INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

# Speech-in-speech recognition is modulated by familiarity to dialect

*Jessica L. L. Chin*[1], *Elena Talevska*[1,2], *Mark Antoniou*[1]

[1]The MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Australia
[2]School of Psychology, Western Sydney University, Australia
jessica.chin@westernsydney.edu.au, 19732826@student.westernsydney.edu.au, m.antoniou@westernsydney.edu.au

## Abstract

Listening to speech in competing background speech can be difficult due to elements such as the linguistic content of the signal. Linguistic release from masking occurs when altering the masker language results in less interference in speech recognition. The greater the linguistic differences between the target and masker, the higher the speech recognition accuracy. However, for dialectal variations of the same language, these patterns are less consistent. This study examined speech-in-speech recognition in Australian English monolinguals when the target speech was in either Australian (AU) or American English, and when the masker speech was in either of the dialects or a foreign language (Swedish). Speech recognition performance was greatest when AU was the target and poorest when AU was the masker. There were fewer differences in performance between the Swedish and dialect maskers. Results indicate that speech recognition is modulated by a listener's familiarity to a dialect.

**Index Terms**: speech-in-speech recognition, target–masker similarity, dialect variation, familiarity to dialect, foreign language

## 1. Introduction

When listening to speech in noisy situations, one must continuously attend to the target talker whilst in the presence of distractors such as environmental noise and other people speaking in the background. These distractors, or maskers, can interfere with the recognition of the target speech in a multitude of ways. This is often categorised as energetic or informational masking [1]. Energetic masking refers to the spectral and temporal overlap between the target and masker, such as observed in loud traffic or industrial noise. Due to these overlaps in signal, individual words spoken by the target talker can become difficult or impossible to comprehend [2]. Informational maskers, on the other hand, cause interference in speech recognition outside of energetic effects of masking (such as the intensity of the masker signal) and can include elements of the masker itself. In the case of distracting background speech, this can refer to the unique information contained in the masker: the language or dialect, or the content of the speech [3]. Effects of both energetic and informational masking interact to reduce speech intelligibility [4].

When comparing between different target–masker conditions in speech-in-speech recognition, we can sometimes observe linguistic release from masking (LRM), which is a reduction in informational masking effects depending on the language of the masker [5]. This is predicated by the linguistic similarity between the target language and that of the masker (the linguistic similarity hypothesis); that is, the more typologically distant the target and masker languages are, the easier it is to distinguish both signals and thus attend to the target [3, 6]. Greater LRM is observed in conditions where the masker language is foreign or unfamiliar to the listener (e.g., English-in-Dutch for American English listeners) compared to conditions where the target and masker are matched (e.g., English–English) [5–8]. It has been contested whether there are differences in performance between two unfamiliar masker languages of varying similarity to the target; more recently, one study found no significant differences between English–Dutch (unfamiliar, but both stress-timed) and English–Mandarin (unfamiliar and mismatched in rhythmic structure) conditions [7].

Linguistic release from masking has also been observed between maskers varying in dialect or accent. In less mutually intelligible dialects of a language (such as Limburgian and Dutch), LRM effects are observed for both L1 Limburgian and L1 Dutch listeners for the mismatched Dutch–Limburgian condition versus the matched Dutch–Dutch condition [3]. For mutually intelligible dialects such as General and Southern American English, a mismatch between the target and masker dialects resulted in greater LRM effects, but only at higher noise levels and for the nonnative Southern US target. Further, the General American English listeners showed higher performance for the Southern US target, which may indicate that phonetic or prosodic features in a dialect may be more intelligible than others [9]. In another study, Southern British and Glaswegian Scottish listeners were presented with familiar, unfamiliar, and foreign-accented speech in nonspeech noise. Performance decreased the more unfamiliar the target dialect was to the listener [4]. LRM effects attributed to unfamiliarity to the masker is present also in foreign-accented speech (e.g., Mandarin-accented English), though the effect is not widespread—while native listeners benefit from the mismatch between native and foreign-accented speech, L2 English listeners and hearing-impaired native listeners do not [10, 11]. In sum, when listeners are exposed to dialects or accents which are intelligible yet unfamiliar, they still experience LRM with mismatched target and masker pairs compared to matched pairs. Listeners also experience greater LRM effects the less familiar they are to the masker. However, it is yet unclear how two intelligible masker dialects (familiar and unfamiliar to the listener) and an unfamiliar foreign masker language would rank in terms of speech recognition performance.

