Performance of single-channel speech enhancement algorithms on Mandarin listeners with different immersion conditions in New Zealand English

Yunqi C. Zhang, Yusuke Hioka, C.T. Justine Hui, Catherine I. Watson

ABSTRACT

Speech enhancement (SE) is a widely used technology to improve the quality and intelligibility of noisy speech. So far, SE algorithms were designed and evaluated on native listeners only, but not on non-native listeners who are known to be more disadvantaged when listening in noisy environments. This paper investigates the performance of five widely used single-channel SE algorithms on early-immersed New Zealand English (NZE) listeners and native Mandarin listeners with different immersion conditions in NZE under negative input signal-to-noise ratio (SNR) by conducting a subjective listening test in NZE sentences. The performance of the SE algorithms in terms of speech intelligibility in the three participant groups was investigated. The result showed that the early-immersed group always achieved the highest intelligibility. The late-immersed group outperformed the non-immersed group for higher input SNR conditions, possibly due to the increasing familiarity with the NZE accent, whereas this advantage disappeared at the lowest tested input SNR conditions. The SE algorithms tested in this study failed to improve and rather degraded the speech intelligibility, indicating that these SE algorithms may not be able to reduce the perception gap between early-, late- and non-immersed listeners, nor able to improve the speech intelligibility under negative input SNR in general. These findings have implications for the future development of SE algorithms tailored to Mandarin listeners, and for understanding the impact of language immersion on speech perception in noise.

1. Introduction

Speech intelligibility, or the ability to understand spoken language, is crucial to effective speech communication. In real-life listening situations, the listening environment is often imperfect, with various factors such as reverberation and background noise affecting the quality of speech signals (Lecumberri et al., 2010; Scharenborg and van Os, 2019). One of the most significant challenges is the degradation of speech caused by noise, which is largely quantified by the signal-to-noise ratio (SNR). Plomp (1977) found that at an SNR above 0 dB (i.e. the speech signal has equal or higher energy than noise) in a cocktail party situation, native listeners are able to understand sentences easily, even if the intelligibility of individual words is low. This phenomenon is attributed to the concept of “glimpse”, which refers to the ability of listeners to utilise gaps and fragments of speech in the presence of fluctuating noise to extract meaning in the speech (Moore, 2003; Srinivasan and Wang, 2005; Cooke, 2006; Lu and Cooke, 2009). However, when the energy of noise surpasses that of the speech, i.e. negative SNR, native listeners are unable to benefit from such glimpses (Brungart, 2001; Ezzatian et al., 2010; Jin and Liu, 2012), resulting in a significant decline in speech intelligibility and leads to difficulties in real-life communication.

Such extremely adverse listening conditions pose challenges not only for native (L1) listeners but also for non-native (L2) listeners. It has been consistently proven that L1 listeners always outperformed L2 listeners in understanding speech in noise as they are more sensitive to acoustic-phonetic cues and more experienced in utilising such cues to facilitate their perception in noise (Flege et al., 1992; van Wijngaarden et al., 2002; Jia et al., 2006; Bradlow and Alexander, 2007; Lecumberri et al., 2010; Broersma and Scharenborg, 2010; Mattys et al., 2010; Jin and Liu, 2012; Mi et al., 2013; Borghini and Hazan, 2020). Furthermore, several researchers also investigated the perception and/or production performance of L2 listeners with different levels of immersion or language proficiency (Flege et al., 1992, 1997; Meador et al., 2000; Jia et al., 2006; Broersma and Scharenborg, 2010; Mattys et al., 2010; Mi et al., 2013; Kilman et al., 2014; Zhang et al., 2014). These studies discovered that L2 listeners with higher proficiency or
who have been immersed/experienced in the target language were able to achieve higher intelligibility scores under noisy conditions. Additionally, listeners’ intelligibility is found to relate to their native language due to contrasts in phonetic familiarity, lexical structure, and acoustic structure between their native language and the target language, etc. (Polka, 1992; Lecumberri et al., 2010). Therefore, it is imperative to fix the language pair when investigating perception differences across languages, as testing across various language pairs may yield dissimilar results.

To address the challenge of understanding speech in noise, speech enhancement (SE) is a widely used technique to improve the quality and intelligibility of speech contaminated by noise (Loizou, 2007; Wang and Wang, 2013; Healy et al., 2013). Ideally, to improve the speech intelligibility of L2 listeners under adverse listening conditions, the perception gap or the L2 disadvantage between the L1 and L2 listeners should be diminished after applying SE algorithms. Whilst a variety of SE algorithms have been designed to improve the intelligibility of noisy speech, previous studies evaluating their performance have focused exclusively on the response from L1 listeners of the target language. To date, to the best of the authors’ knowledge, there is no SE algorithm specifically designed for the needs of L2 listeners. Additionally, most current SE algorithms were designed for and evaluated with moderately noisy speech, i.e. the SNR being positive (above 0 dB), leaving their performance under more severe SNR (below 0 dB) cases unknown.

Given that there are more than 400 million people learning English as a second language in Mainland China, and English has become a compulsory subject since primary school (Wang, 2015), it is worthwhile studying the perception of English sentences by native Mandarin listeners under adverse conditions, specifically, under negative SNR levels. The study by Li et al. (2011) investigated the intelligibility of speech enhanced by single-channel SE algorithms by recruiting three groups of listeners with different native languages (including Mandarin). Nevertheless, in the study, participants were exclusively presented with stimuli in their native language, leaving the effect of SE algorithms on L2 listeners still unknown, which the current study addresses.

