PERCEIVING SPEECH PRODUCED WITH FACE MASKS IN COMPETING TALKER ENVIRONMENTS

Faith Chiu[1], Laura Bartoševičiūtė[2], Albert Lee[3], Yujia Yao[1]

[1]Glasgow University Laboratory of Phonetics, University of Glasgow, [2]University of Essex, [3]The Education University of Hong Kong
faith.chiu@glasgow.ac.uk, albertlee@eduhk.hk

ABSTRACT

This paper examines the perception of speech produced with face masks in everyday multi-talker environments. Three groups of participants listened to English target sentences produced with or without a face mask in the presence of English or Lithuanian competing speech. Participants were monolingual English listeners, and second language English listeners with either Lithuanian or Mandarin Chinese as first language (L1). Lithuanian listeners also completed the experiment with Lithuanian target sentences. Participants were generally more accurate perceiving sentences produced without a face mask, and when listening in L1. Competing speech in a language matching the target lowered perception accuracy. Exceptionally, only when Lithuanian participants (with both English and Lithuanian knowledge) listened for Lithuanian targets was there no added challenge from matching language of target and competing speech. We conclude that acoustic distortions from face masks present an across-the-board difficulty while linguistic knowledge can reduce distraction from competing speech.

Keywords: face masks, competing talker, speech perception

1. INTRODUCTION

The COVID-19 pandemic has reshaped speech communication. Face-to-face communication often includes one or both parties sporting a face mask. The listener's comprehension effort now involves adapting to mask-imposed distortions to the acoustic speech signal [1, 2], not to mention other challenges such as deprivation of visual cues. It has been shown that individuals struggle with understanding speech produced with a face mask when presented in noise [3], even for native speakers [4]. Furthermore, most everyday speech communication also takes place in less than ideal listening environments such as a noisy background with a competing talker. It is also not uncommon these days to converse in one's second or additional language. Previous research has found that non-native listeners perform significantly worse than native listeners in perceiving speech presented with intelligible competing background speech [5, 6]. This study aims to understand the difficulty imposed by speech produced with face masks in an everyday multi-talker environment. Target sentences produced with and without a face mask were presented to listeners in the presence of a competing talker. The competing speech either matched or differed in language from target sentences. Participants' linguistic background determined the intelligibility of the competing talker.

2. METHOD

2.1. Stimuli

The speech material consisted of target sentences in English and in Lithuanian, and competing speech in English and in Lithuanian.

2.1.1. Target sentences

English target sentences were based on the British English version of the International Matrix sentence test for speech audiometry in noise [7]. This standardised test consists of a 50-word base matrix (10 names, 10 verbs, 10 numerals, 10 adjectives, and 10 nouns) from which grammatically correct but contextually unpredictable five-word sentences can be built, using a random combination of one word of each word category. All matrix sentences have fixed syntactic structure ('Alan bought two big beds'; name + verb + number + adjective + noun). Up to 100,000 different sentences can be generated.

Name	Verb	Number	Adjective	Noun
Alan	bought	two	big	beds
Barry	gives	three	cheap	chairs
Hannah	got	four	dark	desks
Kathy	has	five	green	mugs
Lucy	kept	six	large	rings
Nina	likes	eight	old	ships
Peter	sees	nine	pink	shoes
Rachel	sold	ten	red	spoons
Steven	wants	twelve	small	tins
Thomas	wins	some	thin	toys

Table 1: Matrix for English target sentences.

Lithuanian target sentences (only presented to Lithuanian listeners) follow the same format and are constructed as original stimuli for this experiment.

Name	Verb	Number	Adjective	Noun
Ieva	laiko	kelis	pilkus	raktus
Miglė	rodo	visus	juodus	krepšius
Agnė	turi	devynis	žalius	peilius
Dalia	mato	du	sunkius	ratus
Paulius	rado	tris	svarbius	batus
Marius	perka	penkis	mažus	rūbus
Nojus	gavo	šešis	naujus	šaukštus
Benas	neša	keturis	senus	žaislus
Juozas	davė	aštuonis	pigius	stalus
Rūta	piešia	septynis	baltus	daiktus

Table 2: Matrix for Lithuanian target sentences.

Audio stimuli from these target sentences were generated by recording a native female speaker of each language in sound attenuated booths. Individual words were produced in sentence frames specified above, with and without a cotton fabric face mask. The recordings were segmented and recombined on Praat [8] to generate the presented stimuli.

