EFFECT OF SPEECH ENHANCEMENT AT PHONETIC LEVEL PERCEPTION OF ENGLISH SPEECH BETWEEN NATIVE ENGLISH AND MANDARIN LISTENERS

XXX

XXX
XXX

ABSTRACT

This study investigates the effect of commonly used speech enhancement algorithms in improving the intelligibility of noisy New Zealand English speech on native New Zealand English and native Mandarin listeners at the phonetic level. A phonetic error analysis was carried out to analyse the errors made by the listeners from a subjective listening test to find systematic errors. Results show that existing speech enhancement algorithms performed similarly by causing more onset deletion and coda deletion or insertion than when listening to clean speech. Overall, the algorithms did not improve the listeners' speech intelligibility on the phonetic level. Although native New Zealand English listeners always made fewer errors than native Mandarin listeners, they all made similar onset and coda errors. The nucleus errors made by the native Mandarin listeners were mainly caused by unfamiliarity with the accent.

Keywords: phonetic analysis, Mandarin, non-native, speech intelligibility, speech enhancement

1. INTRODUCTION

Speech intelligibility indicates how well a speech can be perceived by listeners. Non-native (L2) listeners have been known to experience more severe speech intelligibility degradation than native (L1) listeners under noisy conditions, such as listening or talking in a busy restaurant [1, 2, 3]. To improve the speech intelligibility of listeners under noisy environments, assistive listening devices implementing speech enhancement (SE) algorithms could be used [4]. While numerous numbers of SE algorithms have been developed and their performance has been thoroughly evaluated, to date their effect on L2 listeners is not well understood.

Listeners' speech perception was found to be influenced by their native language due to various factors such as the contrasts in phonetic familiarity, lexical structure, and acoustic structure between the L1 and L2 languages [3, 5]. Mandarin Chinese, spoken by 920 million people, is the second most widely spoken language in the world after English [6]. Hence, this study focuses on the perception of English speech by native Mandarin listeners.

A previous study [7] investigated the effect of different SE algorithms on two groups of listeners: native English listeners residing in New Zealand and native Mandarin listeners residing in Mainland China who have never been immersed in an English-speaking country. It was found that all SE algorithms tested showed little improvement or even degradation in speech intelligibility compared to the original noisy speech for both groups, and the perception gap between the English group and the Mandarin group was not narrowed. So far, most studies on enhanced speech evaluated speech intelligibility by the correctness of each word, but none has investigated the response error pattern at phonetic level. In addition, the previous study used a corpus in New Zealand English (NZE), whereas most of the similar subjective listening tests used corpus in General American (GenAm) or Received Pronunciation, which are the English varieties reported to be learnt by most Chinese students in Mainland [8]. Despite NZE being less commonly used, there are currently 247,770 Chinese speakers residing in New Zealand [9], showing that NZE ~~was~~ used and can be understood by a part of the local Chinese speakers in noisy environments. NZE, a non-rhotic variety of English, is considered less familiar to Chinese students. NZE shares the same vowel system as other non-rhotic standard varieties of English such as Received Pronunciation (RP), which consists of eleven monophthongs plus a neutral schwa [10], but differs from other non-rhotic varieties from the /ɪ, e, iː, æ/ vowels [11].

The current study analyses the errors made by the two groups of listeners ~~outlined in~~ [7] for each phonetic component. Explores the possible common or unique errors made by each group, and gives a glimpse of the perception of native Mandarin listeners on NZE.

2. METHODOLOGY

Phonetic level analysis was applied to the results of subjective listening tests that used speech sentences processed by different SE algorithms [7]. Two separate online subjective listening tests were conducted: 1) the *baseline test* that evaluated the clean speech in a NZE corpus by native Mandarin listeners and 2) the *speech enhancement test* that investigated the intelligibility of noisy speech processed by selected SE algorithms on native English and native Mandarin listeners. The response keywords were divided into phonetic components for marking and the systematic errors made were analysed. The rest of this section will summarise the listening tests and explain how the phonetic analysis was applied.

The study was approved by the University of Auckland Human Participants Ethics Committee (UAHPEC24202).

2.1. Subjective listening tests

Two online subjective listening tests were conducted where the participants were asked to use their own headphones ~~to perceive~~ the stimuli~~;~~ the speech intelligibility was calculated by counting the number of words correctly answered, which was normalised by the total number of words. Details of the whole test design can be found in [7].