Our preliminary study (Zhang et al., 2022) investigated the performance of existing SE algorithms for native New Zealand English (NZE) listeners and native Mandarin listeners who had never been immersed in English-speaking environments. The result showed a significant difference in the intelligibility score obtained by listeners with different immersion conditions, but the tested SE algorithms were found not to significantly improve the intelligibility or even degrade compared to the unprocessed noisy speech. Moreover, the perception gap between the two groups was hardly narrowed, indicating the ineffectiveness of the SE algorithms on non-immersed L2 listeners. However, these two groups of participants were considered to have distinct differences between their immersion conditions and many of the errors made by the non-immersed native Mandarin group would have been caused by unfamiliarity with the NZE accent. This is because most of the schools in Mainland China teach General American (GenAm) or British English, which matches the demographics data from the previous test. NZE differs from other varieties of English for its raised /æ/ vowel, and the /æ/ vowel is accompanied with a lip rounding (Bauer et al., 2007; Maclagan et al., 2017). NZE is a non-rhotic English like British English, which differs from rhotic GenAm for some specific vowels. For example, NZE merges between vowels /ɔː/ and /oː/, where GenAm merges the /ɔː/ and /oː/ vowels. Moreover, NZE distinguishes between the /ɔː/ and /æː/ vowels and the /oː/ and /ʌː/ vowels, while GenAm distinguishes only the latter one (Rogers, 2000; Hay et al., 2008). According to a series of studies (Eisenstein and Verdi, 1985; Matsuiwa et al., 1999; Clopper and Bradlow, 2009), familiarity with an English dialect significantly influences L2 listeners’ speech intelligibility. This has been further proven by Hui et al. (2023) by conducting a subjective listening test on L2 listeners with different immersion conditions. The result showed that immersed L2 listeners who are familiar with the variety of English of the target language have higher intelligibility than non-immersed L2 listeners. Therefore, the findings for the performance of the SE algorithms in our previous study may not be exclusively influenced by the participants’ speech intelligibility in noise, but also by their familiarity with the accent of the target language, i.e. New Zealand English. Furthermore, due to its nature as a preliminary study (Zhang et al., 2022), the interactions between speech intelligibility, immersion condition, and SE algorithm were not extensively discussed, with a focus primarily on reporting experimental results.

To address the constraints or limitations inherent in our preliminary study, the current study is extended by including another group of native Mandarin listeners who arrived in New Zealand after the age of 15 and have been immersed in NZE for more than one year, i.e. more familiar with the NZE accent. By adding this group, we are able to investigate the speech perception abilities in noise of L2 listeners with different immersion conditions, and further investigate the effectiveness of SE algorithms on L2 listeners. This may provide more insights into the influence of immersion conditions on speech intelligibility and the potential benefits/drawbacks of SE algorithms in this population. Moreover, the way we categorise the participants is improved in this study. As noted by Cheng et al. (2021) in their study, the definition of “nativeness” can be ambiguous and contentious, making cross-study comparisons challenging. This finding also reflects our participants as the native NZE group contains bilingual listeners, who are considered to be better described using immersion age compared to the ambiguous definition of “nativeness”. Therefore, we have revised the group descriptions used in this study. Specifically, we have re-categorised the native NZE group to the “early immersed” group to reflect their lifelong exposure to the NZE accent, and the Mandarin-speaking groups have been renamed as the “late immersed” and “non-immersed” groups, respectively, to reflect their levels of exposure to the NZE accent. These terminologies are adopted from the research by Hui et al. (2022).

In summary, the current study aims to gain insight into how the presence and absence of immersion may impact the effectiveness of speech enhancement algorithms for L2 listeners. To address the limitations of the previous study (Zhang et al., 2022), we explore the interaction between speech intelligibility, immersion condition, and SE algorithm in greater detail. The findings are analysed and discussed regarding the impact on speech perception abilities in noise, potential differences between the effectiveness of SE algorithms among different immersion conditions, and the possible causes of such enhancement/degradation. In addition, a phonetic analysis is performed to gain some insights into how listeners with different levels of immersion perceive English sentences with an NZE accent. Building on our previous findings, we hypothesise that the early-immersed group always performs significantly better than the other groups and the SE algorithms will prove ineffective in improving the intelligibility for all participant groups. We also hypothesise that the late-immersed Mandarin group will perform better than the non-immersed Mandarin group due to the fact that they are more experienced in listening to English speech in noisy environments. In terms of phonetic accuracy, we anticipate fewer nucleus confusions from the late-immersed group compared to the non-immersed group due to their increased familiarity with NZE accents.

2. Methodology

An online subjective listening test was conducted to measure the speech intelligibility of participants with different immersion conditions when listening to noisy speech enhanced by different speech enhancement algorithms. This section presents a comprehensive overview of the test design and data analysis methods to address the research questions.
2.1. Participants

All participants were either at university or had received a university-based education. They were recruited to be in one of three groups:

- Early-immersed group (NZE group): 20 NZE listeners (mean age = 23.3, sd = 4.45, 6 female, 14 male) who had arrived in New Zealand before the age of 12 years old.
- Late-immersed group (NZE group): 19 Mandarin listeners (mean age = 26.5, sd = 3.69, 9 female, 10 male) who had been immersed in New Zealand English for more than one year after the age of 15.
- Non-immersed group (CM group): 19 Mandarin listeners (mean age = 23.5, sd = 1.68, 10 female, 8 male, 1 preferred not to say) residing in Mainland China who had never been immersed in English-speaking environments.