The present study aimed to investigate speech-in-speech recognition in Australian English monolinguals in two mutually intelligible English dialects: Australian English (AU) and American English (US). Participants listened to target sentences in these dialects in the presence of two-talker masker

10.21437/Interspeech.2023-1196

babble in AU, US, and a foreign language, Swedish (SE). The study examined the relationships between the target and masker speech through the linguistic similarity hypothesis, as well as the familiarity of the target and masker languages to the AU listener. The following hypotheses were tested: Firstly, the listeners would recognise the AU targets with higher accuracy than the US targets, regardless of masker. Secondly, performance would be poorest in conditions where the masker is AU, followed by the US maskers, then the SE maskers. Finally, we posited that the effects of target–masker linguistic similarity and listeners' familiarity with the target/masker would be additive, such that greater LRM outcomes will be observed for AU targets in the presence of unfamiliar (US) or foreign (SE) maskers. Alternatively, if the effects of listener familiarity outweigh that of target–masker linguistic similarity, stronger LRM effects will be observed for a familiar, matched target–masker pair (AU–AU) than for an unfamiliar target and native masker pair (US–AU).

# 2. Method

## 2.1. Participants

A total of 56 Australian English monolinguals ($n = 43$ females) aged between 18 and 48 years ($M_{age} = 24.5$, $SD = 7.71$) participated in this study. All participants were native speakers of Australian English and were either born in Australia or arrived in Australia before the age of 10. Three participants were excluded from analyses after failing to follow experiment instructions.

Participants were undergraduate students studying psychology at Western Sydney University. They were reimbursed with credit points for their course. This study was approved by the University's Human Research Ethics Committee.

## 2.2. Stimuli

The experiment followed a (2 × 3) within-subjects design with two independent variables: dialect of the target speech (AU, US), and language/dialect of the masker speech (AU, US, SE). Both English target speech stimuli consisted of 288 sentences taken from the revised Bamford-Kowal-Bench (BKB-R) sentence list [12]. Commonly used in speech-in-speech recognition tasks, the BKB-R sentences contain high frequency English words with a controlled syllable length and morphological/phrasal structure, e.g., "The CLOWN had a FUNNY FACE." One male native AU talker produced the stimuli for the AU target, and one male native US talker produced the stimuli for the US target.

The masker sentences were drawn from the Syntactically Normal Sentence Test (SNST; [13]). The 200 SNST sentences are controlled in word frequency and structure like the BKB-R, but are semantically anomalous, e.g., "The LOW WALK READ the HAT." These masker sentences were produced by two male native AU talkers and two male native US talkers. The masker talkers differed from the AU and US target talkers. For the SE masker condition, the SNST list was translated into Swedish and later produced by two male Swedish talkers.

All recordings took place in an acoustically shielded sound booth at a sampling rate of 44.1 kHz (16 bit) using a Shure SM10A cardioid microphone. Recordings were segmented into individual sentences using Praat [14]. To create the stimuli, the masker sentences were recompiled into a single track for each talker. A selection of each masker track was randomly excised to create the experimental stimuli. The onset and offset of the masker babble was always 500 ms before and after that of the target track. The target sentences were set to 65 dB SPL and the masker sentences were set to 70 dB SPL (-5 dB signal-to-noise ratio).

There were six target–masker combinations in total: AU–AU, AU–US, AU–SE, US–AU, US–US, and US–SE (see Table 1). There were 10 sentences for each of the six conditions, for a total of 60 trials.

Table 1: *Target–masker conditions with information about linguistic similarity and listener familiarity*

| Target–masker pair | Target–masker similarity | Familiarity to listeners |
|---|---|---|
| AU–AU | Matched | Native target and masker |
| AU–US | Different dialects | Native target, unfamiliar masker |
| AU–SE | Different languages | Native target, foreign masker |
| US–AU | Different dialects | Unfamiliar target, native masker |
| US–US | Matched | Unfamiliar target and masker |
| US–SE | Different languages | Unfamiliar target, foreign masker |

## 2.3. Procedure

Participants completed the experiments remotely on a Windows computer running E-Prime Go (version 1.0.2.41), a packaged version of E-Prime [15]. To ensure that participants were Australian English monolinguals and that they had no vision, hearing, or language impairments, participants also completed a questionnaire adapted from the Language Experience and Proficiency Questionnaire [16]. Participants were instructed to complete the task in a quiet environment using wired headphones set to a comfortable listening volume.