2.1.1. Speech Enhancement (SE) Test

Twenty native NZE listeners residing in New Zealand (NZE_{SE} group, called as NZE group in [7]) and nineteen native Mandarin listeners residing in China (C_{SE} group, called as CC group in [7]) were recruited. All participants in the C_{SE} group reported having been learning either British (n = 6) and/or American English (n = 13). Each participant transcribed 108 noisy speech sentences, which included six types of speech: the original noisy speech (before SE was applied) and the enhanced speech processed by five widely used SE algorithms, namely Wiener filter (WF) [12], subspace (SS) [13], non-negative matrix factorization (NMF) [14], Conv-TasNet (Conv) [15], and complex U-Net (Unet) [16].

2.1.2. Baseline Test

Nine native Mandarin listeners based in China (C_{clean} group) were recruited. All participants reported they had learnt or had been learning American (n = 7) or British (n = 2) English. Each participant transcribed 48 clean (without any noise added) English sentences.

2.1.3. Stimuli

To generate the stimuli used in the test, the Bamford-Kowal-Bench (BKB) [17] sentences from the Speech Perception Assessment New Zealand (SPANZ) corpus [18] were used. The corpus is designed for ~~New Zealand~~ participants, ~~and~~ the sentences were recorded in NZE accent and revised to accommodate commonly used expressions in New Zealand. According to the description in the original BKB sentences, it was assumed that the occurrence of the vowels and phonemes in the keywords was balanced. Each sentence contains 3 - 4 keywords to be marked; any words other than the keywords were not marked.

2.2. Phonetic Analysis

The correctness of participants' responses from the subjective listening tests was marked manually by the first author according to the suggestions in the BKB corpus [17]. All keywords that received incorrect answers were collected for phonetic analysis. This is because even though the response words differed for each participant, the phoneme they tend to make mistakes on may be the same.

The same consonant occurring at different locations of a word may be pronounced differently, known as allophones [19], e.g. the pronunciation of consonant /l/ changes when it is placed at the beginning or end of words. Therefore, recording the location of the mistaken phonemes can be crucial for recording the errors accurately. This was realised by dividing every syllable into its phonetic components: onset, nucleus, and coda. For multi-syllable words, the phonetic components were numbered to the syllable they were in.

Overall, each word was divided into their syllables and further divided by phonetic components. Absence of a component was recorded as *NA*. An example is the two syllable word, "happy" [hæ-piː], being separated into onset 1 /h/, nucleus 1 /æ/, coda 1 *NA*, onset 2 /p/, nucleus 2 /iː/, and coda 2 *NA*. Responses with inconsistent number of syllables compared to that of the keywords were recorded with the difference in the number of syllables, where a positive number indicates more syllables were answered than that in the keyword, and vice versa. The phonetic analysis was applied to the baseline and SE test results separately, where the analysis for the SE test also recorded different SE algorithms and groups of participants separately.

This study only investigated the phonetic analysis of the first two syllables in each word as all of the baseline test keywords (83.3% mono-, 16.7% double-syllable) and 98.4% of the SE test keywords were either mono- (69.4%) or double-syllable (28.8%). The responses of three-syllable keywords were excluded as the errors made by the participants were too few to provide any systematic error, i.e. all combinations of keyword and error occurred less than twice. The same phonetic components in different syllables were not combined as the second syllable has a different phoneme error pattern compared to the first syllable.

3. RESULTS AND DISCUSSION

According to the phonetic analysis, the most common errors can be classified into deletion, missing the whole phonetic component; insertion, making up non-existing phonetic component; and substitution, substituting certain phonemes or the whole phonetic component with other phonemes. The errors were analysed by each phonetic component separately and listed in tables. Since the main purpose of this study is to investigate the performance of SE algorithms on native Mandarin listeners, the tables in this paper will focus on the error occurrence of the C_{SE} group. Table 1 shows the three most common nucleus errors for the C_{SE} group for each SE algorithm from the first syllable, where the corresponding error occurrence of the NZE_{SE} group is also ~~included~~. Tables 2, 3, 4 show the most common onset, nucleus, and coda errors of all three groups, respectively, where only the errors occurred over 15 times ~~for the C_{SE} group were~~ presented due to ~~page limitation~~. The onset errors of the C_{clean} group were excluded due to the absence of systematic onset errors, i.e. all combinations of keyword and error appeared only once.