Among the participants in the NZE group, ten out of the 20 participants were monolingual speakers of NZE and the rest were heritage bilingual. The participants had lived in New Zealand for 5–10 years (n = 2), 10 years and longer (n = 18). One participant reported that they lived in Sweden until 15 years old and moved to New Zealand and lived there since then. English was reported as their first and home language. The response of this participant showed no abnormality compared to the others.

The Mandarin groups were asked to report any English proficiency certificates they had obtained. The international certificates they reported included: International English Language Testing System (IELTS) and Test of English as a Foreign Language (TOFEL). For participants who pursued their tertiary study in China, most tertiary students in Mainland China are required to pass at least one English certificate, including the Test of English Majors (TEM) Band 4 and 8 for English major students and the College English Test (CET) Band 4 and 6 for the rest. The higher Band number indicates increasing difficulty.

For the NZM group, all participants reported their first and home language as Mandarin. Ten participants had passed either IELTS (n = 8) or CET6 (n = 2), the rest had been to universities in NZ. Only one participant had neither been to NZ university nor received any English proficiency certificates and had lived in New Zealand for 7 years, who showed an abnormally low intelligibility score during analysis, which was also included in the analysis. Only two of the participants reported their work language as Mandarin, whereas the others were English or both English and Mandarin. The participants self-reported themselves to speak NZE (n = 13), American English (n = 5), and British English (n = 1). The number of years that the participants had lived in New Zealand was: 1–2 years (n = 1), 2–5 years (n = 3), 5–10 years (n = 12), 10 years and longer (n = 3).

For the CM group, two participants reported their L1 as Cantonese and the rest as Mandarin, but all reported themselves to be proficient in Mandarin. For the CM group, only two participants did not have certificates, the rest had passed at least one of CET 4 (n = 2), CET 6 (n = 9), TEM 8 (n = 1), IELTS (n = 4, two scored 6.5 and two not mentioned), and TOFEL (n = 1). All participants from the CM group reported they had learnt either British (n = 6) or American English (n = 13).

2.2. Speech enhancement algorithms

The SE algorithms tested were chosen based on three criteria: (1) being widely used during different periods of SE technology development, (2) based on the fundamental core concept, rather than being a complicated extension, (3) the original implementation is easily accessible. Considering the limitation of time spent for each participant, only a limited number of SE algorithms could be examined. Hence, for machine learning-based algorithms, this study only focused on those that are based on discriminative training and rely on conventional training criteria. Overall, five single-channel SE algorithms were selected for the test, which were:

- A priori Wiener filter (WF) (Scalart and Filho, 1996)
- Generalised subspace (SS) (Hu and Loizou, 2003)
- Unsupervised Bayesian non-negative matrix factorisation (NMF) (Mohammadiha et al., 2013)
- Conv-TasNet (Conv) (Luo and Mesgarani, 2019)
- Deep complex U-net (Unet) (Choi et al., 2019)

Implementations of the algorithms were publicly available, where WF and SS were implemented by the MATLAB code from Loizou (2007), the code for NMF and Unet were provided by the authors of their literature, and Conv was applied through Asteroid (Pariente et al., 2020).

To optimise the SE performance, some parameters of the algorithms were adjusted, except the end-to-end trained Conv and Unet. WF had its smoothing factors, $\beta$ and $\mu$ (Scalart and Filho, 1996), set to 0.96 and 0.99, respectively. For SS, the smoothing factors of $\beta$ and $\mu$ (Hu and Loizou, 2003) were set to 0.96 and 0.99, respectively. For NMF, the main buffer $N_1$ (Mohammadiha et al., 2013) was set to 50.

2.3. Stimuli

Since the study involves late- and non-immersed listeners, the sentences should not confound the performance of such listeners due to their unfamiliarity with complex words or grammar with the results from the SE algorithms or noise, and need to be in the native language of the reference group (NZE group). Hence, Bamford-Kowal-Bench (BKB) sentences in the Speech Perception Assessment New Zealand (SPANZ) corpus (Kim and Purdy, 2014) were used, which are semantically meaningful with simple syntactic structure and frequently used words. The original BKB corpus was designed for hearing-impaired children (Bench et al., 1979) and has been utilised in a number of studies on the perception of L2 listeners (Mayo et al., 1997; Van Engen, 2010; Calandruccio and Smiljanic, 2012; Hui et al., 2022). The SPANZ corpus is specifically designed for studies using New Zealand English as the target language. The corpus re-recorded 15 widely-used speech materials in NZ accent and words were modified to the common expressions in NZ (e.g. “vacation” was replaced by “holiday”). Each sentence contains three to four keywords to be marked, i.e. words other than the keywords are not marked. For example, for the sentence “The dog went for a walk”, the keywords are “dog”, “went”, and “walk”. In our prior study (Zhang et al., 2023), a separated group of non-immersed Mandarin listeners (n = 9) transcribed clean speech from the SPANZ BKB corpus, yielding an impressive 89.56% accuracy rate, thus affirming corpus validity.

The clean speech was contaminated by either stationary speech-shaped noise (SSN) or non-stationary babble noise recorded at the sampling rate of 16 kHz. The SSN was a white noise the spectrum of which was shaped by the power spectrum of the sum of 288 SPANZ BKB sentences. The babble noise was from the NOISEX-92 corpus (Varga and Steeneken, 1993). Since each target sentence had a different length between 3–4 s, the noise signals were truncated to the same length as the speech signal where the starting point of the noise signals was fixed. The noise was added to the clean speech at specific input signal-to-noise ratio (SNR) levels to generate the noisy speech. To avoid significant ceiling/flooring effect across all participant groups, input SNR levels of 0, −3, −6 dB for the babble noise, and −3, −6, −9 dB for the SSN were chosen based on a pilot test. The sound level of the target speech sentences remained the same regardless of the input SNR levels.