The experiment task consisted of a sentence recognition task. Participants heard a target sentence presented in two-talker masker babble, then were instructed to type what they heard to the best of their ability into the text box, before pressing the ENTER key to proceed to the next trial. All 60 sentences were presented in random order, regardless of target–masker condition.

The researchers manually scored the participants' responses. Sentence recognition accuracy was calculated based on whether participants correctly identified the capitalised words in their response, e.g., "The CLOWN had a FUNNY FACE." Leniency was given for responses made by human error, such as spelling mistakes or duplicate words.

# 3. Results

Data were submitted to a (2 × 3) repeated measures ANOVA, with the within-subjects variables of Target dialect (AU, US) and Masker language/dialect (AU, US, SE). There was a significant effect of Target dialect on sentence recognition accuracy $F(1,55) = 64.68$, $p < .001$, $\eta_p^2 = .540$. Listeners showed higher sentence recognition accuracy when the target talker was an AU speaker ($M = 51.5\%$) than a US speaker ($M = 37.3\%$). A significant main effect of Masker language/dialect was also observed, $F(2,110) = 144.12$, $p < .001$, $\eta_p^2 = .724$.

Mean sentence recognition accuracy scores for all six target–masker conditions are shown in Figure 1.

There was an interaction effect between Target and Masker, $F(2,110) = 86.33$, $p < .001$, $\eta_p^2 = .611$. To investigate this interaction effect, a paired samples post-hoc $t$-test was conducted on sentence recognition accuracy, with the Bonferroni adjusted alpha set to .0167. There was a significant difference within the paired AU–AU and US–AU conditions, $t(55) = 6.29$, $p < .001$, indicating a higher mean sentence recognition accuracy for the AU–AU condition ($M$ = 38.3, $SD$ = 8.8) than for the US–AU condition ($M$ = 27.0, $SD$ = 16.1). A medium effect size ($r^2$ = .42) was found for the paired conditions ($M_{diff}$ = 11.29, 95% CI [1.79, 14.89]). A significant difference was also observed for the AU–US and US–US conditions, $t(55) = 13.04$, $p < .001$. Sentence recognition accuracy was higher for the AU–US condition ($M$ = 69.9, $SD$ = 12.6) than the US–US condition ($M$ = 39.7, $SD$ = 12.6), and a large effect size was found ($r^2$ = .76) was found between the conditions ($M_{diff}$ = 30.19, 95% CI [25.55, 34.83]). However, there was no significant difference between for the paired AU–SE and US–SE conditions, $t(55) = .444$, $p = .658$, nor was an effect size present between these two conditions ($r^2$ = .00).

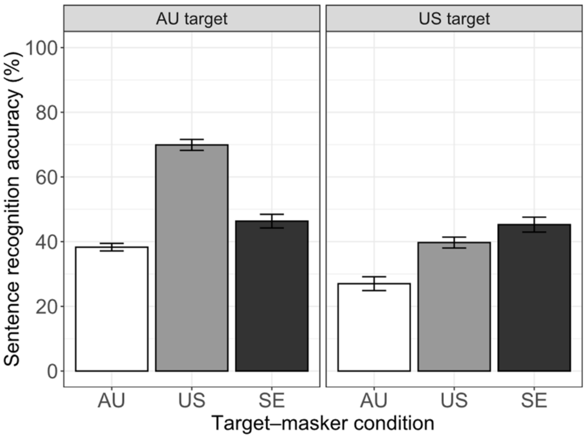

Figure 1: *Sentence recognition accuracy by target–masker condition. Error bars depict standard error of the mean (SEM).*

## 4. Discussion

The results lend partial support to the hypotheses, though some patterns in the findings require further investigation. When comparing between target–masker pairs (e.g., AU–US vs. US–US), there is a noticeable advantage when listening to a native target in the presence of either a native or unfamiliar masker dialect. However, performance was comparable between the AU–SE and US–SE conditions. This was also where results deviated from predictions. It was posited that the masker most unfamiliar to the listener (i.e., SE) would result in greater overall speech recognition performance and LRM between the native and unfamiliar targets, yet this was not the case—the SE masker was only slightly less inhibitive than the other maskers when US was the target dialect. The US masker provided the greatest degree of LRM between target conditions. As expected, the AU masker was most detrimental to the listeners overall.