2.1.2. Competing speech

Competing speech was semantically meaningful sentences in either English or Lithuanian. These sentences were presented continuously at a difficult −10dB Signal-to-Noise ratio, meaning that the competing speech was 10dB louder than the presented target sentences. Male voices were chosen so participants can utilise speaker sex differences as a segregation cue. Low-level acoustic cues such as fundamental frequency differences can help with separating and tracking sentences in competition [9].

English competing speech contained sentences from the IEEE lists [10] produced by a native British English male speaker [11]. Lithuanian competing speech consisted of text being read aloud by a native Lithuanian male speaker from the LIEPA corpus [12]. Both these speakers did not wear a face mask.

2.2. Participants

2.2.1. Experiment 1

Twenty-four native Lithuanian listeners (thirteen female and eleven male) completed Experiment 1. All participants were between 18 and 37 years old and reported no history of hearing or language impairment. All participants had second language knowledge of English, and had completed either high school or tertiary education.

2.2.2. Experiment 2

Participants were twenty-two monolingual British English speakers (sixteen female and six male, age range: 18–34) and twenty-two second language speakers of English with Mandarin Chinese as first language (nineteen female and three male, age range: 20–31). All participants reported no history of hearing or language impairment and had completed either high school or tertiary education.

2.3. Procedure

2.3.1. Experiment 1

The experiment was conducted online using Psytoolkit [13, 14]. Participants were instructed to complete the task in a quiet room with headphones, and at a computer with keyboard input. After initial surveys on demographic information and English experience and proficiency (abbreviating LEAP-Q [15]), the experimental session started with a headphone check [16]. Then, participants were told that they would hear target sentences by a female talker in the presence of a male competing talker. The target sentences were cued by sex of the speaker. Participants were instructed to listen only to the female voice and ignore the male voice. They were then asked to type what they heard after each sentence, and were requested to report individual words if they had not heard the whole sentence.

In the main task, participants heard a total of 160 trials: from 2 Target language conditions (ENGLISH/LITHUANIAN) × 2 Mask conditions (YES/NO) × 2 Competing Speech language conditions (ENGLISH/LITHUANIAN) with 20 sentences each. These 160 trials were presented over four blocks of 40 sentences each. Presentation order of blocks were counterbalanced across participants and trials were randomised within each block. The task was self-paced and participants heard each sentence only once; there was no repetition of sentences in the same configuration. No feedback was provided. The total duration of the experimental session, without breaks, was around forty-five minutes. Participants were allowed to take breaks between blocks.

2.3.2. Experiment 2

The procedure in Experiment 2 was identical to Experiment 1, except that participants heard only English target sentences produced with and without mask, in the presence of either competing English or Lithuanian speech. There were thus a total of 80 main trials in Experiment 2.

2.4. Data scoring

Responses were scored based on the number of words accurately reported in each sentence. There

were five keywords per sentence; a full mark on a sentence was scored as 100%. Each error took away 20% (4 accurate keywords = 80%). Keywords were considered accurate regardless of the position they appeared in the participant's response. If the participant missed some of the keywords in the sentence, the remaining accurately reported words were still scored (e.g. 'Lucy small desks'). Variations in spelling (e.g. 'Kathy' as 'Cathy') were considered an accurate response. Phonologically similar words that change the meaning of the word were considered incorrect (e.g. 'wings' for 'rings').

3. RESULTS

3.1. Experiment 1

Lithuanian listeners' perception performance is given in Figure 1.

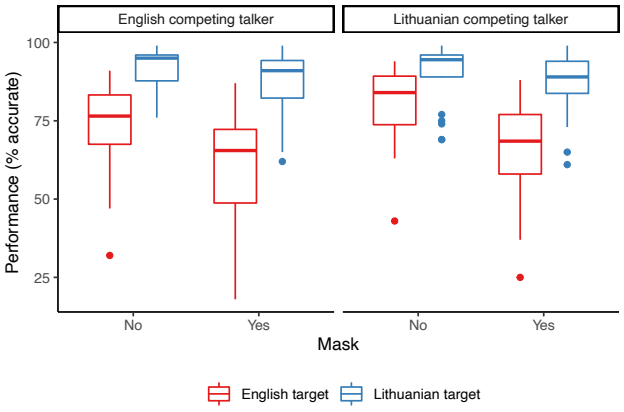

Figure 1: Performance (% accurate) of Lithuanian listeners across all conditions.