The noisy speech was processed by the SE algorithms mentioned in Section 2.2 to obtain the enhanced speech. Even though some SE algorithms may adjust the sound level of the input signal, no manipulation was conducted to the output signals. Besides the enhanced speech, the original noisy speech signals were also included in the test as the baseline reference. Speech signals processed by the same SE algorithms under a certain input SNR level were considered as one condition and each condition was repeated three times with randomly selected different sentences. Therefore, each participant listened to 108
sentences ((5 SE algorithms + 1 noisy speech) × 2 noise types × 3 input SNR levels × 3 repetitions) in the test.

To quantitatively evaluate the performance of the SE algorithms both from objective and subjective points of view, the improvements in the signal-to-distortion ratio (SDR) (Vincent et al., 2006), perceptual evaluation of speech quality (PESQ) (Rix et al., 2001), scale-invariant SNR (SI-SNR) (Luo and Mesgarani, 2018), and the short-time objective intelligibility (STOI) (Taal et al., 2011) were also measured. These four objective metrics have been widely used to identify the effectiveness of SE algorithms by researchers and such metrics have also evaluated most of the SE algorithms we tested in this experiment. Table 1 presents the mean differences in these metrics before and after applying the SE algorithms. The results demonstrate that Conv outperformed the other algorithms for all conditions. The decline in speech quality and intelligibility, as determined by PESQ and STOI, respectively, occurred in the rest of the SE algorithms; however, such occurrences were infrequent. Overall, most of the SE algorithms were able to reduce the amount of noise and/or improve the quality or intelligibility of noisy speech objectively, which supports our purpose of investigating their performance by subjective listening test.

2.4. Test procedure

The test was developed by PsychoPy (Peirce et al., 2019) and was made available online through Pavlovia. The participants were asked to transcribe the sentences played through headphones via the GUI as shown in Fig. 1 by a keyboard. A practice page was given for the participants to adjust the sound level and get familiar with the test. They were instructed to adjust the sound level of their device to the maximum level that they deemed comfortable and were not allowed to change the setting during the formal test. To avoid the sound level of sentences with low input SNR levels being too noisy to tolerate, an unprocessed noisy signal at −9 dB input SNR, which is the case with the highest possible sound level, was included in the practice.

Demographic information of the participants was collected through a questionnaire, including the gender, age, first language, type of English speaking, language background, year of staying in New Zealand (for NZE and NZM groups), proficient language(s), educational background, English language certificates (for NZM and CM groups), and whether they experience hearing loss.

During the formal test, each participant listened to 108 sentences as summarised in Section 2.3. Every sentence was played automatically and was not repeatable. The same set of sentences were used for every participant, but with random parameter combinations and were played in an arbitrary order. The duration of the test was around 30 min excluding the break and the participants were given a break after listening to 54 sentences, with the freedom to take as much time as required.

The participants were rewarded koha (a New Zealand Māori custom which means gift, donation or contribution) as a token of appreciation. The procedure was approved by the University of Auckland Human Participants Ethics Committee (24202).

2.5. Marking rubric

Responses of participants who self-reported that they had no hearing impairment at the time of sitting the experiment were marked manually based on the BKB corpus’ marking criteria (Bench et al., 1979). The root of the word was marked instead of the whole word, e.g. “buying” and “bought” were both marked correct for the keyword “buy”. Homonyms such as “son” responses “sun”, “two” responses “to” were not subject to a penalty. To quantify the speech intelligibility of the participants, the ratio of correct words (proportion correct) was calculated for each participant under each condition which included three sentences as mentioned in Section 2.3.

2.6. Statistical analysis

The speech intelligibility score indicated by the proportion correct obtained by the participants was analysed using the linear mixed effect model (LME) in R with the lme4 package (Bates et al., 2015). Since the experiment involved two types of noise with different input SNR levels, separate model was developed for each noise type. Step function in the lmerTest package (Kuznetsova et al., 2017) was carried out to obtain the best fitted model. The fixed effects were immersion condition (participant group, i.e. NZE, NZM, and CM), input SNR, and SE algorithm, where the baseline (i.e. the unprocessed noisy speech) was also included. The random effect was the participant ID. Likelihood ratio test was conducted to check the significance of the fixed effects. Post-hoc pairwise comparisons between models were achieved by the emmeans package (Lenth et al., 2022) with the p-values adjusted by the Tukey HSD method. Results with p-values being less than 0.01 were considered significant.

2.7. Phonetic analysis

The phonetic analysis summarises the frequent errors made by the participants in each group. All keywords that received incorrect answers were collected since the same phonetic confusion may occur in different response words. A phoneme that occurs at different locations in a word can have various pronunciations depending on their surrounding phoneme environments, which is known as co-articulation (Daniloff and Hammargard, 1979). The root of the word was marked instead of the whole word, e.g. “buying” and “bought” were both marked correct for the keyword “buy”. Homonyms such as “son” responses “sun”, “two” responses “to” were not subject to a penalty. To quantify the speech intelligibility of the participants, the ratio of correct words (proportion correct) was calculated for each participant under each condition which included three sentences as mentioned in Section 2.3.
Fig. 2. Predicted proportion correct from linear mixed effect model under different noise types in terms of immersion condition separated by algorithms. The error bars represent the 95% confidence intervals.

Table 2

Pairwise contrasts of speech intelligibility proportion correct scores among NZE, NZM, and CM groups. The entries for SSN are reduced as SE algorithm and immersion condition have no significant interaction according to the model.