The additive effects of target–masker mismatch and familiarity to dialect seem to be present for the AU–US condition, where performance was highest for a familiar target (AU) and mutually intelligible but unfamiliar masker (US). Conversely, performance was poorest when a mutually intelligible but unfamiliar target (US) was presented alongside a familiar masker (AU). Even when accounting for mismatches between the target and masker, performance in the mismatched US–AU condition was still poorer than the matched US–US or AU–AU conditions. This suggests a greater weighting in performance towards listener familiarity than target–masker mismatch, whereby having an unfamiliar target and native masker results in greater interference over an unfamiliar target–masker pair or a native target–masker pair. This finding contradicts that of Viswanathan et al. [5], where higher performance was observed for a mismatched native target and foreign masker pair (English–Dutch) compared to a matched English–English pair.

Some results in the current study do not reflect findings from previous research. Unlike in Jacewicz and Fox [9], where the unfamiliar target (Southern US) was more intelligible to the General US listeners, the AU listeners in the current study found the AU target more intelligible than the US target, hence the overall higher accuracy in the AU target conditions.

This may be attributed to the fact that AU and US are more distinct phonetically than Southern US and General US. Also, contrary to earlier predictions that a foreign language would provide the greatest LRM outcomes, performance was not the highest for the AU–SE condition. This finding is in conflict with past studies showing greater performance when listening to native target and unfamiliar/foreign masker pairs [2, 5, 6]. A possible explanation for this result is the phonetic features of Swedish: it contains many more vowels than English (many of which both languages share) [17]. This may result in more overlaps between the target and masker signals compared to other foreign languages. A potential future research topic could investigate how the degree of phonetic overlap between a target and masker language system influences speech recognition. This would differ from the previous work conducted on similarities between target and masker languages at the prosodic level (i.e., rhythm), where findings have been contested [7, 18].

Future iterations of this study can improve on the current experiment design. As a study conducted remotely, the listening environment for each participant will vary in testing location and headphones used. More measures could be taken to monitor or control for these variables, including participant self-reports of the above details, as well as screening tasks to confirm whether participants are listening to stimuli through headphones [19]. In addition, normalising the long-term average speech spectra (LTAS) of the maskers would help reduce the impact of energetic masking between conditions. This procedure has been implemented in past speech-in-speech research [6, 7, 20], and is especially relevant for tasks with a higher number of masker language conditions.

## 5. Conclusion

The present study aimed to investigate the effect of dialect variation and foreign language on sentence recognition performance in monolingual Australian English listeners. Results showed that linguistic release from masking can occur between two mutually intelligible dialects of a language with varying levels of familiarity to the listener. This was demonstrated by the differences in performance between target–masker pairs depending on the listeners' familiarity to the target and masker (i.e., AU–AU vs. US–AU, and AU–US vs. US–US). There is an interplay between target–masker similarity and the listeners' familiarity to the competing signals such that the latter holds greater weight than the former for listening outcomes: AU–US was the condition with the highest performance, while US–AU led to the worst performance overall. In sum, a native target and unfamiliar masker was most optimal, while an unfamiliar target and native masker was most detrimental.

It was anticipated that the foreign language masker, Swedish, would provide greater LRM outcomes due to its dissimilarity to both English dialects. However, SE was only least inhibitive in the (unfamiliar) US target condition, and performance between the AU–SE and US–SE pair was comparable. It is possible that the phonetic characteristics of Swedish, such as its large vowel system, made it a difficult masker to separate from the English signal. Future research could compare speech recognition performance between different masker languages by categorising their similarity to the target language at the phonetic (rather than solely prosodic) level.

## 6. Acknowledgements

We would like to thank Jade Nguyen, Gabrielle Teklic, Bishoy Boulos, and Adau Aher for assisting in the development of the experimental stimuli.