A 2 × 2 × 2 analysis of variance (ANOVA) was conducted on the percentage of accurately reported keywords as a function of language of Target (English versus Lithuanian), Mask (with or without a face mask), and language of Competing Speech (English versus Lithuanian). The results revealed two significant two way-interactions: Target × Mask ($F(1, 23) = 26.001$, $p < .001$, $\eta p^2 = .531$) and Target × Competing Speech ($F(1, 23) = 25.123$, $p < .001$, $\eta p^2 = .522$). All three main effects were significant, for Target ($F(1, 23) = 59.409$, $p < .001$, $\eta p^2 = .721$), Mask ($F(1, 23) = 59.536$, $p < .001$, $\eta p^2 = .721$) and Competing Speech ($F(1, 23) = 18.771$, $p < .001$, $\eta p^2 = .449$). The three-way interaction Target × Mask × Competing Speech was not significant, nor was the two-way interaction Mask × Competing Speech (both $p > .050$).

Individual 2 × 2 ANOVAs were performed for each Target language in order to follow up these observations, comparing the percentage of accurately reported keywords as a function of Mask (with or without a mask) and language of Competing Speech (English versus Lithuanian). For English target sentences, the interaction term was not significant ($p > .050$). There was a significant main effect of Mask ($F(1, 23) = 54.448$, $p < .001$, $\eta p^2 = .703$). More keywords were accurately reported on sentences produced without a face mask, in both English and Lithuanian competing speech. There was also a main effect of Competing Speech ($F(1, 23) = 31.304$, $p < .001$, $\eta p^2 = .576$). Lithuanians listeners were less accurate when the competing speech was in a language which matches the English target sentences; this was true both when the targets were produced with and without a mask.

When listening to Lithuanian target sentences, there was only a main effect of Mask on the percentage of accurately identified keywords ($F(1, 23) = 15.544$, $p < .001$, $\eta p^2 = .403$). Lithuanian target sentences produced with a face mask were more poorly perceived, and this was true regardless of whether it was presented in both English and Lithuanian competing speech. Here, unlike the case with English target sentences, there was no main effect of Competing Speech. The interaction term was also not significant. (Both $p > .050$.) Thus, only when Lithuanian listened in their first language was there no added challenge from matching language of target sentences and competing speech.

In addition, Lithuanian listeners reported more accurate keywords when listening to target sentences in their first language (Lithuanian) compared to English. Running planned comparisons, this was found in all four conditions: when target sentences were produced with a mask and presented in English competing speech ($t(23) = 8.730$, $p < .001$) and in Lithuanian competing speech ($t(23) = 6.890$, $p < .001$), and when target sentences were produced without a face mask and delivered with an English ($t(23) = 6.834$, $p < .002$) and Lithuanian competing talker ($t(23) = 4.314$, $p < .001$).

3.2. Experiment 2

Figure 2 shows the percentage of accurately reported keywords by English and Chinese listeners.

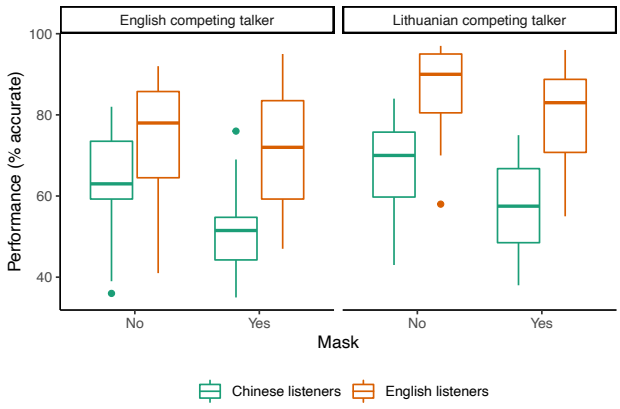

Figure 2: Performance of English listeners and Chinese listeners across all conditions.

A mixed 2 × 2 × 2 ANOVA was conducted on the percentage of accurately reported keywords as a function of Mask (with or without a face mask), language of Competing Speech (English versus Lithuanian), and Group (English versus Chinese listeners). The results indicated a significant three-way interaction of Mask × Competing Speech × Group ($F(1, 42) = 6.497$, $p = .015$, $\eta p^2 = .134$).

Planned comparisons showed English listeners outperforming Chinese listeners in all conditions: when target sentences were produced with a face mask and presented in English competing speech ($t(42) = 5.009$, $p < .001$) and in Lithuanian competing speech ($t(42) = 6.162$, $p < .001$), and when target sentences were produced without a face mask and delivered with an English ($t(42) = 2.767$, $p = .008$) and Lithuanian competing talker ($t(42) = 5.776$, $p < .001$).