<table>
<thead>
<tr>
<th>Babble noise</th>
<th>NZE - CM</th>
<th>NZE - NZM</th>
<th>NZM - CM</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Algorithm</strong></td>
<td><strong>SNR</strong></td>
<td><strong>t.ratio</strong></td>
<td><strong>p.value</strong></td>
</tr>
<tr>
<td>noisy</td>
<td>0</td>
<td>0.53</td>
<td>11.47</td>
</tr>
<tr>
<td>Conv</td>
<td>0</td>
<td>0.41</td>
<td>8.8</td>
</tr>
<tr>
<td>Unet</td>
<td>0</td>
<td>0.33</td>
<td>7.05</td>
</tr>
<tr>
<td>NMF</td>
<td>0</td>
<td>0.47</td>
<td>10.15</td>
</tr>
<tr>
<td>SS</td>
<td>0</td>
<td>0.51</td>
<td>11.11</td>
</tr>
<tr>
<td>WF</td>
<td>0</td>
<td>0.53</td>
<td>11.53</td>
</tr>
<tr>
<td>noisy -3</td>
<td>0</td>
<td>0.42</td>
<td>8.98</td>
</tr>
<tr>
<td>Conv -3</td>
<td>0</td>
<td>0.29</td>
<td>6.31</td>
</tr>
<tr>
<td>Unet -3</td>
<td>0</td>
<td>0.21</td>
<td>4.57</td>
</tr>
<tr>
<td>NMF -3</td>
<td>0</td>
<td>0.36</td>
<td>7.66</td>
</tr>
<tr>
<td>SS -3</td>
<td>0</td>
<td>0.4</td>
<td>8.62</td>
</tr>
<tr>
<td>WF -3</td>
<td>0</td>
<td>0.42</td>
<td>9.04</td>
</tr>
<tr>
<td>noisy -6</td>
<td>0</td>
<td>0.31</td>
<td>6.6</td>
</tr>
<tr>
<td>Conv -6</td>
<td>0</td>
<td>0.18</td>
<td>3.93</td>
</tr>
<tr>
<td>Unet -6</td>
<td>0</td>
<td>0.21</td>
<td>2.19</td>
</tr>
<tr>
<td>NMF -6</td>
<td>0</td>
<td>0.24</td>
<td>5.28</td>
</tr>
<tr>
<td>SS -6</td>
<td>0</td>
<td>0.29</td>
<td>6.24</td>
</tr>
<tr>
<td>WF -6</td>
<td>0</td>
<td>0.31</td>
<td>6.66</td>
</tr>
<tr>
<td>SSN -3</td>
<td>0</td>
<td>0.44</td>
<td>10.6</td>
</tr>
<tr>
<td>−6</td>
<td>0</td>
<td>0.45</td>
<td>10.68</td>
</tr>
<tr>
<td>−9</td>
<td>0</td>
<td>0.31</td>
<td>7.42</td>
</tr>
</tbody>
</table>

1973). Therefore, every syllable was divided into its phonetic components: onset, nucleus, and coda, to record the location of the phonemes to eliminate the ambiguity caused by co-articulation. Phonetic components of multi-syllable words were numbered to the syllables they were in. The missing of a whole syllable and the primary stress in a multi-syllable keyword were also recorded.

3. Result

Both two LME models separated by different noise types had significant two-way interaction between variables according to the likelihood ratio comparison. For the babble noise, the interactions were between the SE algorithm and input SNR ($\chi^2(10) = 4763, p < .0001$), immersion condition (NZE, NZM, and CM groups) and SE algorithm ($\chi^2(10) = 4778.3, p < .0001$), and input SNR and immersion condition ($\chi^2(4) = 4786.5, p < .0001$). For SSN, the interactions were between the SE algorithm and input SNR ($\chi^2(10) = 4701.6, p < .0001$) and between input SNR and immersion condition ($\chi^2(4) = 4697.5, p < .0001$).

Fig. 2 shows the linear prediction of the proportion correct in terms of the immersion condition of the participants separated by SE algorithms. Each subplot represents the performance of a SE algorithm on different groups of participants, where the groups are distinguished by different colours and the error bars display the 95% confidence interval. The plot also allows observation of trends in speech intelligibility among participant groups and different input SNR levels. The
same trend in proportion correct across all immersion conditions can be seen in Fig. 2; the NZE group shows a much higher predicted proportion correct than the Mandarin groups, i.e. followed by the NZM group and the lowest for the CM group. This is supported by Table 2, which presents the results of the post-hoc pairwise contrasts among the NZE, NZM, and CM groups in terms of SE algorithm and input SNR levels. The proportion correct of the NZE group is significantly higher than the Mandarin groups for most cases except for Unet at −6 dB input SNR, while the NZM group performs significantly better than the CM group under high input SNR levels (0 dB under babble noise, −3 and −6 dB under SSN). Both Mandarin groups have similar proportion correct under the lower input SNR, i.e. −9 dB under SSN and −3 and −6 dB under babble noise except for WF.