Errors of the phonetic components were first analysed for each SE algorithm, but little difference was found between algorithms. Table 1 listed the three most occurred nucleus errors for each SE algorithm as an example. It shows that most of the errors repeatedly occur and reappear in Table 3 which summarised the overall nucleus errors. The only nucleus error that is absent in Table 3 was the [/ɜː/-/ou/] confusion in the noisy speech. This was due to limited space and the confusion is ~~actually~~ common which has occurred 14 times in total for the C_{SE} group only. The onset and coda errors caused by each SE algorithm were mostly deletion errors, which were not shown in this paper. The

Table 1: The three most common occurred nucleus errors for C_{SE} group in the first syllable for each SE algorithm.

SE Algorithm	Nucleus error (C_{SE}count, NZE_{SE}count)		
Noisy	eɪ → aɪ (7, 3)	e → iː (4, 1)	ɜː → ou (4, 0)
WF	eɪ → aɪ (7, 7)	ə → *NA* (5, 5)	ɒ→ ou (5, 1)
SS	eɪ → aɪ (9, 9)	e → eɪ (8, 0)	ɒ→ ʌ (5, 0)
NMF	eɪ → aɪ (9, 10)	ə → *NA* (8, 5)	ɒ→ oː (7, 1)
Conv	eɪ → aɪ (11, 9)	ə → *NA* (6, 4)	e → eɪ (6, 0)
Unet	ɒ→ ou (7, 0)	ə → *NA* (6, 2)	ɒ→ oː (5, 0)

Table 2: Most frequent onset errors (C_{clean} group is excluded as no systematic onset errors).

C_{SE} group	Count	NZE_{SE} group	Count
Onset 1 (first syllable)			
ð → *NA*	38	ð → *NA*	39
ð → h	17	p → *NA*	15
k → h	16	f → *NA*	13
p → *NA*	15	h → *NA*	13
f → *NA*	15	p → p l	11
Onset 2 (second syllable)			
d → *NA*	17	l → *NA*	6
s → *NA*	17	r → *NA*	6
l → *NA*	10	b → *NA*	5
d → t	9	d → *NA*	5
r → *NA*	9	s → *NA*	5

Table 3: Most frequent nucleus errors.

C_{clean} group	Count	C_{SE} group	Count	NZE_{SE} group	Count
Nucleus 1 (first syllable)					
e → ɪ	9	eɪ → aɪ	46	eɪ → aɪ	41
ʌ→ æ	7	ə → *NA*	31	ə → *NA*	25
e → iː	6	e → eɪ	26	æ→ aː	14
æ→ e	3	ɒ→ ou	20	æ→ e	11
ɪə → iː	3	**ɒ→ oː**	19	eɪ → ə	11
ou → oː	2	**e → ɪ**	17	au → e	11
æ→ ʌ	2	ɒ→ ʌ	17	ou → ʌ	7
eɪ → aɪ	2	**æ→ e**	16	eɪ → æ	7
ʌ→ aː	2	**e → iː**	16	uː → iː	7
ɪ → iː	2	**ɜː → uː**	16	iː → ə	7
ɪ → ʌ	2	**eɪ → iː**	15	æ→ eɪ	6
Nucleus 2 (second syllable)					
iː → ə	2	ə → *NA*	65	ə → *NA*	32
		ɪ → *NA*	21	ɪ → *NA*	8
		iː → ɪ	21	iː → ɪ	7

only exception was the /ʃ/-/h/ confusion in speech enhanced by NMF for the C_{SE} group (n = 5) in onset. As a result, analysis is ~~done~~ regardless of SE algorithms for the rest of this section.

For common onset errors, as reported earlier, no systematic errors were found from the C_{clean} group. However, the occurrence of the most frequent systematic onset errors increased (n = 39) similar to that of the nucleus (n = 46) and coda errors (n = 32) ~~for the C_{SE} group.~~ This increase is expected as the consonants at the beginning of an English word are often voiceless [19]. Therefore, the onset, especially the onset of the first syllable, can be masked easily by noise and/or removed by the SE algorithm. The most frequent type of onset error in the C_{SE} group was deletion, which was the same for the NZE_{SE} group, indicating that this deletion is not due to the

Table 4: Most frequent coda errors.