Individual 2 × 2 ANOVAs were performed for each listener group. For English listeners, there was a main effect of Mask ($F(1, 21) = 5.439$, $p = .030$, $\eta p^2 = .206$) as well as a main effect of Competing Speech ($F(1, 21) = 78.729$, p. < .002, $\eta p^2 = .789$). The interaction term was not significant ($p > .050$). Chinese listeners were similar; there was a main effect of Mask. They too reported more accurate keywords when the target sentences were produced without a mask. This held, both when the target was presented in English and in Lithuanian competing speech ($F(1, 21) = 21.960$, $p < .001$, $\eta p^2 = .511$). There was also a main effect of Competing Speech in Chinese listeners ($F(1, 21) = 19.869$, $p < .001$, $\eta p^2 = .486$). Like English listeners, there was more of a detrimental effect on perception when the competing talker matched the language of the target sentences produced with and without a mask.

4. DISCUSSION

4.1. Across-the-board difficulty in perceiving speech produced with face masks

Overall, a main effect of Mask was found across all participant groups: in monolingual English listeners, and in second language English listeners with either Lithuanian or Chinese as first language. Target sentences produced with a mask were less accurately perceived. With Lithuanian listeners, the Mask effect was observed for target sentences both in their first (Lithuanian) and second (English) languages.

This finding echoes other reports of decreased perception performance when listening to speech produced with a face mask and presented in noise [3, 4]. Notably, all listener groups in all conditions experienced a drop in performance in perceiving speech produced with masks. This across-the-board effect could be due to attenuation of the acoustic signal from mask-wearing in the form of dampening. In particular, high frequency information is lost [17]. Further data analysis and future iterations of this study could consider specifically the identification of keywords involving fricatives and plosives. These segments tend to be commonly misperceived when produced with a face mask even in quiet [18]. The study could also benefit from additional auditory and acoustic analyses of the target sentence stimuli in terms of speaking style. Clear speech strategies have been documented with speakers wearing face masks and this instead increases signal intelligibility [19].

4.2. Target and competing talker language similarity

We also found that perception accuracy was higher when listening in one's first language. In Experiment 1, Lithuanian listeners reported more accurate keywords to Lithuanian rather than English target sentences. In Experiment 2, the native English group performed better than the second language group which had Chinese as first language. These results corroborate the findings of previous work [5, 6] showing that speech perception with a competing talker is more difficult in one's non-native language.

Additionally, we found that a competing talker in a language which matches the target sentences had more of a detrimental effect on perception accuracy compared to one that was mismatched. This was found in all listener groups (Lithuanian, English and Chinese) when listening to English target sentences. The results with Lithuanian and Chinese listeners on English target sentences replicate existing findings documenting a benefit of linguistic mismatch between target and competing speech for non-native speakers [20, 21, 22]. This is in line with Brouwer and colleagues' linguistic similarity hypothesis [20].

However, there was no effect of Competing Speech language in our study when Lithuanians listened to target sentences in their native language. We believe this is due to the group's knowledge of both the target and competing talker languages. Our study contradicts [20] in their native bilingual group. When Dutch speakers of second language English listened in their native language, they behaved instead like monolinguals presented with a competing talker in a foreign unintelligible language [20] (i.e. less interference listening to English-on-Dutch as opposed to Dutch-on-Dutch). To validate our finding, we could attempt replicating our results with another speaker for Lithuanian target sentences or to include different competitor talkers. This will help ensure that the observed effects are not just due energetic masking differences across our two competing talkers: that the English competing talker was not merely more effective at swamping (rendering more time-frequency regions inaudible from more overlaps of the target signal), and that the Lithuanian competing talker has not just allowed more glimpses of the target signal (regions in which the target is least affected) [23].

Acknowledgements: This research was supported by a British Academy grant awarded to FC and AL (COV19\2008370).

5. REFERENCES

[1] Corey, M. R., Jones, U., Singer, C. A. 2020. Acoustic effects of medical, cloth, and transparent face masks on speech signals. *J. Acoust. Soc. Am.* 148, 2371–2375.

[2] Magee, M., Lewis, C., Noffs, G., Reece, H., Chan, J. C. S., Zaga, C. J., Paynter, C., Birchall, O., Rojas Azocar, S., Ediriweera, A., Kenyon, K., Caverlé, M. W., Schultz, B. G., Vogel, A. P. 2020. Effects of face masks on acoustic analysis and speech perception: Implications for peri-pandemic protocols. *J. Acoust. Soc. Am*. 148, 3562–3568.