### 3.2. Effect of speech enhancement algorithms on speech intelligibility

By comparing the predicted proportion correct in terms of SE algorithms through Fig. 2, the original noisy speech has the highest intelligibility for all groups of participants and under all input SNR conditions for most cases. This is confirmed by Table 3, which displays the post-hoc pairwise contrasts between different SE algorithms in terms of the participants’ immersion condition and input SNR level. We find that noisy speech significantly outperforms the five SE algorithms, especially under low input SNR levels. The pairwise contrasts under SSN in Table 3 are simplified as the SSN LME model shows no significant interaction between the SE algorithm and immersion condition, i.e. changing one of the effects does not influence the other one.
In general, Unet performs the worst under most conditions as shown in Table 3. Moreover, the insignificant contrast in proportion correct among all participant groups in Table 2 and the near-flooring proportion correct of Unet under −6 dB under babble noise for all groups in Fig. 2 also prove the ineffectiveness of Unet, especially under low input SNR. However, Unet shows the second-best intelligibility improvement as shown in Fig. 2, which may be common for listeners who are familiar with other varieties of English. Another common confusion caused by unfamiliarity with the NZE accent was /u/ → /æ/, which occurred frequently in both NZM and CM groups.

### 4. Discussion

In the current study, we investigate how existing speech enhancement algorithms affect the speech intelligibility of listeners with different immersion conditions. This section discusses how SE algorithms and immersion conditions interactively affect speech intelligibility.

Referring to the original literature of the five SE algorithms we evaluated, all of them have been tested using either SSN or babble noise, or both. Both Conv (Luo and Mesgarani, 2019) and NMF (Mohammadiha et al., 2013) were tested down to −5 dB input SNR whereas the others were only tested for positive input SNR conditions (Scalart and Filho, 1996; Hu and Loizou, 2003; Choi et al., 2019). The most prominent objective improvement in both intelligibility and quality as reported in Table 1 was achieved by the DNN-based methods (Conv and Unet).

#### 4.1. Effect of immersion condition on speech intelligibility

The fact that the NZE group always performed significantly better than the Mandarin groups under all conditions as shown in Table 2 is aligned with our hypothesis of expecting the highest intelligibility score from the NZE group. The only exception is for Unet at −6 dB input SNR under babble noise, which is potentially being caused by the flooring effect. Since the sentences in the SPANZ BKB corpus have a relatively high predictability context (Schoof and Rosen, 2015), the contextual information may have been available during speech perception. Bradlow and Alexander (2007), O’Neill et al. (2020) found that semantic information is more accessible to L1 listeners under noisy conditions even if both L1 and L2 listeners are able to benefit from it. They also stated that L2 listeners exhibited reduced sensitivity to the phonetic distinctions of the target language, possibly because of their limited exposure to the complete set of acoustic-phonetic cues. Therefore, it is assumed that the Mandarin groups, especially the CM group, are forced to rely heavily on acoustic-phonetic cues for perception due to their unfamiliarity with NZE.

Contrary to our hypothesis, the performance of the NZM and CM groups had no significant difference at lower input SNR levels (i.e. −3 and −6 dB under babble noise and −9 dB under SSN) while we expected the CM group would always perform worse than the NZM group due to their different levels of immersion (Flege et al., 1992, 1997; Meador et al., 2000; Jia et al., 2006; Mi et al., 2013). One of the possible reasons for the disagreement would be that, by looking at the intelligibility scores in Fig. 2, both Mandarin groups reached almost flooring at the lowest input SNR levels (−6 dB for babble noise and −9 dB for SSN). This indicates that under extremely noisy conditions, the late- and non-immersed listeners are less likely to capture enough acoustic-phonetic cues during perception. In fact, Flege et al. (1992) have investigated
the production of word-final English consonant pair /t/-/d/ by native English and native Mandarin listeners. It was found that experienced L2 speakers did not produce word-final contrast consonants significantly better than the inexperienced ones, which aligns with our findings. Even though the current study focuses on speech intelligibility, it has been proven that the degree of accuracy in production and perception of speech are closely related (Flege et al., 1997). However, other studies have reported that experienced/proficient non-native listeners are able to perform significantly better than inexperienced/less proficient L2 listeners for both production and perception (Flege et al., 1997; Kilman et al., 2014; Jia et al., 2006). One possible explanation for these contradicting results is that the majority of these studies focus on vowels or sentences. Vowels are voiced and have a longer duration than voiceless stop consonants at word-final positions (Rogers, 2000) and sentences allow top-down effects, hence, making the perception cues more accessible to experienced/proficient L2 listeners. In contrast, for the stop consonant pairs, the major way to distinguish and produce them is by making perceptual use of the difference in vowel duration before such consonant pairs. This may indicate that the ability to notice and utilise subtle acoustic-phonetic cues is less likely to be obtained by late- and non-immersed listeners under extremely noisy conditions, especially when such cues may be mitigated or removed by SE algorithms. Moreover, our study tested the intelligibility of enhanced speech, where the speech is very likely to be distorted by SE algorithms, Bradlow and Alexander (2007) demonstrated that L2 listeners were only able to utilise higher-level contextual information when there were sufficient acoustic-phonetic cues. Therefore, auditory cues may be less accessible to the Mandarin groups and limit the accessibility of top-down information from the sentences for the NZM group, leading to an insignificant difference in intelligibility between them is by making perceptual use of the difference in vowel duration before such consonant pairs. This may indicate that the ability to notice and utilise subtle acoustic-phonetic cues is less likely to be obtained by late- and non-immersed listeners under extremely noisy conditions, especially when such cues may be mitigated or removed by SE algorithms. Moreover, our study tested the intelligibility of enhanced speech, where the speech is very likely to be distorted by SE algorithms, Bradlow and Alexander (2007) demonstrated that L2 listeners were only able to utilise higher-level contextual information when there were sufficient acoustic-phonetic cues. Therefore, auditory cues may be less accessible to the Mandarin groups and limit the accessibility of top-down information from the sentences for the NZM group, leading to an insignificant difference in intelligibility between the NZM and CM groups.