## 7. References

[1] I. Pollack, "Auditory informational masking," *The Journal of the Acoustical Society of America*, vol. 57, no. S1, pp. S5–S5, Apr. 1975.

[2] L. Kilman, A. Zekveld, M. Hällgren, and J. Rönnberg, "The influence of non-native language proficiency on speech perception performance," *Frontiers in Psychology*, vol. 5, Jul. 2014.

[3] S. Brouwer, "Masking release effects of a standard and a regional linguistic variety," *The Journal of the Acoustical Society of America*, vol. 142, no. 2, pp. EL237–EL243, Aug. 2017.

[4] P. Adank, B. G. Evans, J. Stuart-Smith, and S. K. Scott, "Comprehension of familiar and unfamiliar native accents under adverse listening conditions," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 35, no. 2, pp. 520–529, Apr. 2009.

[5] N. Viswanathan, K. Kokkinakis, and B. T. Williams, "Spatially separating language masker from target results in spatial and linguistic masking release," *The Journal of the Acoustical Society of America*, vol. 140, no. 6, pp. EL465–EL470, Dec. 2016.

[6] S. Brouwer, K. J. Van Engen, L. Calandruccio, and A. R. Bradlow, "Linguistic contributions to speech-on-speech masking for native and non-native listeners: Language familiarity and semantic content," *The Journal of the Acoustical Society of America*, vol. 131, no. 2, pp. 1449–1464, Feb. 2012.

[7] V. A. Brown, N. H. Dillman-Hasso, Z. Li, L. Ray, E. Mamantov, K. J. Van Engen, and J. F. Strand, "Revisiting the target-masker linguistic similarity hypothesis," *Attention, Perception, & Psychophysics*, Apr. 2022.

[8] A. L. Francis, L. J. Tigchelaar, R. Zhang, and A. A. Zekveld, "Effects of second language proficiency and linguistic uncertainty on recognition of speech in native and nonnative competing speech," *Journal of Speech, Language & Hearing Research*, vol. 61, no. 7, pp. 1815–1830, Jul. 2018.

[9] E. Jacewicz and R. A. Fox, "The effects of dialect variation on speech intelligibility in a multitalker background," *Applied Psycholinguistics*, vol. 36, no. 3, pp. 729–746, May 2015.

[10] L. Calandruccio, S. Dhar, and A. R. Bradlow, "Speech-on-speech masking with variable access to the linguistic content of the masker speech," *The Journal of the Acoustical Society of America*, vol. 128, no. 2, pp. 860–869, Aug. 2010.

[11] L. Calandruccio, A. R. Bradlow, and S. Dhar, "Speech-on-speech masking with variable access to the linguistic content of the masker speech for native and nonnative English speakers," *Journal of the American Academy of Audiology*, vol. 25, no. 04, pp. 355–366, Apr. 2014.

[12] J. Bench, Å. 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, Jan. 1979.

[13] P. W. Nye and J. H. Gaitenby, "The intelligibility of synthetic monosyllabic words in short, syntactically normal sentences," in *Haskins Laboratories Status Report on Speech Resolution, SR-37/38*, 1974, pp. 169–190.

[14] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer." 2022.

[15] Psychology Software Tools, Inc., "E-Prime Go." 2020.

[16] V. Marian, H. K. Blumenfeld, and M. Kaushanskaya, "The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals," *Journal of Speech, Language, and Hearing Research*, vol. 50, no. 4, pp. 940–967, Aug. 2007.

[17] E. Konig and J. van der Auwera, *The Germanic Languages*. London: Routledge, 2013.

[18] L. Calandruccio, S. Brouwer, K. J. Van Engen, S. Dhar, and A. R. Bradlow, "Masking release due to linguistic and phonetic dissimilarity between the target and masker speech," *American Journal of Audiology*, vol. 22, no. 1, pp. 157–164, Jun. 2013.

[19] K. J. P. Woods, M. H. Siegel, J. Traer, and J. H. McDermott, "Headphone screening to facilitate web-based auditory experiments," *Attention, Perception, & Psychophysics*, vol. 79, no. 7, pp. 2064–2072, Oct. 2017.

[20] S. Brouwer, "The role of foreign accent and short-term exposure in speech-in-speech recognition," *Attention, Perception, & Psychophysics*, vol. 81, no. 6, pp. 2053–2062, Aug. 2019.