C_{clean} group	Count	C_{SE} group	Count	NZE_{SE} group	Count
Coda 1 (first syllable)					
l → t	3	*NA* → n	32	*NA* → n	12
NA → d	3	*NA* → d	20	n t → t	11
t → tʃ	3	n → *NA*	15	k → t	8
d → t	2	*NA* → l	15	*NA* → d	8
Coda 2 (second syllable)					
		ŋ → *NA*	27	ŋ → *NA*	10
		NA → ŋ	18	n → *NA*	8
		n → *NA*	16	*NA* → ŋ	8

language ability but the consequence of noise and SE algorithms.

Many of the frequent nucleus errors made by the Mandarin listeners (C_{clean} and C_{SE} groups) can be explained by the unfamiliarity with the NZE accent, which ~~was~~ shown in bold in Table 3, including the well-known centralised /ɪ/, the raised /e/ [10], and the more recently studied merging [/ɜː/-/uː/] [20]. Some vowel confusions were caused by the listeners being more familiar with the rhotic GenAm accent, which is learnt by most of the Mandarin participants, than NZE accent [21]. For example, the vowel confusion [/ʌ/-/æ/] can be explained by the fact that /æ/ in GenAm is usually used as /ʌ/ in NZE; the [/ɒ/-/oː/] confusion may due to the fact that NZE distinguishes /ɒ/ and /oː/ while /ɒ/ and /ɑ/ merged in GenAm [21, 19]. The most common nucleus errors in the second syllable were completely identical between the C_{SE} and NZE_{SE} groups, indicating that under noisy conditions, even the native listeners have difficulty distinguishing the vowel lengths between /ɪ/ and /iː/. Besides, the schwa in both syllables was ~~being~~ ignored frequently, 96 times by the C_{SE} group, 57 times by the NZE_{SE} group in total, as it is an unstressed vowel with a very short duration and usually appears in the unstressed syllable in the corpus used, e.g. "along" to "long". Hence, it can be easily missed or masked by noise. All three groups made two common vowel errors, which were the [/eɪ/-/aɪ/] confusion (e.g. "hay" to "hi") and the [/æ/-/e/] confusion (e.g. "bag" to "bed"), where the latter ~~one~~ is regarded as a characteristic of NZE accent. Both confusions imply that under noisy conditions, even native listeners encounter difficulties in distinguishing certain vowel pairs.

As seen in Table 4, the common coda errors from the C_{clean} group were mainly substitution errors, which can be explained by the language structure of Mandarin in which no obstruent consonant in the word-final position is allowed [22]. Hence, the C_{clean} group may be less sensitive to the word-final sound even when listening to clean speech. For the SE test, most coda errors by both the NZE_{SE} and C_{SE} groups can be concluded as deletion or insertion. This can be explained by the same reason for the missing onset, as the consonants in English at the end of a word are also mostly voiceless [19]. The NZE_{SE} group showed much fewer coda errors than the C_{SE} group, which may due to the fact that the NZE_{SE} group was able to notice "small but significant" vowel duration differences to distinguish voicing minimal pairs compared to the C_{SE} group [23]. The coda errors for the second syllable for both C_{SE} and NZE_{SE} groups were the same, related to the deletion or insertion of /n/ and /ŋ/. One possible explanation of this finding is that the SE algorithms tend to manipulate the noisy signal in a way that the nasal consonants sound similar to noise. The high occurrence of /ŋ/ may be due to the fact that many double-syllable keywords in the corpus are gerund, e.g. "playing".

4. CONCLUSION

This study analysed the phonetic errors from three groups of results from listening tests that evaluated the intelligibility of speech: native Mandarin listeners' responses to clean NZE speech (C_{clean} group), native Mandarin (C_{SE} group) and English listeners' (NZE_{SE} group) responses to enhanced NZE speech. The errors between different algorithms showed little difference, indicating that the existing SE algorithms did not improve the speech intelligibility of either native English or native Mandarin listeners at phonetic level. The result showed that the NZE_{SE} group had much fewer errors than the C_{SE} group for all phonetic components. Compared to the C_{clean} group, the C_{SE} group tended to have much more deletion errors for onset, and deletion and insertion errors for coda. Such error patterns in the onset and coda were also similar for the NZE_{SE} group.