To further prove the influence of the immersion condition, phonetic analysis was conducted to compare the performance difference between the Mandarin groups on phonetic level. According to Table 5, some of the common nucleus errors can be attributed to unfamiliarity with the NZE accent. These errors include the well-reported raised /o/ and /æ/ to /I/ and /æ/, respectively (Watson et al., 2000), and the more recently studied merging /[x]/-/[u]/ (Maclagan et al., 2017). As reported by our previous phonetic study (Zhang et al., 2023), non-immersed Mandarin listeners also tend to mishear /o/ to /I/ when listening to clean NZE speech. Hence, the transcription errors of the CM group in the current study are very likely to have resulted from a confluence of the unfamiliarity with NZE vowels, the noise, and the distortion caused by the SE algorithms. Interestingly, except for the /[x]/-/[u]/ confusion where the NZM group made fewer mistakes than the CM group, the two Mandarin groups performed similarly in terms of the other errors mentioned above. This finding contradicted with our hypothesis; the NZM group would have fewer nucleus confusions due to unfamiliarity with the NZE accent since they have been immersed in NZE for at least a year. This may indicate that even experienced late-immersed listeners may find the characteristics of NZE accent challenging in noise. Even though the NZM group showed little advantage in terms of distinguishing NZE vowels, the higher intelligibility score achieved by the NZM group at higher input SNR levels indicates that they may still be able to perceive universal phonemes more accurately than the CM group due to their extensive exposure to NZE. Additionally, some vowel confusions may be due to the listeners' greater familiarity with the rhotic General American (GenAm) accent, which occurs with a lower but notable occurrence. For instance, the confusion of /[x]/-/[u]/ (NZM occurrence = 7/475, CM occurrence = 15/475) may be explained by the fact that /æ/ in GenAm is often used in place of /æ/ in NZE. The confusion of /[x]/-/[u]/ (NZM occurrence = 12/608, CM occurrence = 19/608) may result from the fact that NZE distinguishes between /o/ and /æ/, whereas these vowels merge in GenAm (Rogers, 2000; Hay et al., 2008). Such confusion caused by the difference between NZE and GenAm may be due to the fact that participants from the CM group had never been exposed to NZE, and most of them had received their English education in GenAm. Also, both confusions occurred for the NZM group, but with slightly lower occurrences, possibly because the participants had learnt GenAm in their early English education, at least until the age of 15. Therefore, there is a possibility that they may still be affected by the English learnt at an earlier age. However, to the best of our knowledge, this topic has not been investigated in prior research. All three groups exhibited two common nucleus confusions, specifically confusion between the vowel pairs /[æ]/-/[au]/ (e.g. “hi” to “hay”) and /[æ]/-/[ɒ]/ (e.g. “bed” to “bag”), where the latter pair is likely due to the raised /æ/ in NZE accent which often causes such confusion to non-NZE speakers. These confusions indicate that even early-immersed listeners face challenges in distinguishing certain vowel sounds under noisy conditions. The most frequent nucleus error for all participant groups was the deletion of schwa. This would have been mainly caused by the extremely short duration of this unaccented vowel having been often located in the unstressed syllable in the corpus we used, e.g. “around” to “round”. Hence, the vowel may be easily masked by the noise and removed by the SE algorithms. These findings highlight the impact of familiarity with a variety of English on phonetic perception.

4.2. Effect of speech enhancement algorithms on speech intelligibility

The unprocessed noisy speech is slightly but statistically significantly more intelligible than any of the enhanced speech according to Table 3, which aligns with the finding in Li et al. (2011). Although the test conducted in this early research was limited to five traditional SE algorithms with input noisy speech signals at positive input SNRs for listeners with different native languages targeted in their native languages, the result also showed that most of the SE algorithms suppress the background noise with the cost of distorting the speech. Similarly, Cooke and Lecumberri (2016) pointed out that “while enhanced styles lead to gains by reducing the effect of masking noise, the same styles distort the acoustic-phonetic integrity of the speech signal”. This previous study found that modified speech styles which are beneficial in noise failed to achieve higher intelligibility in quiet conditions. Our current study revealed that the current SE algorithms yielded lower intelligibility than the unprocessed noisy speech. Hence, the results from both research showed that such manipulation of speech might result in damaging its acoustic-phonetic structure.

From the STOI improvements shown in Table 1, we expected that the proportion correct of all the SE algorithms apart from NMF, would be higher than that of the original noisy speech, while the result showed the contrary. A plausible explanation for the degradation can be drawn from the design of these algorithms; most of them were designed in such a way as to improve the speech with respect to the objective metrics (SNR, SDR, SSNR, etc.), but not the actual speech intelligibility perceived by listeners. Therefore, these methods may have focused too much on eliminating the noise from the signal, resulting in removing some critical phonetics features (e.g. initials or ends of a word) or distorting the speech signal, which will be explained in detail in the following section. The exception is Conv-TasNet, where a subjective test was also conducted in the original study. However, the test asked the participants to mark the quality of the enhanced speech but did not measure how well the participants could understand them, i.e. intelligibility (Choi et al., 2019). Also, the input SNR levels used in the current study were lower than that assumed in the original literature, hence, such algorithms may not be designed to perform well under negative input SNR levels.

Another interesting finding is that, when listening to the signals processed by the worst performer, Unet, the signals sounded very clean with most of the noise removed, which may have led to the over-removal of phonetic components. The phonetic analysis further proves this observation, where most of the onset and coda errors made by the participants were missing the whole phonetic component regardless of immersion condition. Since some onset and coda consist of voiceless
consonants, which have a shorter duration (Rogers, 2000), it is very likely that background noise may partially or completely mask them, especially under low input SNR conditions. Moreover, the SE algorithms are designed to suppress and remove the noise and may treat these phonetic components as part of background noise and remove them in the process.