[3] Toscano, J.C., Toscano, C.M. 2021. Effects of face masks on speech recognition in multi-talker babble noise. *PLOS ONE* 16, e0246842.

[4] Yi, H., Pingsterhaus, A., Song, W. 2021. Effects of wearing face masks while using different speaking styles in noise on speech intelligibility during the COVID-19 pandemic. *Front. Psychol.* 12, 682677.

[5] Bradlow, A. R., Alexander, J. A. 2007. Semantic and phonetic enhancements for speech-in-noise recognition by native and non-native listeners. *J Acoust. Soc. Am.* 121, 2239–2249.

[6] Cooke, M., Garcia Lecumberri, M.L., Barker, J. 2008. The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception. *J. Acoust. Soc. Am.* 123, 414–427.

[7] Hall, S. J. 2006. *The development of a new English Sentence in Noise Test and an English Number Recognition Test*. MSc thesis. University of Southampton.

[8] Boersma, P., Weenink, D. 2022. Praat: doing phonetics by computer [Computer program]. Version 6.3.03.

[9] Brungart, D.S. 2001. Informational and energetic masking effects in the perception of two simultaneous talkers. *J Acoust Soc Am.* 109, 1101–1109.

[10] Rothauser, E. H., Chapman, W. D., Guttman, N., Hecker, M. H. L., Nordby, K. S., Silbiger, H. R., Urbanek, G. E., Weinstock, M. 1969. IEEE recommended practice for speech quality measurements. *IEEE Trans. Audio Electroacoust*. 17, 225–246.

[11] Millman R.E., Mattys S.L. 2017. Auditory Verbal Working Memory as a Predictor of Speech Perception in Modulated Maskers in Listeners With Normal Hearing. *J. Speech Lang. Hear. Res.* 60, 1236–1245.

[12] Laurinčiukaitė, S., Telksnys, L., Kasparaitis, P., Kliukienė, R., Paukštytė, V. 2018. Lithuanian Speech Corpus Liepa for Development of Human-Computer Interfaces Working in Voice Recognition and Synthesis Mode. *Informatica* 29, 487–498.

[13] Stoet, G. 2010. PsyToolkit - A software package for programming psychological experiments using Linux. *Behav. Res. Methods* 42, 1096–1104.

[14] Stoet, G. 2017. PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teach Psychol.* 44, 24–31.

[15] Marian, V., Blumenfeld, H.K., Kaushanskaya, M. 2007. The Language Experience and Proficiency Questionnaire (LEAP-Q): assessing language profiles in bilinguals and multilinguals. *J. Speech Lang. Hear. Res.* 50, 940–967.

[16] Woods, K.J.P., Siegel, M.H., Traer, J., McDermott, J.H. 2017. Headphone screening to facilitate web-based auditory experiments. *Atten. Percept. Psychophys.* 79, 2064–2072.

[17] Corey, R.M., Jones, U., Singer, A.C. 2020. Acoustic effects of medical, cloth, and transparent face masks on speech signals. *J. Acoust. Soc. Am.* 148, 2371–2375.

[18] Llamas, C., Harrison, P., Donnelly, D., Watt, D. 2008. Effects of different types of face coverings on speech acoustics and intelligibility. *York Papers in Linguistics* 2, 80–104.

[19] Cohn, M., Pycha, A., Zellou, G. 2021. Intelligibility of face-masked speech depends on speaking style: Comparing casual, clear, and emotional speech. *Cognition* 210, 104570.

[20] Brouwer, S., Van Engen, K.J., Calandruccio, L., Bradlow, A.R. 2012. Linguistic contributions to speech-on-speech masking for native and non-native listeners: language familiarity and semantic content. *J. Acoust. Soc. Am.* 131, 1449–64.

[21] Calandruccio, L., Zhou, H. 2014. Increase in speech recognition due to linguistic mismatch between target and masker speech: monolingual and simultaneous bilingual performance. *J. Speech Lang. Hear. Res.* 57, 1089–1097.

[22] Van Engen, K.J. 2010. Similarity and familiarity: Second language sentence recognition in first- and second-language multi-talker babble. *Speech Commun.* 52, 943–953.

[23] Cooke, M. 2006. A glimpsing model of speech perception in noise. *J. Acoust. Soc. Am.* 119, 1562–73.