	400–600	15	50.0	43.3	28.3	23.3
	600–800	32	65.6	48.4	39.1	30.5
	800–990	8	78.1	62.5	37.5	28.1

Table 4: Results of the second test of the 55 learners

Part	TOEIC	N	Orig.	GF	ATC	GF+ATC
A	400–600	15	70.0	66.7	26.7	35.0
	600–800	32	73.4	73.4	40.6	32.8
	800–990	8	96.9	96.9	75.0	40.6
B	400–600	15	66.7	48.3	38.3	23.3
	600–800	32	61.7	51.6	42.2	35.2
	800–990	8	87.5	84.4	62.5	31.3

Table 5: Error reduction rate (ERR)
Larger ERR values than 40 are shown in bold.

Part	TOEIC	N	Orig.	GF	ATC	GF+ATC
A	400–600	15	9.9	35.6	2.3	-11.5
	600–800	32	-17.2	22.7	3.7	9.4
	800–990	8	**80.1**	**80.1**	**55.5**	23.9
B	400–600	15	33.4	8.8	13.9	0.0
	600–800	32	-11.3	6.2	5.1	6.8
	800–990	8	**42.9**	**58.4**	**40.0**	4.5

and Table 4, which indicate directly effectiveness of ATC-based HVPT, are quantified relatively as error reduction rate (ERR) in Table 5. It is defined as

$$\text{ERR} = \frac{\text{ER of the pre-test} - \text{ER of the post-test}}{\text{ER of the pre-test}},$$

where ER is error rate (=100−accuracy rate). In [15], /a/-/ae/ identification test was adopted with HVPT and ERR was reported to be about 40%. In Table 5, larger ERRs than 40 are shown in bold.

Since all the materials used in the 18-day listening drill were ATC-based distorted materials, we firstly focus on ERR in the case of ATC. Irrespective of proficiency level, ERR is always positive, which means ER (error rate) is reduced after the listening drill. However, effectiveness is much larger for advanced learners. With them, about half of errors were corrected by the listening drill.

Next, we focus on the results of GF. Also in this case, ERR is always positive and it is surprising that the values of ERR in GF are higher than those in ATC. We can say that robust listening skills acquired through the listening drill with ATC-based distortion are transferred and exploited when listening to differently distorted speech. However, transfer of robustness is not always effectively made. The ERRs are very small in Part B of beginning learners and intermediate learners. We have to admit that stable and good transfer of robust listening is found only in the case of advanced learners.

It seems to be the case with Original questions. Large ERRs are only found again in the case of advanced learners. Even for intermediate learners, the ERRs are negative for unknown reasons. From these results, it can be said the ATC-based listening drill is generally effective but highly effective only for advanced learners, and effective transfer of robust listening is also found only for them. This is probably because listening to speech materials in ATC may require well-integrated knowledge of English (phonology, syntax, semantics, pragmatics, etc). Or only advanced learners could keep motivated during the listening drill. The authors wonder whether similar effects can be observed in the case of non-advanced learners when we introduce much milder ATC distortions or untested types of distortion.

Although promising results are obtained in the experiment, to guarantee this effectiveness, the authors have to answer several questions in the future: 1) whether learners with the same amount of listening drills composed only of original EIKEN G2 tests will not show large robustness improvement and 2) whether advanced learners can keep high robustness, once they acquired, with no more difficult drills.

7. CONCLUSIONS

With advanced speech modification technology, a difficult listening test was designed and conducted with three types of novel acoustic distortions. Results showed learners' weakness and natives' robustness with respect to ATC-based distortions. Four and a half months afterwards, a listening drill for ATC-distorted materials was made and the drill was found to be very effective to improve the robustness of listening in the case of advanced learners. However, the authors do not claim that the ATC-distorted speech is the best speech to enhance listening robustness. It is also true that the distorted speech samples used in this paper seemed too difficult for non-advanced learners. Further, we can say that although the performance of advanced learners was improved, their performance in Original is just comparable to that of native speakers in GF+ATC. This means that a huge gap of listening performance still exists between learners and native speakers. As well as the future work listed in the previous section, we are interested in measuring native speakers' listening comprehension in a classroom environment and in investigating strategic differences of listening between native speakers and learners.

This work was supported by MEXT KAKENHI JP26118002 and JSPS KAKENHI JP26240022.

8. REFERENCES

[1] Broersma, M., Scharenborg, O. 2010. Native and non-native listeners' perception of english consonants in different types of noise. *Speech Communication* 52(11), 980–995.

[2] Cutler, A., Weber, A., Smits, R., Cooper, N. 2004. Patterns of english phoneme confusions by native and non-native listeners. *J. Acoust. Soc. Am.* 116, 3668–3678.

[3] Eiken tests. http://www.eiken.or.jp/eiken/en/grades/.

[4] Hwang, H., Lee, H. Y. 2015. The effect of high variability phonetic training on the production of english vowels and consonants. *Proc. ICPhS* P1.50.

[5] Lecumberri, M. L. G., Cooke, M. 2006. Effect of masker type on native and non-native consonant perception in noise. *J. Acoust. Soc. Am.* 119, 2445–2454.

[6] Lecumberri, M. L. G., Cooke, M., Cutler, A. 2010. Non-native speech perception in adverse conditions: A review. *Speech Communication* 52(11), 864–886.

[7] Lively, S. E., Longan, J. S., Pisoni, D. B. 1993. Training japanese listeners to identify english /r/ and /l/: the role of phonetic environment and talker variability in learning new perceptual categories. *J. Acoust. Soc. Am.* 94, 1242–1255.

[8] Masuda, H., Arai, T. 2011. Perception of english voiceless fricatives by japanese and english native listeners under various signal-to-noise ratios. *Proc. Spring Meet. Acoust. Soc. Jpn.* 471–474.

[9] Masuda, H., Arai, T. 2012. Perception of /r/ and /l/ in quiet and multi-speaker babble noise by japanese and english native listeners. *Proc. Spring Meet. Acoust. Soc. Jpn.* 477–480.

[10] Masuda, H., Arai, T. 2012. Perception of voiced english consonants in quiet and multi-speaker babble noise by japanese and english native listeners. *Proc. Fall Meet. Acoust. Soc. Jpn.* 361–364.

[11] Masuda, H., Arai, T., Kawahara, S. 2013. Preliminary analysis on the identification of english consonants in noise and/or reverberation by native japanese and english listeners. *Proc. Fall Meet. Acoust. Soc. Jpn.* 417–420.

[12] Mattys, S. L., Davis, M. H., Bradlow, A. R., Scott, S. K. 2012. Speech recognition in adverse conditions: A review. *Language and Cognitive Processes* 27(7), 953–978.

[13] Pitz, M., Ney, H. 2005. Vocal tract normalization equals linear transformation in cepstral space. *IEEE Trans. on Speech and Audio Processing* 13(5), 930–944.

[14] Van Dommelen, W. A., Hazan, V. 2010. Perception of english consonants in noise by native and norwegian listeners. *Speech Communication* 52(11), 968–979.

[15] Wong, J. W. S. 2014. The effects of high and low variability phonetic training on the perception and production of english vowels /e/-/æ/ by cantonese esl learners with high and low l2 proficiency levels. *Proc. INTERSPEECH* 524–528.