Overall, Table 3 shows that the NZE group is more capable of capturing the small but statistically significant difference between SE algorithms under babble noise. Such behaviour is not found in the NZM nor CM groups, which may reflect that late- and non-immersed listeners perceive enhanced and noisy speech with similar intelligibility and are less able to benefit from the SE algorithms than the early-immersed listeners under non-stationary noise. Such inability to observe the difference between noisy and enhanced speech is found under SSN for all participant groups, with the difference between algorithms being more significant under the lowest input SNR level (−9 dB).

According to Fig. 2, the perception gaps among the three groups at each input SNR level were barely narrowed by applying SE algorithms. This corroborates our hypothesis about the ineffectiveness of SE algorithms on all participants regardless of immersion conditions. The only exception is for Unet at −6 dB under babble noise, where the difference in intelligibility score among the participant groups is insignificant according to Table 2. However, such a narrowed perception gap may be due to the significant drop in the intelligibility score caused by the flooring effect (i.e. proportion correct lower than 0.25 as shown in Fig. 2).

4.3. Limitations and future work

The limitations of the research can mainly be attributed to three factors. Firstly, due to the COVID-19 pandemic, the test had to be conducted online. According to the research by Cooke and García-Lecumberri (2021), online speech perception tests are reliable when performed by known participants and avoid low-quality headphones. Since most of the participants in the Mandarin groups were known by the authors and were instructed to avoid low-quality headphones, the results from this study should be appropriate. However, since no information about the quality of the participants’ headphones was collected, it is still possible that the speech perception of some participants’ was influenced. Additionally, the participants’ listening environments varied and the sound level of the stimuli was not controlled. While participants were instructed not to adjust the device’s sound volume during the formal test, this could not be monitored. Since the volume of the enhanced speech differed significantly among certain SE algorithms, e.g. signal processed by Conv had a much higher energy than that processed by NMF, this would have potentially impacted the participants’ responses.

Secondly, since meaningful sentences were used in the test, listeners with higher proficiency would have taken advantage of using semantic cues to guess the correct keywords. Due to such sentences, responses treating multiple words as a single word were common, e.g. “cost a lot” as “customer”. Additionally, the corpus was not optimised for phonetic analysis as test sentences contain an unbalanced distribution of phones, which results in a less comprehensive analysis of individual phonemes. This is the main reason why the study did not focus on the phonetic analysis and only briefly covered it in Section 4.1. Furthermore, since the participants transcribed whatever they heard, part of the marking involved guessing their corresponding response to the keyword. Therefore, the marked results may not fully reflect their responses.

Thirdly, the experiment only tested noisy speech with negative input SNR conditions to avoid the intelligibility score of the early-immersed listeners suffering from ceiling effect. However, as shown in Fig. 2, the intelligibility of the Mandarin listener groups (NZM and CM) tested at the highest input SNRs (0 dB for babble noise, −3 dB for SSN) exhibited a large potential for improvement under positive input SNR conditions.

Since this is the early study to investigate the performance of single-channel SE algorithms on non-native listeners, it primarily serves to broaden our understanding of the general effects of SE on this listener group. Future studies are required to understand the effect of SE algorithms on listeners with different language exposure and language familiarity. This includes using a different corpus with low predictability sentences, controlling the listening environment (e.g. by conducting the test in person), testing non-native listeners with different first languages under positive input SNR conditions, and conducting such subjective listening tests to evaluate SE algorithms designed to enhance speech intelligibility, as well as machine learning-based methods with diverse training and learning models.

5. Conclusion

This study investigated the effect of different speech enhancement (SE) algorithms on speech intelligibility with a glimpse of the phonetic perception of early-immersed New Zealand English (NZE) listeners and native Mandarin listeners with varying degrees of immersion in NZE in noisy environments. The results indicated that the early-immersed group consistently performed better than the late- and non-immersed groups in terms of speech intelligibility, and that noisy speech was generally more intelligible than enhanced speech. However, the late-immersed group did show a higher intelligibility score than the non-immersed group at higher input signal-to-noise ratios (SNR), although this advantage diminished at extremely low SNR levels (i.e. at input SNRs lower than −6 dB). The SE algorithms tested showed insignificant differences from each other, with the exception of the deep complex U-net (Unet) algorithm performing worse. Furthermore, the study found that the Mandarin groups were less able to detect intelligibility differences between speech processed by different SE algorithms, suggesting that these algorithms may not effectively improve speech intelligibility for late- and non-immersed listeners. Finally, phonetic errors in noisy and enhanced speech were found to be similar, where the unfamiliarity with the NZE accent led to nucleus errors for both Mandarin groups, where the late-immersed listeners were also unable to take advantage of their immersion experience. Overall, these findings have implications for the development of SE algorithms, especially for non-native listeners, and the understanding of the effects of language immersion on speech perception in noise.

CRediT authorship contribution statement

Yunqi C. Zhang: Conceptualization, Methodology, Software, Validation, Investigation, Resources, Formal analysis, Data curation, Writing – original draft, Writing – reviewing & editing, Visualization.
Yusuke Hioka: Conceptualization, Methodology, Writing – original draft, Writing – reviewing & editing, Supervision, Project administration.
C.T. Justine Hui: Conceptualization, Methodology, Resources, Writing – review & editing, Supervision.
Catherine I. Watson: Conceptualization, Methodology, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgements

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