The main limitation of this study is that the corpus is not designed for phonetic analysis. Since meaningful sentences were used in the test, participants with higher language proficiency would have been able to use semantic cues to guess the correct keywords. This top-down effect should have been considerably reduced as most of the keywords in a sentence had multiple possible substitutions. The meaningful sentences also lead to the test containing unbalance phones, resulting in a less comprehensive analysis for each phoneme. Also, part of the manual marking involved guessing the response, so the result may not fully reflect the participants' perception. Further experiments with a corpus that is designed for phonetic analysis would be recommended.

ACKNOWLEDGEMENT

We thank the participants for their participation, Dr. Suzanne Purdy for providing the SPANZ corpus, and Dr. Elaine Ballard for her insightful advice.

5. REFERENCES

[1] M. Cooke, M. L. G. Lecumberri, and J. Barker, "The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception." *The Journal of the Acoustical Society of America 123, no. 1*, pp. 414–427, 2008.

[2] S. L. L. M. C. Mattys, C. K. W. Li, and S. L. Y. Chan, "Effects of energetic and informational masking on speech segmentation by native and non-native speakers." *Speech Communication, Non-native Speech Perception in Adverse Conditions, 52, no. 11*, pp. 887–99, 2010.

[3] M. L. G. Lecumberri, M. Cooke, and A. Cutle, "Non-native speech perception in adverse conditions: A review." *Speech Communication, 52, no. 11*, pp. 864–86, 2010.

[4] P. Loizou, *Speech Enhancement: Theory and Practice*, 2007.

[5] L. Polka, "Characterizing the influence of native language experience on adult speech perception." *Perception & Psychophysics 52, no. 1*, pp. 37–52, 1992.

[6] "What is the most spoken language?" https://www.ethnologue.com/guides/most-spoken-languages, 2018.

[7] Y. C. Zhang, C. J. Hui, Y. Hioka, and C. I. Watson, "Performance of speech enhancement algorithms for native mandarin listeners on english perception," in *Proc. 15th ICA International congress on acoustics*, South Korea, 2022, pp. 111–116.

[8] W. Wenfeng and G. Xuesong, "English language education in china: A review of selected research." *Journal of Multilingual and Multicultural Development 29, no. 5*, pp. 380–99, 2008.

[9] StatsNZ. Chinese ethnic group. [Online]. Available: https://www.stats.govt.nz/information-releases/transport-broken-page/chinese

[10] C. I. Watson, M. Maclagan, and J. Harrington, "Acoustic evidence for vowel change in new zealand english." *Language Variation and Change 12, no. 1*, pp. 51–68, 2000.

[11] M. Maclagan and J. Hay, "Getting fed up with our feet: Contrast maintenance and the new zealand english 'short' front vowel shift." *Language Variation and Change 19, no. 1*, pp. 1–25, 2007.

[12] P. Scalart and J. Filho, "Speech enhancement based on a priori signal to noise estimation," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 2, 1996, pp. 629–632 vol. 2.

[13] Y. Hu and P. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 4, pp. 334–341, 2003.

[14] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2140–2151, 2013.

[15] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[16] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex U-Net," *ICLR*, p. 20, 2019.

[17] J. Bench, Ã. Kowal, and J. Bamford, "The BKB (bamford-kowal-bench) sentence lists for partially-hearing children," *British Journal of Audiology*, vol. 13, no. 3, pp. 108–112, 1979.

[18] J.-H. Kim and S. Purdy, "Speech perception assessments new zealand (SPANZ)," *New Zealand Audiological Society Bulletin*, vol. 24, pp. 9–16, 2014.

[19] H. Rogers, *The Sounds of Language: An Introduction to Phonetics.* Taylor & Francis Group, London, United Kingdom, 2000.

[20] M. Maclagan, C. I. Watson, R. Harlow, J. King, and P. Keegan, "Investigating the sound change in the new zealand english nurse vowel /ɜː/," *Australian Journal of Linguistics 37, no. 4*, pp. 465–85, 2017.

[21] M. Hay, J.and Maclagan, E. Gordon, J. Beal, P. Honeybone, and A. McMahon, *New Zealand English.* Edinburgh University Press, Edinburgh, United Kingdom, 2008.

[22] C.-C. Cheng and C.-C. Cheng, *A Synchronic Phonology of Mandarin Chinese.* De Gruyter, Inc., 1973.

[23] J. E. Flege, M. J. Munro, and L. Skelton, "Production of the word-final english /t/-/d/ contrast by native speakers of english, mandarin, and spanish." *The Journal of the Acoustical Society of America 92, no. 1*, pp. 128–43, 1992.